Statistics 301 and 371, Blended: Course Notes [1]

Robert L. Wardrop

May 23, 2015

Contents

Part I

Inference Based on Randomization

Chapter 1

The Completely Randomized Design with a Numerical Response

A **Completely Randomized Design** (**CRD**) is a particular type of comparative study. The word *design* means that the researcher has a very specific protocol to follow in conducting the study. The word *randomized* refers to the fact that the process of **randomization** is part of the design. The word *completely* tells us that *complete randomization* is required, in contrast to some form of incomplete randomization, such as the randomized pairs design we will study later in these notes. What is a numerical response? See the following section.

1.1 Comparative Studies

So, what is a **comparative study**? Let's look at its two words, beginning with the word *study*. According to dictionary.com (`http://dictionary.reference.com`) the fifth definition of *study* is:

> Research or a detailed examination and analysis of a subject, phenomenon, etc.

This reasonably well fits what I mean by a study. Next, again according to dictionary.com, the first definition of compare (the root word of comparative) is:

> To examine (two or more objects, ideas, people, etc.) in order to note similarities and differences.

Because of time limitations, for the most part in these notes we will restrict attention to *exactly two*, as opposed to *two or more*, *things* being compared.

In the examples of the first two chapters, **Dawn** wants to compare two flavors of cat treats; **Kymn** wants to compare two settings on an exercise machine; **Sara** wants to compare two golf clubs; and **Cathy** wants to compare two routes for jogging. In the practice and homework problems of these first two chapters you will be introduced to several other comparative studies. Indeed, a large majority of the chapters in this book are devoted to comparative studies. Why? Two reasons:

1. Comparative studies are extremely important in science.

2. The discipline of Statistics includes several good ideas and methods that help scientists perform and analyze comparative studies.

Next, some terminology: the two things being compared are called the **two levels** of the **study factor**. For our examples we have the following study factors and levels.

- Dawn's study factor is the *flavor of the cat treat*, with levels equal to *chicken-flavored* and *tuna-flavored*.

- For Kymn's study, her exercise apparatus is called an ergometer which requires two choices by its operator. Kymn's study factor is the machine setting with first level defined as *small gear with the vent closed*; her second level is *large gear with the vent open*.

- Sara's study factor is the golf club she used with levels *3-Wood* and *3-Iron*,

- Cathy's study factor is the route for her one mile run with levels *at her local high school* and *through the park*.

The remaining components of a comparative study are:

- The **units** that provide the researcher with information.

- The **response** which is the particular information from the unit of interest to the researcher.

- One of the following **methods**:

  - The researcher *identifies* each unit with its level of the study factor, or
  - The researcher *assigns* each unit to a level of the study factor.

I choose to introduce you to the units and the response for each of our studies in the various sections below. I do want to say a bit about the **method** in the last bullet of the above list.

Examples of *identifying*, sometimes called *classifying*, are: comparing men and women; comparing old people with young people; comparing residents of Wyoming with residents of Wisconsin. Our development of randomization-based inference—beginning with Chapter 3—in Part I of these notes, **will not consider any studies that involve identifying** units with levels.

As the last sentence implies, randomization-based inference is restricted to studies in which the researcher has the **option** of assigning units to levels. In fact, as the name suggests, we attend only to those studies in which the researcher *exercised the option* of assignment by using a method called **randomization**. You will learn about randomization in Chapter 3.

1.2 Dawn's Study; Various Tools

Dawn completed my class several years ago. In this section you will be introduced to Dawn's project.

The choice of a project topic, or, indeed, any research, should begin with a curiosity about how the world operates. Here is Dawn's motivation for her study, as she wrote in her report.

> I decided to study my cat's preference to two different flavors of the same brand-name cat treats. I was interested in this study because I figured that Bob, my cat, would prefer the tuna-flavored treats over the chicken-flavored because Bob absolutely loves canned tuna with a passion. My interest came when I looked at the ingredients on the two labels. I noticed that the percentage of real chicken in the chicken-flavored treats was larger than the percentage of real tuna in the tuna-flavored treats.

Thus, Dawn had a pretty good idea of what she wanted to study. Her next step was to operationalize the above notions into the standard language of a comparative study. We know her study factor and its levels from the previous section. Now, we need to specify: the definition of the **units** and the **response**.

A unit consisted of presenting Bob with a pile of ten treats, all of the same flavor. The flavor of the treats in the pile determined the level of the study factor for that unit: either chicken (level 1) or tuna (level 2). The response is the number of treats that Bob consumed in five minutes.

The technical term unit is not very descriptive. In this course there will be two types of units: **trials** and **subjects**. Dawn's units are trials. Essentially, we have trials if data collection involves doing something repeatedly. In Dawn's case this something is setting out a pile of ten treats. And then doing it again the next day. By contrast, in many studies the different units are different people. When the units are people, or other distinct objects, we call them subjects. As we will see later in these notes, sometimes the distinction between trials and subjects is blurry; fortunately, this doesn't matter.

Dawn decided to collect data for 20 days, with one trial per day. Dawn further decided to have 10 days assigned to each level of her study factor. (Sometimes, as here, I will speak of assigning units to levels. Other times I will speak of assigning levels to units. These two different perspectives are equivalent.) Dawn had to decide, of course, *which* days would be assigned to chicken-flavored and which days would be assigned to tuna-flavored. She made this decision by using a method called **randomization**. Randomization is very important. We will learn what it is in Chapter 3. In fact, the word *randomized* in CRD emphasizes that trials are assigned to levels by randomization. Without randomization, we have some other kind of comparative study; not a CRD.

For example, suppose a researcher wants to compare the heights of adult men and women. The study factor would be sex with levels male and female. Note that the researcher **most definitely cannot assign** subjects (individual adults) to level (female or male) by randomization or any other method! Sex as a study factor is an example of a **classification factor**, also called **observational factor** because each unit is *classified* according to the level *it possessed before entry into the study*. Thus, for example, Sally is a female before the study begins; she is not assigned *by the researcher* to be a female.

Table 1.1: Dawn's data on Bob's consumption of cat treats. 'C' ['T'] is for chicken [tuna] flavored.

Day:	1	2	3	4	5	6	7	8	9	10
Flavor:	C	T	T	T	C	T	C	C	C	T
Number Consumed:	4	3	5	0	5	4	5	6	1	7

Day:	11	12	13	14	15	16	17	18	19	20
Flavor:	C	T	C	T	C	C	T	C	T	T
Number Consumed:	6	3	7	1	3	6	3	8	1	2

Whenever units are assigned by randomization to levels, we call the levels **treatments**. Thus, if you hear a researcher talking about the treatments in a study, you may conclude that randomization was utilized.

When Dawn randomized, she arrived at the following assignment:

> Trials 1, 5, 7, 8, 9, 11, 13, 15, 16 and 18 are assigned to chicken-flavored treats and the remaining ten trials are assigned to tuna-flavored treats.

Very imprecisely, the purpose of a CRD is to learn whether the treatments influence the responses given by the units. For Dawn, this becomes: Does the flavor of the treat influence the number of treats that Bob consumes?

Here is an aside, especially for readers with a strong background in experimentation: A key word in the above is *influence*. In many types of studies, hoping to find influence is too optimistic; in these cases we can seek only an *association* between the two levels being compared and the response. As we will see later, randomization plays a key role here. Roughly speaking, with randomization we might find influence; without randomization, the most we can hope for is association.

There are several other important features of *how* Dawn performed her study, but I will defer them for a time and introduce the numbers she obtained. Dawn's data are presented in Table 1.1. Let me make a few brief comments about this display.

I use the terms specific to Dawn's study as my labels in this table; namely, day, flavor and number of treats consumed as opposed to trial (or unit), treatment and response. A major goal of mine is to develop a unified approach to CRDs and for this goal, general language is preferred. When we are considering a particular study, however, I prefer language that is as descriptive of the study's components as possible.

Take a few moments to make sure you understand the presentation in Table 1.1. For example, note that on day 6, Bob was presented tuna-flavored treats and he consumed four of them.

My next step in presenting Dawn's data is to separate, by treatment, the list of 20 numbers into two groups of 10. These appear in Table 1.2. Note that in this table I preserve the order in which the data, within treatment, were collected; e.g., the first time Bob was presented with chicken (day #1), he consumed 4 treats; the second time (day #5) he consumed 5 treats; and so on. I have done

Table 1.2: Dawn's data on Bob's consumption of cat treats, by treatment.

Observation Number:	1	2	3	4	5	6	7	8	9	10
Chicken :	4	5	5	6	1	6	7	3	6	8
Tuna:	3	5	0	4	7	3	1	3	1	2

Table 1.3: Dawn's data on Bob's consumption of cat treats, sorted within each treatment.

Position:	1	2	3	4	5	6	7	8	9	10
Chicken:	1	3	4	5	5	6	6	6	7	8
Tuna:	0	1	1	2	3	3	3	4	5	7

this because *sometimes* a researcher wants to explore whether there is a *time-trend* in the data. We won't look for a time-trend in this study, in part, to keep this presentation simple.

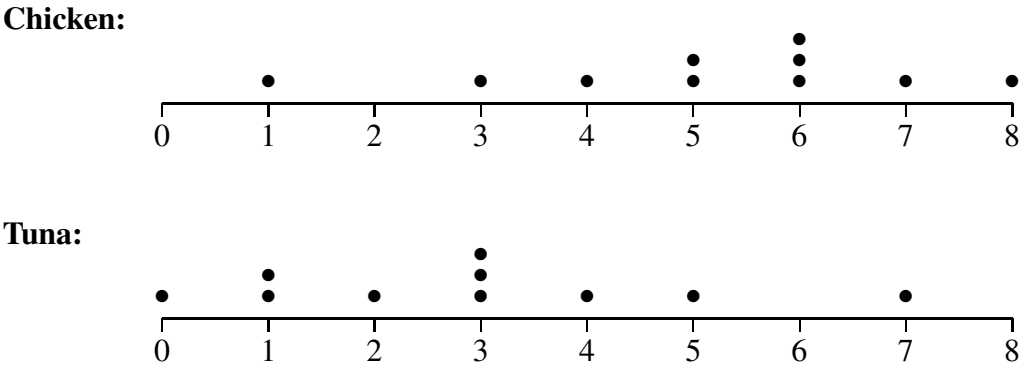
Usually it is useful to sort—always from smallest to largest in Statistics—the data within each treatment. I present these new lists in Table 1.3. Of the three tables I have created with Dawn's data, I find the sorted data easiest to interpret quickly. For example, I can see easily the smallest and largest response values for each treatment and, more importantly, that, as a group, the chicken responses are substantially larger than the tuna responses.

In Statistics we talk a great deal about *within group* and *between group* variation. *Within* the sorted list of chicken data, I see variation. Indeed, I would say that there is a great deal of day-to-day variation in Bob's consumption of chicken-flavored treats. Similarly, I see a great deal of variation within the sorted tuna data. Finally, I see a substantial amount of variation between the flavors: the responses to chicken are clearly larger, as a group, than the responses to tuna. In fact, you can easily verify that overall Bob ate 51 chicken treats compared to only 29 tuna treats.

Statisticians and scientists find it useful to draw a picture of data. We will learn a variety of pictures, starting with the **dot plot** also called the **dot diagram**. The dot plots for Dawn's data are in Figure 1.1. Some of you are already familiar with dot plots. Others may find them so obvious that no explanation is needed, but I will give one anyways. I will *explain* a dot plot by telling you how to construct one. Look at the dot plot for chicken. First, I draw a segment of the number line that is sufficient to represent all data, using the method described below. The plot contains 10 dots, one for each of the 10 chicken responses. Dots are placed above each response value. When a particular response value occurs more than once, the dots are stacked so that we can see how many are there. For example, there are three dots above the number '6' in the dot plot of the chicken data because on three chicken-days Bob consumed six treats.

Statisticians enjoy looking at a dot plot and labeling its shape. I don't see a shape in either of these dot plots. Indeed, I would argue that it is extremely unusual to see a shape in a small amount of data. One thing that I do see is that **neither of these dot plots is symmetric**. Left-to-right symmetry is a big deal in Statistics-pictures (as we will see). With real data perfect left-to-right symmetry is extremely rare and we are usually happy to find approximate symmetry. In fact, I

Figure 1.1: The dot plots for the cat treat study.



would say that the tuna dot plot is approximately symmetric. You may reasonably disagree and it turns out not to matter much in this current example.

Do you recall my earlier remarks about within and between variation in these data? I comment now that these features are easier to see with the dot plots than with the listings of sorted data. This is a big reason why I like dot plots: Sometimes they make it easier to *discover* features of the data.

It is, of course, a challenge to look at 10 (for either treatment) or 20 (for comparing treatments) response values and make sense of them. Thus, statisticians have spent a great deal of time studying various ways to summarize *a lot of numbers* with *a few numbers.* This can be a fascinating topic (well, at times, for statisticians, if not others) because of the following issues:

1. Are there any justifications for selecting a particular summary?

2. For a given summary, what are its strengths and desirable properties?

3. What are the weaknesses of a given summary?

Statisticians classify summaries into three broad categories. (There is some overlap between these categories, as you will learn later.)

1. Measures of center. Examples: mean; median; mode.

2. Measures of position. Examples: percentiles, which include quartiles, quintiles and median. Also, percentiles are equivalent to quantiles.

3. Measures of variation (also called measures of spread). Examples: range; interquartile range; variance and standard deviation.

Don't worry about all of these names; we will learn a little bit about some of them now and will learn more later. I suspect that many of you already know a little or a lot about several of these summaries. (For example, my granddaughter learned about medians at age nine in fourth grade.)

For Dawn's data we will learn about the three measures of center listed above. To this end, it will be convenient to adopt some mathematical notation and symbols. We denote the data from the first [second] treatment by the letter x [y]. Thus, Dawn's chicken data are denoted by x's and her tuna data are denoted by y's. We distinguish between different observations by using subscripts. Thus, in symbols, Dawn's chicken data are:

$$x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}.$$

Obviously, it was very tedious for me to type this list and, thus, in the future I will type simply $x_1, x_2, \ldots x_{10}$. Similarly, Dawn's tuna data are denoted by $y_1, y_2, \ldots y_{10}$.

Our next bit of generalization is needed because we won't always have 10 observations per treatment. Let n_1 denote the number of observations on treatment 1 and n_2 denote the number of observations on treatment 2. For Dawn's data, of course, $n_1 = n_2 = 10$. Whenever $n_1 = n_2$ we say that the study is **balanced**.

The subscripts on the x's and y's denote the order in which the data were collected. Thus, for example, $x_1 = 4$ was the response on the first chicken-day; $x_2 = 5$ was the response on the second chicken-day; and so on.

We will need notation for the sorted data too. We try to avoid making notation unnecessarily confusing. Thus, similar *things* have similar notation. For the sorted data we still use x's and y's as above and we still use subscripts, but we denote sorting by placing the subscripts inside parentheses. Thus, Dawn's sorted chicken data are:

$$x_{(1)}, x_{(2)}, x_{(3)}, x_{(4)}, x_{(5)}, x_{(6)}, x_{(7)}, x_{(8)}, x_{(9)}, x_{(10)}.$$

More tediously, for Dawn's chicken data

$$x_{(1)} = 1, x_{(2)} = 3, x_{(3)} = 4, x_{(4)} = 5, x_{(5)} = 5, x_{(6)} = 6, x_{(7)} = 6, x_{(8)} = 6, x_{(9)} = 7, x_{(10)} = 8.$$

Note that the collection of sorted values

$$x_{(1)}, x_{(2)}, \ldots x_{(n_1)}$$

is called the **order statistics** of the data.

The **mean** of a set of numbers is its arithmetic average. For example, the mean of 5, 1, 4 and 10 is:

$$(5 + 1 + 4 + 10)/4 = 20/4 = 5.$$

We don't really need a mathematical formula for the mean, but I will give you one anyways. Why? Well, later you will need to be comfortable with some formulas of this *type*, so we might as well introduce an easy one now.

Suppose we have m numbers denoted by

$$w_1, w_2, w_3 \ldots w_m.$$

The mean of these m numbers is

$$\bar{w} = \frac{w_1 + w_2 + w_3 + \ldots + w_m}{m} = \frac{\sum_{i=1}^{m} w_i}{m} = \frac{\sum w}{m}. \tag{1.1}$$

As you can see, we denote the mean by $\bar{w}$, read w-bar. For our notation for a CRD, we will have means denoted by $\bar{x}$ and $\bar{y}$. Often in these notes I will be informal in my use of summation notation; for example, $\sum w$ in these notes will be an instruction to sum all the w's in the problem at hand.

Note, of course, that for any set of data, summing sorted numbers gives the same total as summing unsorted numbers. You may easily verify that for Dawn's data:

$$\bar{x} = 51/10 = 5.1 \text{ and } \bar{y} = 29/10 = 2.9.$$

In words, in Dawn's study, the mean number of chicken treats eaten by Bob is larger than the mean number of tuna treats eaten by Bob. Usually (but not always; exceptions will be clearly noted), we compare two numbers by subtracting. Thus, because $5.1 - 2.9 = 2.2$ we will say that the mean number of chicken treats eaten by Bob is 2.2 larger than the mean number of tuna treats eaten by Bob.

The idea of the **median** of a set of numbers is to find the number in the center position of the sorted list. This requires some care because the method we use depends on whether the sample size is an odd number or an even number. For example, suppose we have five sorted numbers: 1, 3, 6, 8 and 8. There is a unique center position, position 3, and the number in this position, 6, is the median. If, however, the sample size is even, we need to be more careful. For example, consider four sorted numbers: 1, 4, 5 and 10. With four positions total, positions 2 and 3 have equal claim to being a center position, so the median is taken to be the arithmetic average of the numbers in positions 2 and 3; in this case the median is the arithmetic average of 4 and 5, giving us 4.5.

For both sets of Dawn's data (chicken and tuna) there are 10 numbers; hence, there are two center positions, namely positions 5 and 6. If you look at Table 1.3 on page 7 again, you will see that the median for Dawn's chicken data is $(5 + 6)/2 = 5.5$ and the median for her tuna data $(3 + 3)/2 = 3$.

If we have m numbers denoted by w's then the median is denoted by the symbol $\tilde{w}$, which is read as *w-tilde*. There are two formulas for calculating the median.

- If the sample size m is an odd integer, define $k = (m + 1)/2$, which will be an integer.

$$\tilde{w} = w_{(k)} \tag{1.2}$$

- If the sample size m is an even integer, define $k = m/2$, which will be an integer.

$$\tilde{w} = (w_{(k)} + w_{(k+1)})/2 \tag{1.3}$$

You are *not required* to use these formulas. Usually, I find it easier to visually locate the center position(s) of a list of sorted data.

I have one additional comment on Equation 1.2, which applies to many of the equations and formulas in these notes. When you are reading this equation, **do not** have your inner-voice say, "w-tilde equals w-sub-parentheses-k." This sounds like gibberish and won't help you learn the material. Instead, read what the equation *signifies*. In particular, I recommend reading the equation as "We obtain the median by finding the number in position k of the sorted list." Similarly, I read Equation 1.3 as, "We obtain the median by taking the arithmetic average of two numbers. The first

of these numbers is the number in position k of the sorted list. The second of these numbers is immediately to the right of the first number."

We won't use the mode much in these notes, but I will mention it now for completeness. It is easiest to explain and determine the mode by looking at a dot plot of the data. Refer to Figure 1.1 on page 8. In the chicken dot plot the tallest stack of dots occurs above the number 6. Thus, 6 is the mode of the chicken data. Similarly, the mode of the tuna data is 3. If two (or more) response values are tied for being most common, they are both (all) called modes. As an extreme case, with a set of m distinct numbers, every response value is a mode. It seems bizarre to claim that reporting m values of the mode is a *summary* of the m observations!

1.2.1 A Connection Between Pictures and Numbers

For any set of numbers, there is a very strong connection between their dot plot and their mean.

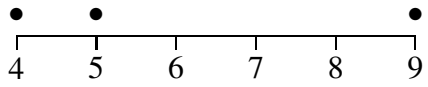
Suppose that we have m numbers denoted by $w_1, w_2, \ldots w_m$. As usual, let $\bar{w}$ denote the mean of these numbers. From each w_i we can create a new number, called its **deviation**, which is short for **deviation from the mean**. We create a deviation by taking the number and subtracting the mean from it. Symbolically, this gives us the following m deviations:

$$w_1 - \bar{w}, w_2 - \bar{w}, \ldots, w_m - \bar{w}.$$

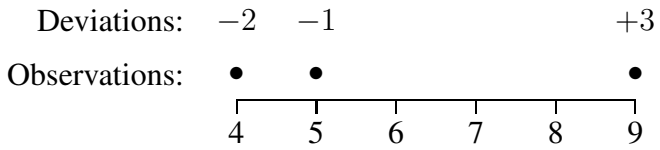
Let's have a quick example. Suppose that $m = 3$ and the three observations are 4, 5 and 9, which gives a mean of $\bar{w} = 6$. Thus, the deviations are

$$4 - 6, 5 - 6 \text{ and } 9 - 6; \text{ or } -2, -1 \text{ and } 3.$$

Below is a dot plot of our three numbers.



Below is the same dot plot with the deviations identified.



The following points are obvious, but I want to mention them anyways:

- The observation 4 has deviation -2 because it is two units smaller than the mean.
- The observation 5 has deviation -1 because it is one unit smaller than the mean.
- The observation 9 has deviation $+3$ because it is three units larger than the mean.

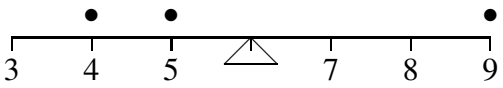
In general, **a non-zero deviation** has a sign (positive or negative) and a magnitude (its absolute value). Thus, the deviations for the three observations 4, 5 and 9 have signs: negative, negative and positive and magnitudes 2, 1 and 3, both respectively. The sign of a deviation tells us whether its observation is smaller than (negative) or larger than (positive) the mean. The magnitude of a deviation tells us *how far* its observation is from the mean, regardless of direction. Note, of course, that an observation has deviation equal to zero (which has no sign, being neither positive nor negative, and magnitude equal to 0) if, and only if, it is equal to the mean of all the numbers in the data set.

You have probably noticed that the three deviations for my data set sum to zero; this is not an accident. For any set of data:

$$\sum_{i=1}^{m}(w_i - \bar{w}) = \sum_{i=1}^{m} w_i - m\bar{w} = m\bar{w} - m\bar{w} = 0.$$

In words, for any set of data, the sum of the deviations equals zero. Statisticians (and others) refer to this property by saying that the mean is equal to the **center of gravity** of the numbers. If this terminology seems a bit mysterious or arbitrary, perhaps the following will help.

Below I have once again drawn the dot plot of my data set of $m = 3$ numbers: 4, 5 and 9, with one addition to the picture.



I have placed a fulcrum at the value of the mean, 6. Imagine the number line as the *board* on a seesaw and imagine that this board has no mass. Next, imagine each dot as having the same mass. Next, view the three dots as three equal weight (equal mass) *children* sitting on the board. We can see that with this scenario that with the fulcrum placed at the mean, 6, the seesaw will balance.

Look at the dot plots for Dawn's data in Figure 1.1 on page 8. First, consider the chicken data. We can see quickly that if we placed a fulcrum at 5, the picture would *almost balance*; in fact, recall, that $\bar{x} = 5.1$. Similarly, looking at the tuna data, we can see quickly that if we placed a fulcrum at 3, the picture would *almost balance*; in fact, recall, that $\bar{y} = 2.9$. Thus, we can look at a dot plot and get a quick and accurate idea of the value of the mean.

1.3 The Standard Deviation

I have mentioned, for example, the within-chicken variation in Dawn's data. We need a number that summarizes this variation. The summary we choose is a function of the deviations defined above. Actually, there are two measures of variation, also called spread, that we will investigate: the **variance** and the **standard deviation**. These two measures are very closely related; the standard deviation is the square root of the variance. Or, if you prefer, the variance is the square of the standard deviation. As a rough guide, statisticians prefer to measure spread with the standard deviation while mathematicians prefer to use the variance. As the course unfolds, you can decide

Table 1.4: The computation of the variances and standard deviations for Dawn's data on Bob's consumption of cat treats.

	Chicken			Tuna		
Observation	x	$x-\bar{x}$	$(x-\bar{x})^2$	y	$y-\bar{y}$	$(y-\bar{y})^2$
1	4	-1.1	1.21	3	0.1	0.01
2	5	-0.1	0.01	5	2.1	4.41
3	5	-0.1	0.01	0	-2.9	8.41
4	6	0.9	0.81	4	1.1	1.21
5	1	-4.1	16.81	7	4.1	16.81
6	6	0.9	0.81	3	0.1	0.01
7	7	1.9	3.61	1	-1.9	3.61
8	3	-2.1	4.41	3	0.1	0.01
9	6	0.9	0.81	1	-1.9	3.61
10	8	2.9	8.41	2	-0.9	0.81
Total	51	0.0	36.90	29	0.0	38.90
Mean	$\bar{x}=51/10=5.1$			$\bar{y}=29/10=2.9$		
Variance	$s_1^2=36.9/9=4.100$			$s_2^2=38.9/9=4.322$		
Stand. Dev.	$s_1=\sqrt{4.1}=2.025$			$s_2=\sqrt{4.322}=2.079$		

which measure you prefer. It's not really a big deal; as best I can tell, mathematicians prefer the variance because lots of theorems are easier to remember when stated in terms of the variance. (For example, *under certain conditions, the variance of the sum is the sum of the variances* it certainly easier to remember than the same statement in terms of standard deviations. If you doubt what I say, try it!) On the other hand, for the types of work we do in Statistics, the standard deviation makes more sense.

Our approach will be to calculate the variance. Once the variance is obtained, it is just one more step—taking a square root—to obtain the standard deviation. I will introduce you to the computational steps in Table 1.4. Let's begin by looking at the treatment 1 (chicken) data. In the x column I have listed the ten response values. I placed these numbers in the order in which they were obtained, but that is not necessary. If you want to sort them, that is fine. I sum the x's to find their total, 51, and then divide the total by $n_1 = 10$ to obtain their mean, $\bar{x} = 5.1$. Next, I subtract this mean, 5.1, from each observation, giving me the column of deviations, $x - \bar{x}$. As discussed earlier a deviation is positive [negative] if its observation is larger [smaller] than the mean.

For the chicken data, five deviations are negative and five are positive. In terms of magnitude, two deviations are very close to 0 (their magnitudes are both 0.1, for observations 2 and 3); the deviation for observation 5 has the distinction of having the largest magnitude, 4.1. The idea is that we want to summarize these ten deviations to obtain an overall measure of spread within the

chicken treatment. In my experience many people consider it natural to compute the mean or median of the magnitudes of the deviations. Neither of these operations—calculating the mean or median magnitude—is shown in the table because neither turns out not to be particularly useful in our subsequent work. What turns out to be very useful, as we shall see throughout these notes, is to *square each deviation*. The squared deviations appear in the $(x - \bar{x})^2$ column of the table.

We find the total of the squared deviations, which appears in the table as 36.90 for the chicken data.

Now, another strange thing happens. (Squaring the deviations was the first strange thing.) Mathematicians and statisticians disagree on what to do with the total of the squared deviations, again 36.90 for the chicken data. Mathematicians argue in favor of calculating the mean of the squared deviations; i.e., to divide 36.90 by $n_1 = 10$ to obtain 3.690. Statisticians divide by

$$(n_1 - 1) = 9 \text{ to obtain } 36.90/9 = 4.100.$$

In these notes we will follow the lead of statisticians and divide the sum of squared deviations by the sample size minus one. The resultant number is called the **variance of the data** and is denoted by s_1^2 for our x's and s_2^2 for our y's. Let me summarize the above with the following formula.

Definition 1.1 *Suppose that we have m numbers, denoted by*

$$w_1, w_2, \ldots, w_m$$

with mean denoted by $\bar{w}$. The variance of these numbers is denoted by s^2 and is computed as follows:

$$s^2 = \frac{\sum_{i=1}^{m}(w_i - \bar{w})^2}{m-1} = \frac{\sum(w - \bar{w})^2}{m-1} \tag{1.4}$$

In particular, for the data from treatment 1, the variance is denoted by s_1^2 and is computed as follows:

$$s_1^2 = \frac{\sum_{i=1}^{n_1}(x_i - \bar{x})^2}{n_1 - 1} = \frac{\sum(x - \bar{x})^2}{n_1 - 1}. \tag{1.5}$$

For the data from treatment 2, the variance is denoted by s_2^2 and is computed as follows:

$$s_2^2 = \frac{\sum_{i=1}^{n_2}(y_i - \bar{y})^2}{n_2 - 1} = \frac{\sum(y - \bar{y})^2}{n_2 - 1} \tag{1.6}$$

Why do statisticians divide by the sample size minus one? As discussed earlier—when talking about the center of gravity interpretation of the mean—I noted that for any data set the sum of the deviations equals 0. This fact is illustrated for both the chicken and tuna data sets in Table 1.4. Let's focus on the chicken data. Each deviation gives us information about the spread in the data set. Thus, initially, one might think that there are 10 items of information in the 10 deviations. But, in fact, there are only nine items of information because once we know any nine of the deviations the value of the remaining deviation is determined; it equals whatever is needed to make the 10 deviations sum to 0.

Here is a simpler example: suppose that I have $m = 3$ numbers. Two of the deviations are: $+4$ and -7. Given these two deviations, we **know** that the third deviation must be $+3$. A picturesque way of saying this is that for $m = 3$ observations, "Two of the deviations are free to be whatever number they want to be, but the third deviation has no freedom." In other words, the three deviations have two **degrees of freedom**. Thus, in our chicken or tuna data, the ten deviations have nine degrees of freedom. In general, for m observations, the m deviations have $(m - 1)$ degrees of freedom.

Let's return to the question:

> Why do statisticians divide by the sample size minus one?

The answer is: Because statisticians divide by the degrees of freedom. There are many reasons why statisticians divide by the degrees of freedom and we will learn some of them in these notes. I won't, however, introduce new concepts in this chapter simply to explain why,

The variance of the chicken data is 4.100. You may follow the presentation in Table 1.4 and find that the variance of the tuna data is 4.322. This measure of spread is nearly identical for the two data sets; in words, the two data sets have almost exactly the same amount of spread; well, at least as measured by the variance.

As stated earlier, statisticians prefer to take the (positive) square root of the variance and call it the standard deviation. There are three main reasons statisticians prefer using the standard deviation to measure spread rather than the variance.

1. In many of the formulas we will see in these notes, especially those for population-based inference, the standard deviation appears, not the variance.

2. In Chapter 2, I will give you a guide, called the **empirical rule**, which allows us to interpret the value of the standard deviation. There is no such useful guide for interpreting the variance. I am saving this for Chapter 2 in order to keep the size of the current chapter—your first after all—less daunting.

3. The standard deviation gets the *units of the variable* correct; the variance does not. I will explain this below for the data from Dawn's study.

Regarding the third reason above, consider the *unit of the variable* for the study of Bob the cat. (Yes, this is unfortunate language. The unit of the variable is not the units of the study. This is one reason I prefer to call the units of the study either trials or subjects.) Each observation counts the number of cat treats consumed. For example, on the first chicken day, four cat treats were consumed by Bob. The mean for the chicken data is 5.1 *cat treats*. Each deviation is measured in cat treats: the chicken day when Bob consumed 7 treats has a deviation of $7 - 5.1 = 1.9$ cat treats. This day gives a squared deviation of $(1.9)^2 = 3.61$ *cat treats squared*, whatever they are. Thus, the variance for the chicken data is 4.100 cat treats squared. When we take the square root of 4.100 to obtain the standard of 2.025, we also take the square root of *cat treats squared*, giving us a standard deviation of 2.025 cat treats.

1.4 Computing

WARNING: In this course I will direct you to several websites for computing. In my experience, some of these websites do not work for all web servers. My recommendation is to use Firefox, Safari or Chrome. If you have difficulties, contact your instructor for this course.

In this chapter we have seen several tools for presenting and summarizing data: dot plots, means, medians, variances and standard deviations. I have presented these tools as if we perform all the necessary operations by hand. Obviously, we need to reduce the tedium involved in using these tools. Before I discuss the specifics of computing for this chapter, I want to give you a brief overview of computing in this course.

For simple computations, I recommend that you use an inexpensive handheld calculator. For example, I use the calculator on my cellphone; it performs the four basic arithmetic operations and takes square roots. Thus, if you tell me that the variance of a set of data equals 73.82, I pull out my cell phone and find that the standard deviation is $\sqrt{73.82} = 8.592$. Similarly, if you tell me that $x = 5$, $b_0 = 20$ and $b_1 = -1$, I can determine that the value of

$$y = b_0 + b_1 x \text{ is } y = 20 - 1(5) = 20 - 5 = 15.$$

There are literally dozens of *approaches* you could use to perform more involved computations required in this course; five approaches that come to mind are:

1. Using a variety of **websites**.
2. Using the statistical software package **Minitab**.
3. Using some other statistical software package. Of special interest is the open-source software package **R**. (Yes, its name is a single letter.)
4. Using a sophisticated **hand-held calculator**.
5. Using a spreadsheet program, for example **Excel**.

In these notes I will provide guidance on the first of these approaches. I use Minitab extensively to produce output that is not available from any website. If you are interested in learning about Minitab, let me know. No promises as to what we will do, but I would like to know. Neither the TA nor I will provide any guidance on the other three options above or, indeed, any other option you might know. Thus, please do not ask us to do so.

The websites are great because:

- They are free.
- They have some quirks, but, for the most part, require little or no training before they are used.

The websites, however, have two potential problems.

- I **cannot** guarantee that they will remain available because I am not the tsar of the internet.

- Whereas I personally have a great deal of faith in the validity of answers provided by Minitab and R, *I don't really know about these sites*. I have found a serious error in the *Binomial Probabilities* website and will warn you about it when the time comes. I have found some other errors that we can *work around* and will mention them when the time is appropriate. Are there other errors? Who knows? If you find or suspect an error, please let me know.

I have used Minitab in my teaching and research since 1974. Perhaps obviously, I am very satisfied with its performance. Advantages of Minitab include:

- If you do additional course work in Statistics, eventually you will need to learn a statistical software package.
- *Knowledge of Minitab* might be a useful addition to any application for employment. Might; no guarantee.
- If you enjoy programming, Minitab will give you a good understanding of the steps involved in a statistical analysis.

The two main drawbacks to learning Minitab are:

- It is not free. At the time of my typing this chapter, I do not know the price of Minitab for my course. The number I have heard is $30 which would buy you a six-month rental of Minitab.
- Compared to the websites, Minitab requires more time before you can *get started*. In my experience, a great feature of Minitab is this amount of time is much smaller than for any other statistical software package.

Now I will discuss each of the tools mentioned above and how I expect you to make use of them. By the way, in these *Course Notes* I focus on what I expect you to know in order to do the homework and to be successful on the exams. (By *exams* I mean the midterm(s) and final exams.) If you choose to submit project reports—details to be provided—then you *might* need to do some work by hand.

Dot plots. There is a website that will draw a dotplot of a set of data. You won't need to use it in this course, but I include it for completeness. In general, if I want you to see a dotplot, I will provide it for you. The website is:

`http://rossmanchance.com/applets/DotPlotApplet/DotPlotApplet.html`.

If you are interested in this website, I suggest you try it with the chicken data from Dawn's study.

Median. The difficulty lies in taking the set of data and sorting its values. This is no fun by hand, but is easy with a spreadsheet program. Once you have the sorted data, you may use Equation 1.2 or Equation 1.3, depending on whether your sample size is odd or even, respectively. (Both equations are on page 10.) If you don't know how to use a spreadsheet program, no worries; on exams, except possibly for very small data sets, I will give you the sorted list of data.

Mean, variance and standard deviation. As with the median, you may perform the arithmetic by using a spreadsheet program. In particular, if you look again at Table 1.4 on page 13, you can visualize how to create these columns using a spreadsheet. Again as with the median, if you don't know how to use a spreadsheet program, no worries; on exams, I will give you the value of the mean and the value of either the variance or standard deviation. For homework, the computation of the mean, variance and standard deviation can be achieved by using our first website. Go to the site:

`http://vassarstats.net`

On the left side of the page is a blue border, with links in white type. About 75% of the way down the list, click on:

t-Tests & Procedures.

Click on the third of the four paragraphs that appear, **Single Sample t-Test**. You will be taken to a page with a heading **Procedure** followed by a *Data Entry* box. This website is a bit nasty, meaning that you need to be very careful how you enter the data. I entered the chicken data in Table 1.2 into the box and clicked on the calculate box. The website produced quite a collection of statistics, including the following:

- The sample size, 10; the sum of the observations, 51; the sum of squared (SS) deviations, 36.9; the variance, 4.1; the standard deviation, 2.0248; the mean, 5.1; and the degrees of freedom (df), 9.

If you use this website, **you need to be very careful with data entry**.

1. **If you enter your observations by typing:** After typing each observation, hit the enter key; i.e., you may not enter more than one observation per line.

2. **If you enter your observations by 'cutting and pasting:'** You must cut and paste a column of numbers; one number per row, as described above for typing. If you paste a row that includes more than one observation, it won't work.

1.5 Summary

Scientists use comparative studies to investigate certain questions of interest. A comparative study has the following components:

- **Units:** Units are either trials or subjects. The researcher obtains information—the value(s) of one (or more) feature(s)—from each unit in the study.
- **Response:** The feature of primary interest that is obtained from each unit.
- The scientist wants to investigate whether the **level** of a **study factor** influences (strong) or is associated with (weak) the values of the responses given by the units. Almost always in these notes, a comparative study will have two levels.
- Of very great importance to the scientist is the **method** by which units are *identified with* or *assigned to* levels of the study factor.

Regarding this last item, if the method is:

> Units are assigned to levels (or vice versa) by the process of randomization (as described later in Chapter 3)

then the comparative study is called a Completely Randomized Design (CRD). For a CRD the levels of the study factor are called the treatments.

In the first few chapters of these notes the response always will be a number; hence, it is called a numerical response. When a CRD is *performed* or *conducted*, the result is that the researcher has data. Our goal is to discover how to learn from these data.

The data can be displayed in tables, in three ways of interest to us.

1. The table can present the data exactly as collected. An example of this is Table 1.1 on page 6.
2. The data can be presented as above, but separated into groups by treatment. An example of this is Table 1.2 on page 7.
3. The data in the previous table can be sorted, from smallest to largest, within each treatment. An example of this is Table 1.3 on page 7.

It is instructive to draw pictures of the data, one picture for each treatment. The picture we learned about in this chapter is the dot plot. An example of a dot plot is in Figure 1.1 on page 8.

Finally, we learned about four numbers that can be computed to summarize the information in a set of data. They include two measures of center: the mean and the median; and they include two mathematically equivalent measures of variation (spread): the variance and the standard deviation.

There is an exact connection between the dot plot and the mean; namely, the mean is the center of gravity of the dot plot.

Table 1.5: Sorted speeds, in MPH, by time, of 100 cars.

Speeds at 6:00 pm

26	26	27	27	27	27	28	28	28	28	28	28	28	28	28	28	28	28
28	29	29	29	29	29	29	29	29	29	29	29	30	30	30	30	30	30
30	30	31	31	31	31	32	33	33	33	34	34	35	43				

Speeds at 11:00 pm

27	28	30	30	30	31	31	31	32	32	32	32	32	32	32	33	33	33
33	33	33	33	34	34	34	34	34	34	35	35	35	35	36	36	36	37
37	37	37	37	37	38	38	39	39	40	40	40	40	40				

1.6 Practice Problems

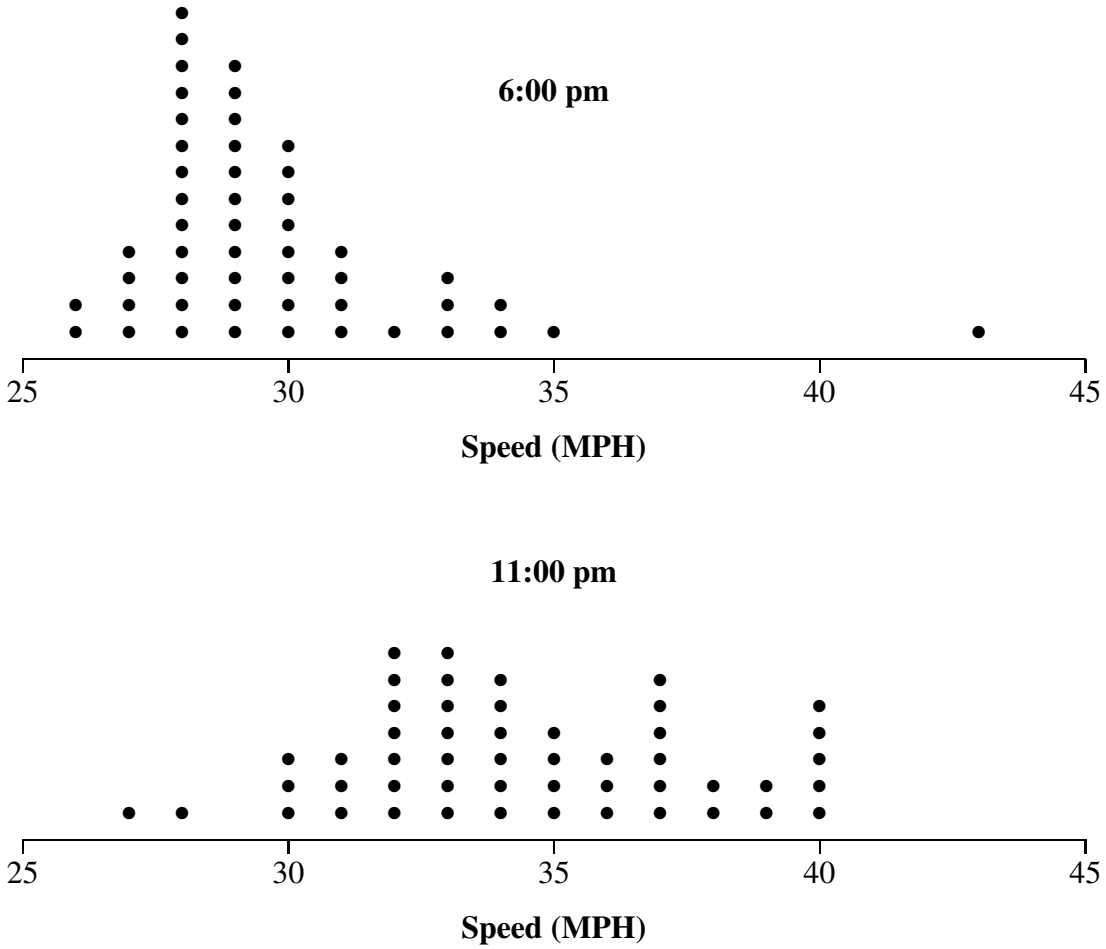
The idea behind this section is to give you an additional example that highlights the many ideas and methods that you need to learn from this chapter.

First, let me describe the data set we will use in this section. On a spring evening, a Milwaukee police officer named Kenny measured the speeds of 100 automobiles. The data were collected on a street in a "warehouse district" with a speed limit of 25 MPH. Fifty cars were measured between roughly 5:45 and 6:15 pm, referred to below as 6:00 pm. The remaining 50 cars were measured between roughly 10:40 and 11:20 pm, referred to below as 11:00 pm.

Each car's speed was measured to the nearest MPH. The sorted data, by time, are in Table 1.5. The dot plots of the speeds, by time, are given in Figure 1.2. These speed data will be used to answer questions 1–6 below.

1. This is a comparative study, but not a CRD. Identify the following components of this study.
   (a) What are the units? Are the units trials or subjects?
   (b) What is the response?
   (c) What is the study factor? What are its levels?
   (d) Explain why this is not a CRD.
2. Look at the two dot plots in Figure 1.2. Write a few sentences that describe what the pictures reveal. You should discuss each picture separately and you should compare them.
3. Calculate the mean, median and standard deviation of the 6:00 PM data.
4. Calculate the mean, median and standard deviation of the 11:00 PM data.
5. Briefly discuss your answers to questions 2–4.
6. We will see repeatedly in these notes that the presence of even one outlier *might* have a big impact on our analysis. Let's explore this topic a bit. Delete the largest observation from the

Figure 1.2: Dot plots of speeds, by time.



6:00 data set and recalculate the mean, median and standard deviation of the remaining 49 observations. Discuss your answers.

1.7 Solutions to Practice Problems

1. (a) The units are the cars driving past the police officer. I think of each car driving past as a trial. If you knew that the 100 cars were driven by 100 different people, you could view the units as subjects. (To paraphrase a well-known national association—channeling Harry Potter, that whose name we do not mention—cars don't speed, drivers speed.) This is an example where either designation—trials or subjects—has merit. It really is not a big deal whether we call call the units trials or subjects.

   (b) The response is the speed of the car, measured to the nearest integer miles per hour.

(c) The study factor is the time of day, with levels 6:00 PM and 11:00 PM.

(d) In my experience, many students find this question to be difficult. Some have said, "Yes, it's a CRD because the cars are driving past at random." This is an example of a very important issue in this class. Randomization has a very specific technical meaning. We must follow the meaning exactly in order to have randomization. Admittedly, I have not told you what randomization is, so you might think I am being unfair; if this were an *exam*, I would be unfair, but this is a *practice problem*. The key point is that in order to have randomization the police office first had to have *control* over when the cars drove past. He had to have a list of the 100 cars (drivers) and say, "You 50, drive past me at 6:00; the remaining 50, you drive past me at 11:00." Clearly, he did not have this control; he *observed* when the cars drove past.

As is rather obvious from the dot plots, cars at the later hour are driven at substantially higher speeds than cars driven at the earlier hour. But—as we will see later and you can perhaps see now—does this mean that a given person tends to drive faster at the later time **or** does this mean that fast drivers come out late at night? You might have a strong feeling as to which of these explanations is better (or you might have some other favorite), but here is my point: The data we have will not answer the question of why. In my earlier language, we *cannot say* that time-of-day *influences* speed; we only *can say* that time-of-day is *associated* with speed.

2. Obviously, there are many possible *good* answers to this question. My answer follows. Don't view this as the *ideal*, but rather try to understand why my comments *make sense* and think about ways to improve my answer.

6:00 PM data: Everybody is driving faster than the speed limit, 25. A substantial majority (32 of 50, if one counts) of the cars are traveling at 28, 29 or 30 MPH. There is not a very much spread in these 50 observations, **except** for the isolated large value of 43. (A large [small] isolated value is called a large [small] **outlier**.)

11:00 PM data: Everybody is driving faster than the speed limit; in fact, all but two drivers exceed the limit by at least 5 MPH. There is a lot of spread in these response values. There are three clear peaks: from 32–34; at 37; and at 40. The peak at 40 is curious; lots of people (well, five) drive 40, but nobody drives faster. Previous students of mine (I do like this example!) have opined that the drivers are trying to avoid a big increase in penalties for being caught driving more than 15 MPH over the speed limit.

Comparing dot plots: The most striking feature is that the speeds are substantially larger at 11:00. Also, there is more spread at 11:00 than at 6:00.

3. I used the website `http://vassarstats.net`, following the method described in Section 1.4. I entered the 6:00 PM data and obtained:

$$\bar{x} = 29.68 \text{ and } s_1 = 2.810.$$

To obtain the median we note that $n_1 = 50$ is an even number. Following Equation 1.3 on page 10, we compute $k = 50/2 = 25$ and $k + 1 = 26$. From Table 1.5, the response 29 is in both positions 25 and 26. Thus, the median $\tilde{x}$ equals 29.

4. I used the website `http://vassarstats.net`, following the method described in Section 1.4. I entered the 11:00 PM data and obtained:

$$\bar{y} = 34.42 \text{ and } s_2 = 3.252.$$

From Table 1.5, the response 34 is in both positions 25 and 26. Thus, the median $\tilde{x}$ equals 34.

5. The mean [median] speed at 11:00 is 4.74 [5.00] MPH larger than the mean [median] speed at 6:00. The differences in these measures of center agree with what we see in the dot plots. The ratio of the standard deviations is $3.252/2.810 = 1.157$. Thus, as measured by the standard deviation, there is almost 16% more spread in the later data.

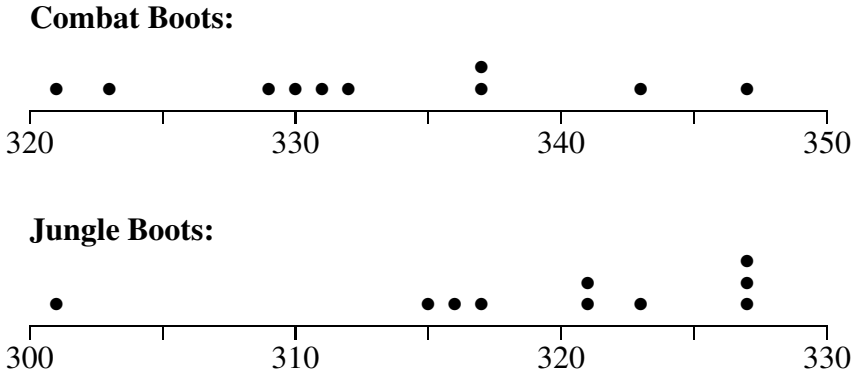
6. With the help of the website, $\bar{x} = 29.408$ and $s_1 = 2.071$. For the median, the sample size is now 49, an odd number. From Equation 1.2 on page 10, we find that $k = (49 + 1)/2 = 25$. The observation in position 25 is 29 and it is the median.

   The deletion of the outlier has left the median unchanged. The mean decreased by $29.68 - 29.41 = 0.27$ MPH; or, if you prefer, the mean decreased by 0.9%. The standard deviation decreased by 26.3%! As we will see repeatedly in these notes, even one outlier can have a huge impact on the standard deviation.

   I *do not advocate* casually discarding data. If you decide to discard data, you should always report this fact along with your reason for doing so.

   Beginning with Chapter 3 we will devote a great deal of effort into learning how to quantify uncertainty, using the language, techniques and results of probability theory. It is important to learn how to quantify uncertainty, *but it is equally important to realize that there are many situations in which we cannot, in any reasonable way, quantify uncertainty*. Often we just need to accept that our answers are uncertain. In the current case, there is uncertainty about who drives down a street on any given night. I don't know why the driver responsible for the large outlier at 43 MPH decided to drive down the street being studied when Kenny was collecting data. But it's certainly possible that he/she could have chosen a different route or a different time. Thus, I think it is interesting to see what would happen to our analysis if one of the subjects/trials had *not* been included in the study.

Figure 1.3: The dot plots of Brian's running times, by type of boots.



1.8 Homework Problems

Brian performed a balanced Completely Randomized Design with 20 trials. His response is the time, measured to the nearest second, he needed to run one mile. Wearing combat boots, his sorted times were:

321 323 329 330 331 332 337 337 343 347

Wearing jungle boots, Brian's sorted times were:

301 315 316 317 321 321 323 327 327 327

Figure 1.3 presents the two dot plots for Brian's study. Use Brian's data to solve problems 1–4.

1. Calculate the mean, median and standard deviation for the combat boots data.

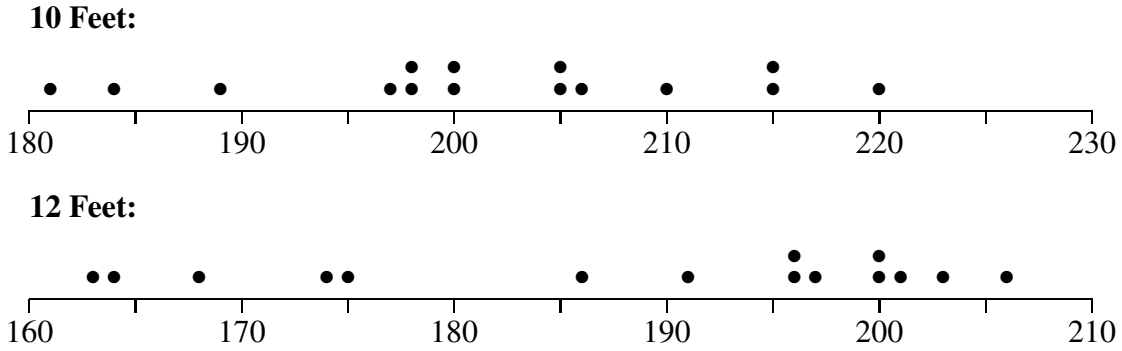
2. Calculate the mean, median and standard deviation for the jungle boots data.

3. Recalculate the mean, median and standard deviation for the jungle boots data after you delete the small outlier (leaving a set of nine observations).

4. Write a few sentences to explain what you have learned from your answers to problems 1–3 as well as an examination of these dot plots.

   Note: There is no unique correct answer to this problem. I don't put questions like this on my exams—grading such questions is very subjective and I try to avoid such grading issues. Also, I don't want you to feel you are at a disadvantage compared to the other students in this class who, surprisingly, are all majoring in military footware. Seriously, I try to avoid grading you based on your scientific knowledge in any particular field that I choose to present. Answering this question is, however, good practice for your projects. In a project, you choose the topic; if you choose a topic for which you have no knowledge, no interest and no aptitude, then your grade will suffer!

Figure 1.4: The dot plots of Reggie's dart scores, by his distance from board.



Reggie performed a balanced CRD of 30 trials. Each trial consisted of a game of darts, where a game is defined as throwing 12 darts. Treatment 1 [2] was throwing darts from a distance of 10 [12] feet. Reggie's response is the total of the points obtained on his 12 throws and is called his score, with larger numbers better. Below are Reggie's sorted scores from 10 feet:

181 184 189 197 198 198 200 200 205 205 206 210 215 215 220

Below are Reggie's sorted scores from 12 feet:

163 164 168 174 175 186 191 196 196 197 200 200 201 203 206

Reggie's two dot plots are presented in Figure 1.4. Use Reggie's data to answer problems 5–7.

5. Calculate the mean, median and standard deviation for the scores from 10 feet.

6. Calculate the mean, median and standard deviation for the scores from 12 feet.

7. Write a few sentences to explain what you have learned from your answers to problems 5 and 6 as well as an examination of Reggie's dot plots.

   Keep in mind my comments in problem 4 above.

Chapter 2

The CRD with a Numerical Response: Continued

This chapter continues the theme of Chapter 1. I begin with another example of a student project.

2.1 Kymn the Rower

Kymn was a member of the women's varsity crew at the University of Wisconsin-Madison. When she could not practice on a lake, she would work out on a rowing simulation device called an ergometer. One does not simply sit down at an ergometer and begin to row. It is necessary to choose the *setting* for the machine. There are four possible settings, obtained by combining two dichotomies:

- One can opt for the *small gear* setting or the *large gear* setting.
- One can choose to have the vent *open* or *closed*.

Kymn decided that she was not interested in two of these settings: the large gear with the vent closed would be too easy and the small gear with the vent open would too difficult for a useful workout. As a result, Kymn wanted to compare the following two settings:

- **Treatment 1:** The small gear with the vent closed, and
- **Treatment 2:** The large gear with the vent open.

For her response, Kymn chose the time, measured to the nearest second, she required to row the equivalent of 2000 meters.

In the above, I have implicitly defined Kymn's trial as sitting on the erg and rowing the equivalent of 2000 meters. Kymn decided to perform a total of 10 trials in her study.

Kymn's data are in Table 2.1, with dot plots in Figure 2.1. Look at these data for a few minutes. What do you see? Below are some features that I will note.

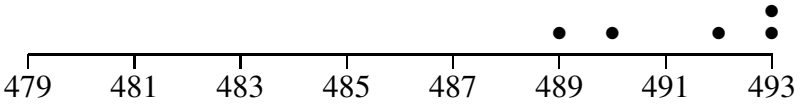
1. Every response on treatment 2 is smaller than every response on treatment 1.

Table 2.1: Kymn's times, in seconds, to row 2000 meters on an ergometer. Treatment 1 is the small gear with the vent closed; and treatment 2 is the large gear with the vent open.

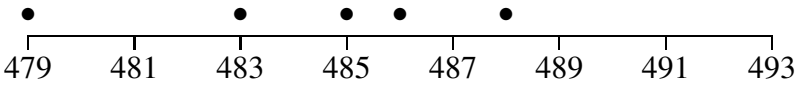
Trial:	1	2	3	4	5	6	7	8	9	10
Treatment:	2	1	1	1	2	2	1	2	2	1
Response:	485	493	489	492	483	488	490	479	486	493

Figure 2.1: The dot plots for Kymn's rowing study.

Treatment 1: Small Gear, Vent Closed:



Treatment 2: Large Gear, Vent Open:



2. The variation in treatment 2 is larger than the variation in treatment 1. Having noted this fact, in both treatments there is very little within-treatment variation. It is impressive, yet perhaps unsurprising for a well-conditioned athlete, that in response times of slightly more than 8 minutes, there is so little variation in trial-to-trial performance.

If one looks at the dot plots, and remembers the center of gravity interpretation of the mean, one can see that the mean on treatment 1 is a bit larger than 491 seconds and that the mean on treatment 2 is a bit smaller than 485 seconds; these visual conclusions are supported by computation. In particular, for future reference note that the means, medians and standard deviations of these data are:

$$\bar{x} = 491.4, \tilde{x} = 492, s_1 = 1.817, \bar{y} = 484.2, \tilde{y} = 485 \text{ and } s_2 = 3.420.$$

2.2 Sara's Golf Study; Histograms

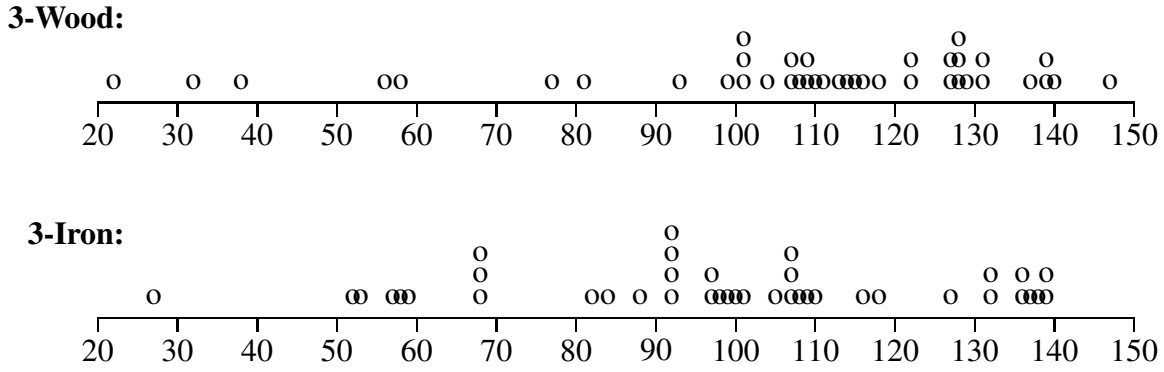
Sara performed a balanced CRD with 80 trials. Her response was the distance—in yards—that she hit a golf ball at a driving range. (She hit the ball into a net which displayed how far the ball would have traveled in *real life*. I have no idea how accurate these devices are.) Sara had two treatments: hitting the ball with a 3-Wood (treatment 1) and hitting the ball with a 3-Iron (treatment 2). If you don't know much about golf, don't worry; all that matters is that Sara wanted to compare two clubs with particular interest in learning which would lead to a larger response.

Table 2.2: The distance Sara hit a golf ball, in yards, sorted by treatment.

3-Wood									
22	32	38	56	58	77	81	93	99	101
101	101	104	107	107	108	109	109	110	111
113	114	115	116	118	122	122	127	127	128
128	128	129	131	131	137	139	139	140	147

3-Iron									
27	52	53	57	58	59	68	68	68	82
84	88	92	92	92	92	97	97	98	99
100	101	105	107	107	107	108	109	110	116
118	127	132	132	136	136	137	138	139	139

Figure 2.2: The dot plots for Sara's golf study.



Sara's data, sorted by treatment, are presented in Table 2.2. Even a cursory examination of this table reveals that, within each treatment, there is a huge amount of variation in Sara's responses. Dot plots of Sara's data are presented in Figure 2.2.

I don't like these dot plots very much, but let me begin by mentioning their good features. As with all dot plots, each plot is a valid presentation of its observations. If you want to see the *exact* values of all of the observations and how they relate spatially, the dot plot is great. In addition, a dot plot is good at revealing outliers: we can see the three very small response values with the 3-Wood and the one very small value with the 3-Iron. Now I will discuss, briefly, what I don't like about these dot plots.

The 3-Wood data range from a minimum of 22 yards to a maximum of 147 yards. This distance, 125 yards, towers over the number of observations, 40. As a result, there must be, and are, a large number of **gaps** in our picture and *usually* (there are weird exceptions) with so little data spread out so far, the peaks are very short and, hence, likely have no scientific meaning. There is another way to view the above comments: the dot plot is very bumpy; i.e., it is not very *smooth*. As I will

Table 2.3: Frequency tables of the distances Sara hit a golf ball, by treatment.

		3-Wood			**3-Iron**		
Class Interval	Width (w)	Freq. (f)	Rel. Freq. ($r_f = f/n_1$)	Density ($d = r_f/w$)	Freq. (f)	Rel. Freq. ($r_f = f/n_2$)	Density ($d = r_f/w$)
0–25	25	1	0.025	0.001	0	0.000	0.000
25–50	25	2	0.050	0.002	1	0.025	0.001
50–75	25	2	0.050	0.002	8	0.200	0.008
75–100	25	4	0.100	0.004	11	0.275	0.011
100–125	25	18	0.450	0.018	11	0.275	0.011
125–150	25	13	0.325	0.013	9	0.225	0.009
Total	—	40 (n_1)	1.000	—	40 (n_2)	1.000	—

discuss later in the subsection on *kernel densities*, smoothness is very important to scientists.

Here is what I mean by bumpy. Imagine the number line is a road and the dots are bumps in the road. Driving a car (or if you prefer a greener example, riding a bike) along the road will result in a flat road (the gaps) interrupted by numerous bumps.

Finally, it is very difficult to see a shape in either of these dot plots. This is disturbing because the quest for a shape is one of the main reasons that scientists draw pictures of data.

Admittedly, all of the dot plots we have seen in these notes have been bumpy. Most of our dot plots have not had a recognizable shape, but that is to be expected with a small amount of data, as we had in Dawn's and Kymn's studies in our exposition, as well as Brian's and Reggie's studies in the homework to Chapter 1. Arguably, policeman Kenny's dot plots (see Figure 1.2 on page 21 in the Chapter 1 Practice Problems), based on a large number of observations and a small range of values, did reveal shapes. Below we will introduce histograms, which are smoother (statisticians often prefer this more *positive* term to *less bumpy*) than dot plots and usually reveal shape better. Finally, I will introduce you briefly to kernel density estimates that are, in the sense we will learn, better than histograms both on smoothness and revealing shapes.

In the excellent movie *Amadeus* (1984), a dramatization of the life of composer Wolfgang Amadeus Mozart (1756–1791), a jealous competitor derides one of Mozart's works as having *too many notes*. In a similar spirit, one can criticize a dot plot for having *too much detail*. Our next picture, the **histogram** sacrifices some of the detail of a dot plot. The reward, one hopes, is a better, more useful, picture.

The first thing to note is that to refer to *the histogram* for a set of data is a bit misleading. The definite article—the—is inappropriate because *many* histograms can be drawn for any set of data, for two reasons. First, as we will see below, a histogram is dependent on our choice of *class intervals*, and there are always many possible choices for these. Second, for a given choice of class intervals, there are three possible histograms: the frequency histogram, the relative frequency histogram and the density histogram.

The first step in creating a histogram is to create a frequency table. Table 2.3 presents fre-

quency tables for both treatments for Sara's data. Let me carefully explain these tables. The first column presents my choices for the class intervals. Because I am very interested in comparing the responses on the two treatments, I am using the same class intervals for both tables. This isn't necessary, but I do think it's a good idea.

In this course—exams and these notes, including homework—I will always give you the class intervals. (If you perform a project that requires a frequency table, then you will need to choose the class intervals.) Our class intervals will always follow the rules listed below. (Thus, if you need class intervals for your project, please follow these rules.) As you no doubt have already surmised, these rules do not rate high on the excitement-o-meter, but they are necessary. And there is a really annoying feature: There are two other versions of these rules—far inferior to the rules below—each of which appears in many textbooks of introductory Statistics. (Don't let the adjective *many* dismay you; there are hundreds, if not thousands, of introductory Statistics texts and nearly all should be avoided. But that's another topic.)

A valid collection of class intervals will satisfy the following five rules.

1. Each class interval has two endpoints: a lower bound(ary) and an upper bound(ary). As in Table 2.3, when one reads down the first column, the lower (and upper) bounds increase.

2. The smallest class interval boundary must be *less than or equal to* all observations in the data set.

3. The largest class interval boundary must be *greater than or equal to* all observations in the data set.

4. The upper bound of a class interval must *equal exactly* the lower bound of the next class interval.

5. Because adjacent class intervals have an endpoint in common, we need the following **endpoint convention**:

   > When determining the frequencies of the class intervals (see below), each interval includes its left endpoint, but not its right endpoint.

   There is one exception to this endpoint convention: The last class interval includes both of its endpoints.

Let me say a few more words about our fourth rule. In Table 2.3, the first two class intervals are 0–25 and 25–50. There are two ways that the fourth rule could be violated; here are examples of each:

- If these intervals were changed to, say, 0–25 and 30–50, then there would be a *gap* between the intervals.

- If these intervals were changed to, say, 0–25 and 20–50, then these intervals would *overlap*.

We allow neither gaps nor overlap; either of these features would ruin our histograms.

We have spent a lot of effort on the class intervals! Let's return to Table 2.3 and examine its remaining columns.

The second column presents the width (w) of each class interval. The width of a class interval is the distance between its endpoints. For example, the first class interval, 0–25, has $w = 25-0 = 25$, as printed. Note that in this table, all class intervals have the same width. There are reasons that a researcher may prefer to have variable-width class intervals (see Practice Problems below), but if one chooses to have variable-width class intervals, then one must use the density histogram because both of the other histograms are *misleading* (again, see Practice Problems below). *Misleading* is perhaps a bit strong. Statisticians agree that they are misleading, but, for all I know, you might be a person who is *never* misled by a picture.

The frequency counts (f) are pretty evident. For example, in the 3-Wood data set, four observations—77, 81, 93 and 99—fall in the interval 75–100; hence, the frequency of this interval is 4. An interval's relative frequency (r_f) is obtained by dividing its frequency by the number of observations in the data set. Thus, for example, the relative frequency for the 75–100 class interval in the 3-Wood data is $r_f = 4/40 = 0.100$. Finally, an interval's density (d) is obtained by dividing its relative frequency by its width.

Let me give you an example in which the endpoint convention in the fifth rule above comes into play. In Sara's 3-Iron data set, the observation 100 is counted as a member of the interval 100–125, not the interval 75–100.

I will now use the frequency tables for Sara's data to draw frequency histograms. These histograms are presented in Figure 2.3. First, I will discuss how to draw a frequency histogram by hand. Second, I will discuss the information revealed by Sara's frequency histograms. Finally, I will discuss Sara's relative frequency and density histograms.

Drawing a Frequency Histogram. We proceed as follows.

1. Draw a portion of the number line and locate the various class interval boundaries on it.

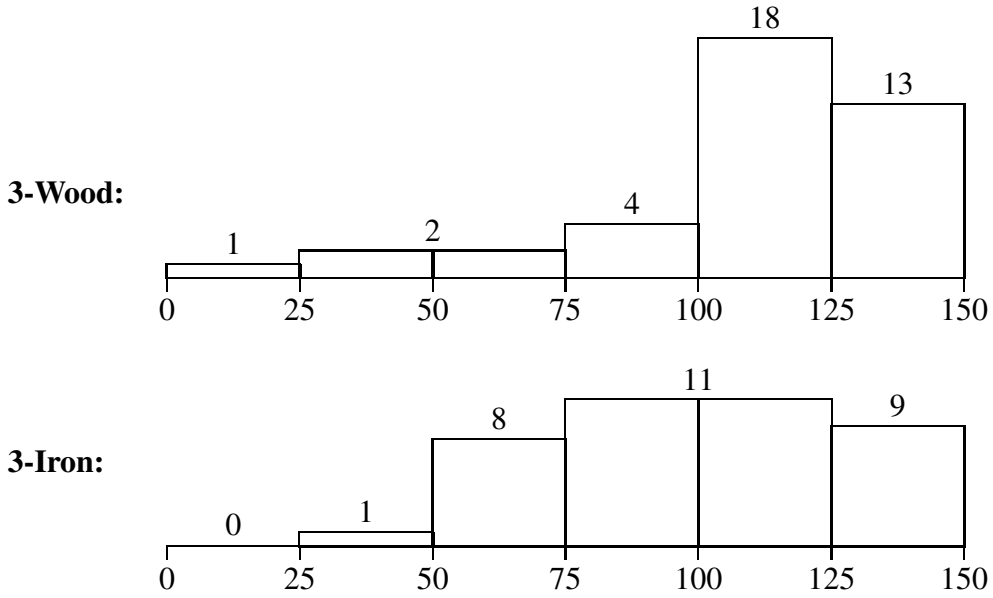
2. Above each class interval draw a rectangle whose height is equal to the frequency of the class interval.

What do we learn by inspecting a frequency histogram? Whenever we have histograms of data from two groups we can look to see if the groups differ substantially in centers and/or spreads. For Sara's data, we see that she definitely hit the ball farther with the 3-Wood than with the 3-Iron. This conclusion is supported by computing means and medians for her data. I also include the standard deviations below.

$$\bar{x} = 106.875, \tilde{x} = 112.0, s_1 = 29.87, \bar{y} = 98.175, \tilde{y} = 99.5 \text{ and } s_2 = 28.33.$$

Statisticians and scientists are particularly interested in assessing the **shape** of a histogram, although this is a very inexact activity. Below are my comments on Sara's two histograms:

Figure 2.3: Frequency histograms for Sara's golf study.



1. **3-Wood Histogram:** There is one tallest rectangle, above the class interval 100–125 yards. Thus, this is the most popular class interval for Sara's 3-Wood responses. The rectangle(s) to the right [left] of this peak rectangle is called the **right tail** [**left tail**] of the histogram. The left tail is much longer than the right tail (100 yards versus 25), but the right tail is heavier (13 observations versus 9). Because of the longer left tail, we label this histogram **skewed to the left**.

2. **3-Iron Histogram:** This histogram exhibits almost perfect left-to-right symmetry.

Note the following facts:

- The 3-Wood histogram is skewed to the left and its mean is smaller than its median: $106.875 < 112.0$.

- The 3-Iron histogram is approximately symmetric and its mean and median are approximately equal: $98.175 \approx 99.5$.

These are two examples of the following famous *Result that is not quite a Theorem*.

Result 2.1 *The following are* usually *true:*

- *If the dot plot or histogram of a set of data is approximately symmetric, then its mean and median are approximately equal.*

- *If the dot plot or histogram of a set of data is clearly skewed to the* right*, then its mean is larger than its median.*

- *If the dot plot or histogram of a set of data is clearly skewed to the* left*, then its mean is smaller than its median.*

Let's apply this result to some of our data sets.

Brian's Running Data. In the Chapter 1 homework, we learned about Brian's study of running. Dot plots of his two data sets are in Figure 1.3 on page 24. Brian's combat boots data look neither symmetric nor skewed to me. The mean, 333.0, is very similar to the median, 331.5. Brian's jungle boots data look strongly skewed to the left, but the mean, 319.5, is only somewhat smaller than the median, 321.0.

Brian's data sets help me to illustrate a common misconception about Result 2.1. For example, in my experience, sometimes a person calculates the mean and median of a data set and finds that they are equal or similar in value. The person then asserts, *without drawing a picture—dot plot or histogram or any other picture—of the data*, that the distribution of the data is symmetric. This can be wrong as illustrated by both of Brian's data sets. (We can also find data sets for which the mean is larger [smaller] than the median but the distribution of the data would not be described as skewed to the right [left].)

The message is: Do not confuse an *if . . . then* result with an *if and only if* result. In math, all definitions and some results (theorems) are if and only if. Many results, in math or other disciplines, are if . . . then results. For example, at the time of my typing these words, the following is a true statement.

> If a person is or has been the President of the United States, then the person is male.

The reverse is not true. I—and millions of other men—have never been the President of the United States.

Reggie and Darts. In the Chapter 1 homework, we learned about Reggie's study of darts. Dot plots of his two data sets are in Figure 1.4 on page 25. Reggie's data from 10 feet look approximately symmetric to me. The mean, 201.53, is very similar to the median, 200.0. Reggie's data from 12 feet look strongly skewed to the left. In agreement with Result 2.1, the mean, 188.0, is smaller than the median, 196.0.

I conclude that Result 2.1 is accurate for Reggie's data.

Kenny and Fast Cars. In the Chapter 1 practice problems, we learned about Kenny's study of car speeds. Dot plots of his two data sets are presented in Figure 1.2 on page 21. Kenny's 6:00 data are strongly skewed to the right with a large outlier, but the mean, 29.68, is only a bit larger than the median, 29.0. Kenny's 11:00 data are not approximately symmetric, yet the mean, 34.42, is very similar to the median, 34.0.

I conclude that Result 2.1 is not very accurate for Kenny's data.

Relative Frequency and Density Histograms. In the above, I stated that for a frequency histogram the height of a rectangle is equal to the frequency of its class interval. As you might guess or already know, a relative frequency histogram differs from a frequency histogram in only one way: The height of any of its rectangles is equal to the relative frequency of the corresponding class interval. For Sara's data, this involves taking her frequency rectangles and dividing each height by 40, in order to convert to relative frequencies. Even if you have not seen the movie, *Honey, I Shrunk the Kids*, you likely realize that this shrinkage of each rectangle (in going from frequency to relative frequency) has no impact on the shape of the histogram. Thus, in terms of shape, it does not matter which of these two histograms we use.

Similarly, a density histogram differs from the previous two histograms only in terms of the heights of its rectangles: The height of any of its rectangles is equal to the density of the corresponding class interval. For Sara's data, this involves taking her relative frequency rectangles and dividing each height by the constant width, 25, in order to convert to densities. Again, as in the movie, *Honey, I Shrunk the Kids*, this shrinkage (because $w > 1$) has no impact on the shape of the histogram. Thus, in terms of shape, for histograms with constant-width class intervals, it does not matter which of these three histograms we use. Please note the following items:

1. As we will see in the Practice Problems, if the class intervals do not have constant widths, you should not use the frequency or relative frequency histogram. They will have the same misleading shape. In this situation, the density histogram will have a different shape and should be used.

2. If we have constant-width class intervals and if $w < 1$, then the densities are larger than the relative frequencies, but all three histograms still have the same shape. (See the totally unnecessary sequel *Honey, I Blew Up the Kid*.)

3. In a frequency or relative frequency histogram we look at the *height* of a rectangle. The height reveals either *how many* or *what proportion* of observations are in the class interval, depending on the histogram. For a density histogram, one should look at the *area* of a rectangle, not its height. In particular, for a density histogram, the area of a rectangle is the relative frequency of the class interval:

$$\text{Area} \;=\; \text{Base} \;\times\; \text{Height} \;= w(r_f/w) = r_f.$$

4. In view of the previous item, we see that the total area of a density histogram equals 1.

In view of the above, which of the three histograms is best? Or does it matter?

1. If one chooses to have variable-width class intervals, the density histogram must be used.

2. If for theoretical or other reasons—see later developments in these notes—one wants the total area of the picture to equal one, the density histogram must be used.

3. If neither of the above apply, then all three histograms give the same picture. In this situation, I avoid the extra work involved in constructing the density histogram. So, how do I choose between frequency and relative frequency histograms?

(a) If the number of observations is small, I prefer frequency: with, say, 10 observations I prefer to say, "Four of the observations are ..." rather than "Forty percent of the observations are ...."

(b) If the number of observations is large, I prefer relative frequency: with, say, 16,492 observations I prefer to say, "Twenty-five percent of the observations are ..." rather than "Four thousand one hundred twenty-three of the observations are ...."

(c) If I am comparing two data sets, as we always do in a comparative study, and the sample sizes n_1 and n_2 are substantially different, then I prefer relative frequency histograms over frequency histograms.

2.2.1 Kernel Densities

Let us return to my earlier discussion of Sara's dot plots being bumpy. Look again at her histograms in Figure 2.3 and remember my *road analogy* on page 30. Here each *road* has long expanses that are flat, making it much smoother than its corresponding dot plot. Unfortunately, these flat, smooth roads result in the careless travel periodically hitting a wall or going over a cliff! (Well, periodically, provided such encounters prove to be neither incapacitating nor sufficiently discouraging to end one's journey.) Clearly, these roads require warning signs and elevators (lifts in the U.K.)! A solution to these dangers is provided by what I will simply call **kernel density histograms**, or **kernel densities**, for short. Later in these notes you will learn that a better name is **kernel density histogram estimates**. But right now we don't know what estimates are and we certainly don't know what kernels are.

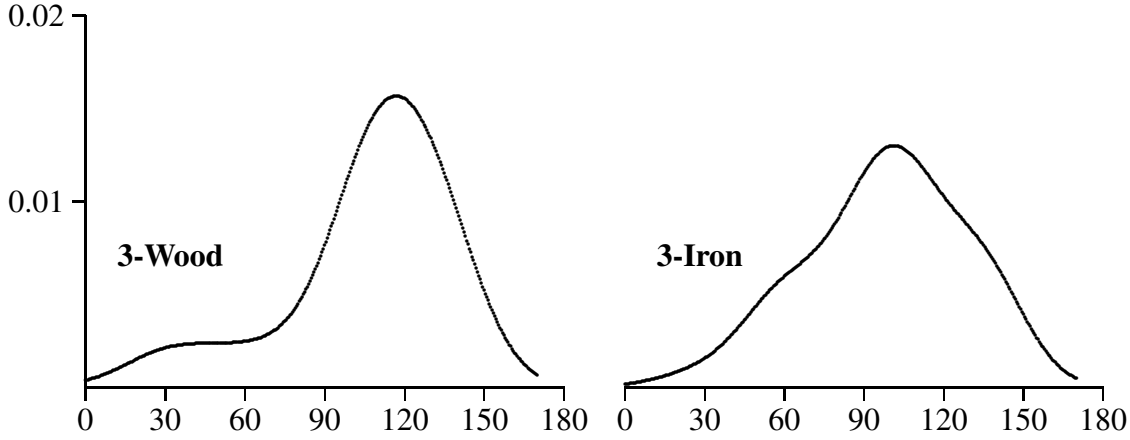
Kernels are fairly easy to explain—though not until we develop several more ideas. They are somewhat difficult to implement and require some careful computations. Software packages exist that will perform the computations for you, but we won't be covering them in this course. I use the statistical software package Minitab to create all of the kernel densities in this course. For our purposes, a kernel density provides a *smoother picture* of our data than either a dot plot or a histogram. In my experience, kernel densities frequently appear in online articles and in published reports—books, journals and magazines. As a result, even though I won't teach you how to construct a kernel estimate, there is value in my introducing you to the idea of one.

Figure 2.4 provides kernel densities for both of Sara's data sets. I want to make several comments on these pictures.

1. For a given set of data, there is not **the** kernel density, there are many. Think of kernel densities as being a range of possibilities between two extremes: the dot plot which is very bumpy and a histogram that has one class interval for all data which, of course, will be one rectangle and, hence, smooth. The kernel densities I have plotted are in some sense the best kernel densities. If the idea of best interests you, read the next item; if not, you may ignore the next item.

2. If you want more information on kernel densities, see the Wikipedia entry for *kernel density estimation*. Following the terminology in Wikipedia, I chose the Normal kernel (also called the Gaussian kernel) with bandwidth $h = 15$ for both pictures. This choice of bandwidth is

Figure 2.4: Kernel densities for Sara's 3-Wood and 3-Iron data.



close to the values (which are similar, but slightly different for the two data sets) given under the heading *Practical estimation of the bandwidth*.

3. The 3-Wood kernel density is skewed to the left, in agreement with my histogram. The 3-Iron kernel density is approximately symmetric, again in agreement with my histogram.

4. Given the small amount of time we will spend on this topic, I don't want you to be concerned about too many issues. Essentially, kernel estimates are good because they give us a smooth picture of the data. They can also be used to calculate areas; hence, my inclusion of a vertical scale on these pictures. The area under a kernel density equals one, which explains why they include the word density in their name.

5. Kernel densities are a reasonable descriptive tool provided *the response is a measurement*. They should not be used if the response is a count. Thus, I would definitely avoid creating a kernel density for Dawn's study of Bob the cat. Some statisticians relax this directive if the count variable exhibits a large amount of variation. For example, some statisticians might use a kernel density to describe Reggie's dart scores. Sadly, we can't spend additional time on this subject.

2.3 Interpreting the Standard Deviation

Please excuse a slight digression. Years ago at a conference on Statistics education, I heard a wonderful talk about students' understanding of the standard deviation. The speaker had interviewed her students approximately one month after the end of her one semester course on introductory Statistics. She found that very few students—even among the students who earned an A in her course—could adequately explain the meaning of the standard deviation. I really admired the speaker for talking about something that many (most? all?) of us teachers suspect: There is

something about the standard deviation that our students *just don't get*. I conclude that teachers—including me—need to improve our explanations of the standard deviation. This section is my attempt to do better than I have in the past.

Note: When we have a comparative study, for example a CRD, we have two standard deviations—one for each set of data—and we distinguish them with subscripts. Similarly, we distinguish our two means by using different letters of the alphabet, x and y. When I discuss means and standard deviations in general terms, my data are represented by

$$w_1, w_2, \ldots, w_m, \text{ with mean } \bar{w} \text{ and standard deviation } s.$$

I hope that this won't be confusing.

Let's revisit what we know. We have decided to measure spread by looking at how much each observation differs from our arithmetic summary, the mean. The discrepancy between an observation and the mean is called the deviation of the observation. A deviation can be negative, zero or positive. The sign of a deviation tells us whether its observation is smaller than—if negative—or greater than—if positive—the mean. The magnitude (absolute value) of a deviation tells us how far the observation is from the mean, regardless of direction. Our goal is to find a way to summarize all these deviations with one number which will be our measure of the spread in the data.

It is a waste of time to summarize deviations by calculating their mean because for every data set the mean deviation is 0. It seems to make sense to summarize deviations by computing the mean of the magnitudes, but, alas, this summary is of no use in Statistics. Instead, we do something very strange. Something I never saw in all my years of studying math until I took my first Statistics course.

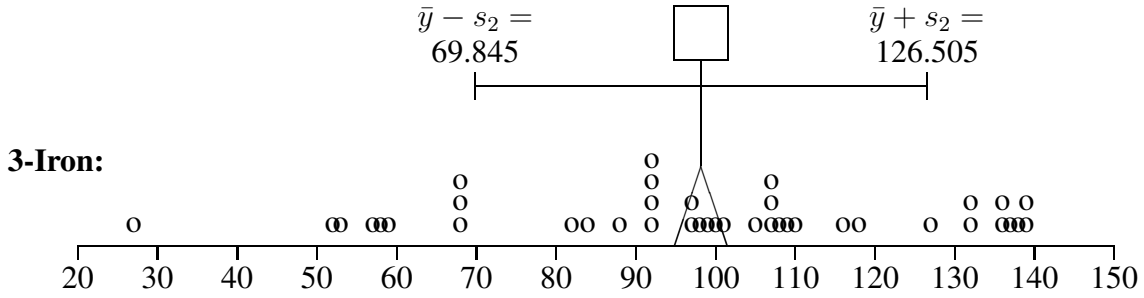
- We square each deviation (equivalently, square each magnitude);
- We compute *almost the mean* (remember: we divide by degrees of freedom, not sample size) of the squared deviations, calling the resulting number the variance;
- We compensate for having squared the deviations by next taking the square root of the variance and call the result the standard deviation.

Now, consider the name: *standard deviation*. Why this name? Well, the *deviation* part makes sense; the summary we obtain is function of the set of deviations. In my experience, it's the word *standard* that befuddles people. Why the modifier standard? I actually don't know. My guess is that we say standard because, as we will see repeatedly in these notes, the standard deviation is essential for the process of *standardizing*. We will see that *standardizing* is very useful. But, for me, this is a chicken versus egg situation: my guess is that the idea of *standardizing* is more basic and, hence, a key number in its process is called the *standard* deviation. But, I might have this backwards; or the truth might be something else entirely. If the proper person sees these notes, perhaps he/she will tell me the answer and I can improve this presentation!

Let us agree to accept that, perhaps, standard deviation is a strange name for s, and let's proceed to learning what it means. For example, I stated earlier that for Sara's 3-Iron data,

$$\bar{y} = 98.175 \text{ and } s_2 = 28.33.$$

Figure 2.5: Elastic Man capturing approximately 68% of Sara's 3-Iron data.



(Recall that the 3-Iron was Sara's treatment 2; thus, we use y's for the data and a subscript of 2 on the s.) How do we interpret the value 28.33 for s_2? First, recall that we have an exact interpretation of the value 98.175 for $\bar{y}$: namely, 98.175 is *exactly* the center of gravity of the dot plot of the data. Our interpretation of s_2 is weaker in that it is not exact, it is only an approximation. To make matters worse, sometimes it's a bad approximation. One positive note: if we have access to a picture of the distribution of the data, then we will know whether the approximation is bad and how it is bad. (See the Practice Problems.)

Our approximation is given in the result below. I recommend that you quickly skim this result and read my motivation which follows it.

Result 2.2 The Empirical Rule for interpreting the value of the standard deviation. *Suppose that we have a set of data. Denote its mean by $\bar{w}$ and its standard deviation by s. The three approximations below collectively are referred to as the Empirical Rule.*

1. *Approximately 68% of the observations lie in the closed interval $[\bar{w} - s, \bar{w} + s]$,*
2. *Approximately 95% of the observations lie in the closed interval $[\bar{w} - 2s, \bar{w} + 2s]$, and*
3. *Approximately 99.7% of the observations lie in the closed interval $[\bar{w} - 3s, \bar{w} + 3s]$.*

Here is the idea behind the *Empirical Rule*. The superhero *Elastic Man* has the ability to stretch his arms as much as he desires. He is standing on the number line at the mean of the data. He poses the following question to himself:

How far do I need to stretch my arms in order to encompass 68% of Sara's 3-Iron data?

The first approximation in the Empirical Rule answers this question; it tells Elastic Man to stretch enough so that one hand is at $(\bar{y} - s_2)$ and the other hand is at $(\bar{y} + s_2)$. This activity is pictured in Figure 2.5. This picture is a bit busy, so let me spend a few minutes explaining it. Elastic Man (a.k.a. *square-headed man*) is standing above the mean of the data, at 98.175. His hands extend from

$$\bar{y} - s_2 = 98.175 - 28.33 = 69.845 \text{ to } \bar{y} + s_2 = 126.505.$$

Table 2.4: The performance of the Empirical Rule for Sara's golf data. The values in this table are the number (percentage) of observations, out of 40, in each interval. Remember that the mean and standard deviation (SD) are $\bar{x} = 106.175$ [$\bar{y} = 98.175$] and s_1 [s_2] for the 3-Wood [3-Iron] data.

	Interval:		
Data Set	Mean $\pm$ SD	Mean ± 2 SD	Mean ± 3 SD
3-Wood	29 (72.5)	37 (92.5)	40 (100)
3-Iron	22 (55.0)	39 (97.5)	40 (100)

According to the Empirical Rule, his reach encompasses approximately 68% of the data. Let's see whether the Empirical Rule is accurate. If you look at Sara's 3-Iron data in Table 2.2 on page 29, you will see that nine observations are smaller than 69.845 and nine observations are larger than 126.505. Thus, in actuality, $40 - (9 + 9) = 22$ observations lie within the reach of Elastic Man. Sadly, $22/40 = 0.55 = 55\%$. The Empirical Rule's 68% is a poor approximation of 55%. But is it really that bad? Looking at the list of observations again, we see that Elastic Man barely misses three observations at 68 yards and one observation at 127 yards. Add these four observations to the previous total of 22 and we get 26 of 40 observations, which is 65% of the 40 observations and *is close* to the Empirical Rule's approximation of 68%. Thus, the approximation fails for these data because Elastic Man needs to stretch a bit farther than s_2 yards in each direction in order to encompass approximately 68% of the data.

I did additional arithmetic and counting to investigate the Empirical Rule's performance for both sets of Sara's data. The results are in Table 2.4. My recommendation: Don't bother checking these numbers; you will get your chance to create such a table in a Homework problem. The Empirical Rule states that the calculated intervals will encompass approximately 68%, 95% and 99.7% of the data. For five of the intervals the approximations are good and I have already discussed the other interval.

Actually, the Empirical Rule tends to work better for larger amounts of data; 40 observations really aren't many in this setting. But even with thousands of observations, there are situations in which the Empirical Rule, gives one or more poor approximations. The first interval, mean $\pm$ SD, is particularly problematic, just as it was for Sara's 3-Iron data. This topic is *not central* to our development in these notes; thus, I will save further examples to the Practice Problems and Homework.

2.4 Cathy's Running Study

I end this chapter with a very small balanced CRD. This will serve as a simple, yet real, example for several ideas presented later in these notes.

Cathy was a very busy student, wife and mother enrolled in my class. One of her favorite *escapes* was to run one mile. She had two routes that she ran: one through a park and one at her local high school. She decided to use her project assignment to compare her two routes. In

Table 2.5: Cathy's times, in seconds, to run one mile. HS means she ran at the high school and P means she ran through the park.

Trial:	1	2	3	4	5	6
Location:	HS	HS	P	P	HS	P
Time:	530	521	528	520	539	527

particular, Cathy performed a balanced CRD with six trials. A trial consisted of Cathy running one mile and the response was the time, measured to the nearest second, required for Cathy to complete her run. Her treatments were: running at the high school (treatment 1) and running through the park (treatment 2). She assigned trials to treatments by randomization. Her data are presented in Table 2.5. Below are the means, medians and standard deviations for Cathy's data.

$$\bar{x} = 530, \tilde{x} = 530, s_1 = 9.00, \bar{y} = 525, \tilde{y} = 527 \text{ and } s_2 = 4.36.$$

2.5 Computing

Given a set of numerical data, the website:

`http://www.wessa.net/rwasp_density.wasp#output`.

will create a kernel density. Time limitations prevent me from discussing this site further.

I could find no websites that create histograms for a set of data.

2.6 Summary

For many data sets, a dot plot does not provide a satisfactory picture of the distribution of the data. In such cases, a researcher might opt for a histogram.

The quest for a histogram begins with the construction of the frequency table. A frequency table consists of five columns, with headings: class interval, width, frequency, relative frequency and density. Note the following:

1. **Class Interval:** There are many valid choices for the class intervals in a frequency table. The collection of class intervals, however, must satisfy the five rules listed on page 31. In these notes I will always provide you with class intervals that obey these rules.

2. **Width:** The width of a class interval is equal to its upper bound minus its lower bound. It is worth noting whether a frequency table has *constant-width* or *variable-width* class intervals.

3. **Frequency:** For each class interval count the number of observations that lie within it, using our endpoint convention: every class interval includes its left endpoint, but not its right, with the exception that the last class interval includes both of its endpoints. The frequencies sum to the number of observations in the data set.

4. **Relative Frequency:** Divide each frequency by the number of observations in the data set; the result is the relative frequency. The relative frequencies for any table sum to one.

5. **Density:** The density of a class interval is equal to its relative frequency divided by its width.

To draw a frequency histogram, follow the two steps on page 32. In particular, the height of a rectangle equals the *frequency* of its class interval. By contrast, in a relative frequency histogram the height of each rectangle equals the *relative frequency* of its class interval. For a density histogram, the height of each rectangle equals the density of its class interval which implies that the area of each rectangle equals the *relative frequency* of its class interval.

For a frequency table with *constant-width* class intervals, all three histograms have the same shape. For a frequency table with *variable-width* class intervals, one should use the density histogram; the other two types of histograms are misleading.

We learned in Chapter 1 that the mean of a set of data is exactly equal to the center of gravity of the data set's dot plot. Thus, for example, if a dot plot is exactly symmetric then the mean (and median) equal the point of symmetry. Result 2.1 on page 33 extends this relationship to dot plots that are not exactly symmetric. This is not a totally satisfactory result because the best I can say is that its conclusions are *usually true*. Despite this weakness, Result 2.1 is considered to be useful.

As discussed earlier in this chapter, a dot plot can be a very *bumpy* picture of a distribution of data. A histogram replaces the bumps with a picture that is flat between (up or down) jumps. A kernel density goes one step further: it has neither bumps nor jumps; it is a smooth picture of the distribution of the data. You will **never** be asked to construct a kernel density.

The Empirical Rule (Result 2.2) provides us with an interpretation of the value of the standard deviation, s: Approximately 68% [95%; 99.7%] of the deviations have a magnitude that is less than or equal to s [$2s$; $3s$].

Table 2.6: Sorted speeds, in MPH, by time-of-day, of Kenny's 100 cars.

Speeds at 6:00 pm

26	26	27	27	27	27	28	28	28	28	28	28	28	28	28	28	28	28
28	29	29	29	29	29	29	29	29	29	29	29	30	30	30	30	30	30
30	30	31	31	31	31	32	33	33	33	34	34	35	43				

Speeds at 11:00 pm

27	28	30	30	30	31	31	31	32	32	32	32	32	32	32	33	33	33
33	33	33	33	34	34	34	34	34	34	35	35	35	35	36	36	36	37
37	37	37	37	37	38	38	39	39	40	40	40	40	40				

2.7 Practice Problems

Recall from Chapter 1 that Kenny the policeman conducted a comparative study on the speeds of cars. Kenny's data are reprinted in Table 2.6. We will use these data in some of the problems below.

1. Using class intervals 26–29, 29–32, 32–35, 35–38, 38–41, and 41–44, construct the frequency tables for both sets of Kenny's data. Remember to use the endpoint convention; for example, the observation 32 is placed in the interval 32–35.

2. Using your tables from problem 1, draw the frequency histograms for both sets of Kenny's data.

3. Kernel densities for Kenny's data are in Figure 2.6 on page 47. Comment on these pictures.

4. The means and standard deviations of Kenny's data are:

   $$\bar{x} = 29.68, s_1 = 2.81, \bar{y} = 34.42 \text{ and } s_2 = 3.25.$$

   Use these values and Kenny's actual data in Table 2.6 to check the performance of the Empirical Rule.

5. The purpose of this problem is to give you some practice working with the three types of histograms. The parts of this problem presented below are variations on the theme *my dog ate my homework* in the sense that each part provides only partial information about a histogram.

   (a) I have a single rectangle from a histogram. Its endpoints are 12 and 15 and its height is 0.03. Given that there are $n = 1000$ observations in the data set, *how many* observations are in this interval (remembering our endpoint convention) if it is:

      i. A frequency histogram?

      ii. A relative frequency histogram?

iii. A density histogram?

(b) I have a single rectangle from a histogram. Its endpoints are 10.00 and 10.05 and its height is 3. Given that there are $n = 500$ observations in the data set, *how many* observations are in this interval (remembering our endpoint convention) if it is:

i. A frequency histogram?

ii. A relative frequency histogram?

iii. A density histogram?

(c) I have a single rectangle from a histogram. Its endpoints are 20 and 22, but I am not going to tell you its height. Given that there are $n = 600$ observations in the data set, and that 10% of these observations are in this interval, *how tall* is the rectangle if it is:

i. A frequency histogram?

ii. A relative frequency histogram?

iii. A density histogram?

The remaining *Problems* in this section do not follow my usual *Question–Answer* format. Rather, they are extended examples, where I illustrate some ideas, but don't ask you to do any work. These remaining *problems are important*, so please read them carefully.

6. This is an extreme example of skewed data, but the data are real. I want to use strongly skewed data because, as you will see, having variable-width class intervals is very useful for skewed data. One of my favorite books for skimming through is *The Baseball Encyclopedia*. Before the internet this book was the first place I would look if I had a question about the history of American professional baseball. A huge segment of *The Baseball Encyclopedia* is devoted to the career statistics of every man who has played major league baseball, dating back to the 1880s, as I recall. The men are separated into two sections, players and pitchers. (A few men appeared in both sections, most notably Babe Ruth who had substantial, and glorious, years as both a pitcher and a player.) A few years ago I selected 200 men at random from the tens of thousands in the player section. I suspect that you have a good idea what I mean by *at random*; I will discuss the concept carefully in Part II of these notes. For my present purposes, your understanding of this issue is not important.

   For each man selected, I recorded a simple statistic: the total number of games in which he appeared during his major league career. I won't list the 200 observations here, but I will give you some highlights:

   (a) The shortest career was one game, a value possessed by 11 players in my sample.

   (b) The longest career in my sample was 3,308 games.

   (c) My three favorite summary statistics are below. Because these data are **not** from a comparative study, I will denote my data by w's, the mean by $\bar{w}$, the standard deviation by s and the sample size by m.

$$\bar{w} = 354.1, \tilde{w} = 83.5 \text{ and } s = 560.4.$$

Note that for this nonnegative response, the ratio of the mean to the median is

$$354.1/83.5 = 4.24!$$

(d) In a data set for which almost one quarter (actual count, 47) of the observations are fewer than 10 games, 21 players had careers of more than 1,000 games, of which six had careers of more than 2,000 games. In fact, the 21 players with the longest careers played a total of 35,151 games, which is almost one-half of the total of 70,810 games played by all 200 players! Any way you look at this data set, these data are strongly skewed to the right.

Figure 2.7 presents a constant-width frequency histogram of these data. The class intervals are: 0–100, 100–200, 200–300, . . . , 3300–3400 games. I do not like this picture! Here is the main feature that I don't like about it:

> Despite having 34 class intervals, over one-half of the data (103 observations) are placed into one interval. At the other extreme, there are twelve class intervals with no observations, six class intervals with one observation and a total of 24 class intervals with three or fewer observations!

This is a bit like having a road atlas (we used these before google maps and gps) with hundreds of pages devoted to Alaska and one-quarter page to New York City. Well, in my actual road atlas, two pages are devoted to New York City and only one page to Alaska. My road atlas *puts emphasis on* the place with lots of streets and people and *discounts* the state with only a handful of highways. We should do the same in Statistics. We accomplish this goal by using variable-width class intervals. The guiding principle is:

- In regions of the number line in which data are plentiful, we want detail. Thus, we make the class intervals narrow.
- In regions of the number line in which data are scarce, we group data more coarsely. Thus, we make the class intervals wide.

Following this principle, I drew a density scale histogram for these baseball data with the following class intervals:

- Four intervals with width $= 25$: 0–25, 25–50, 50–75 and 75–100;
- Four intervals with width $= 100$: 100–200, 200–300, 300–400 and 400–500;
- One interval with width $= 500$: 500–1000;
- One interval with width $= 1000$: 1000–2000; and
- One interval with width $= 1500$: 2000–3500.

This new histogram is presented in Figure 2.8. I will make two comments about it.

(a) In the earlier picture, 103 observations were grouped together into the first class interval, 0–100 games. In the new histogram, this interval has been divided into four narrower intervals. With this extra detail, we can see that almost two-thirds (actual count, 67 out of 103) of these careers were shorter than 25 games (remember out endpoint convention: 0–25 does not include 25). In fact, the rectangle above 0–25 has area:

$$25(0.0134) = 0.335.$$

Recall that for a density histogram, *area equals relative frequency*. Thus, 33.5% of the observations are in the interval 0–25. Finally, 33.5% of 200 equals 67 players, as I mention above parenthetically.

(b) Beginning with the class interval 500–1000 and moving to the right, this new histogram is much smoother than the earlier constant-width frequency histogram. I like this because, as a baseball aficionado, I can think of no reason other than *chance variation* for the numerous bumps in the earlier histogram. Note that the area of the rectangle above 1000–2000 is:

$$1000(0.000075) = 0.075.$$

Thus, 7.5% of the 200 players—i.e., 15 players—had careers of 1000–2000 games.

Finally, Figure 2.9 is the frequency histogram for the same class intervals as in Figure 2.8. When you compare these two figures you will see why statisticians label the frequency histogram *misleading*. It is misleading because even if we are told to focus on the height of each rectangle, we see the area.

7. The goal of this *problem* is to give you additional insight of the Empirical Rule, Result 2.2. I used Minitab to generate three artificial data sets, each of size 1000. I will **not** give you a listing of these data sets, **nor** will I draw their histograms. Instead, note the following:

   (a) The data sets all have the same mean, 500, and standard deviation, 100.

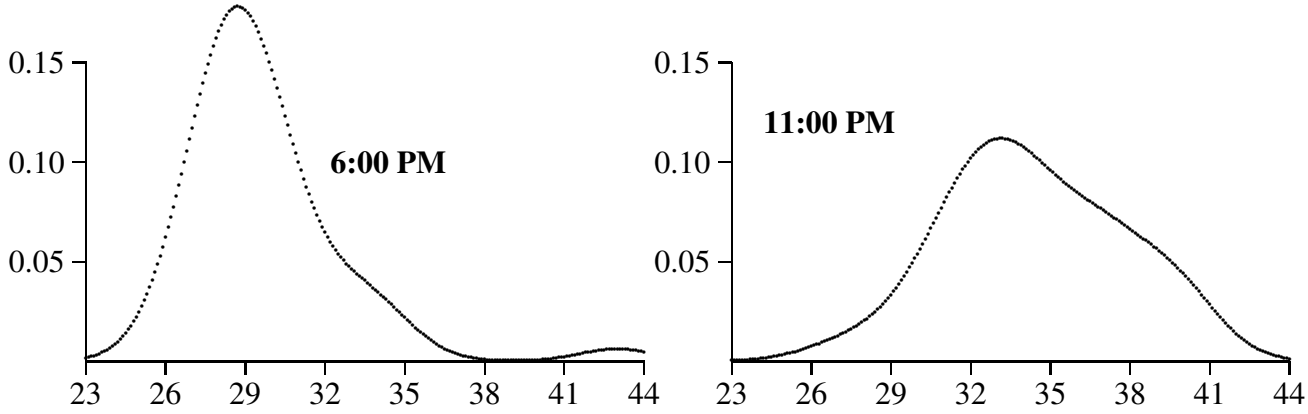
   (b) The first data set has a symmetric, bell-shaped histogram. The second data set has a symmetric, rectangular-shaped histogram. The third data set is strongly skewed to the right.

   Table 2.7 presents a number of summaries of these three data sets. Please examine this table before reading my comments below. Remember: I have **not** given you enough information to verify the counts in this table; trust me on these please.

   (a) The Empirical Rule approximations are nearly exact for the symmetric, bell-shaped histogram. The Empirical Rule does not work well for the other shapes.

   (b) For the symmetric, rectangular-shaped histogram the Empirical Rule approximation count for the interval $\bar{w} \pm s$ is much larger than the actual count. For the interval $\bar{w} \pm 2s$ the Empirical Rule approximation count is much smaller than the actual count.

Figure 2.6: Kernel estimates for Kenny's 6:00 PM and 11:00 PM data.



(c) For the strongly skewed histogram, the Empirical Rule approximation count for the interval $\bar{w} \pm s$ is much smaller than the actual count. Because of this discrepancy, statisticians sometimes abuse the language and say that for skewed data, the standard deviation is *too large*. Obviously, the standard deviation is simply an arithmetic computation; it is what it is and is neither too large nor too small. But, in my opinion, the misspeak has some value. First, in the Empirical Rule, instructing Elastic Man to reach s units in both directions in order to encompass 68% of the data is, indeed, telling my favorite superhero to *reach too far*. Second, as we will see often in these notes when looking at real data, even one extreme value in a data set has a large impact on the value of the standard deviation—making it larger, often much larger. Skewed data almost always contain at least one extreme value.

Also, the Empirical Rule approximation count for the interval $\bar{w} \pm 2s$ is quite close to the actual count. Finally, the Empirical Rule approximation count for the interval $\bar{w} \pm 3s$ is substantially larger than the actual count.

(d) Table 2.7 also presents counts for the data set on length of baseball careers that we studied in the previous problem. These baseball data had 200 observations, not 1,000, so I needed to adjust my Empirical Rule approximation counts. The pattern for these baseball data matches the pattern for the artificial strongly skewed data set.

Figure 2.7: Frequency histogram of the number of games played by $n_1 = 200$ major league baseball players.

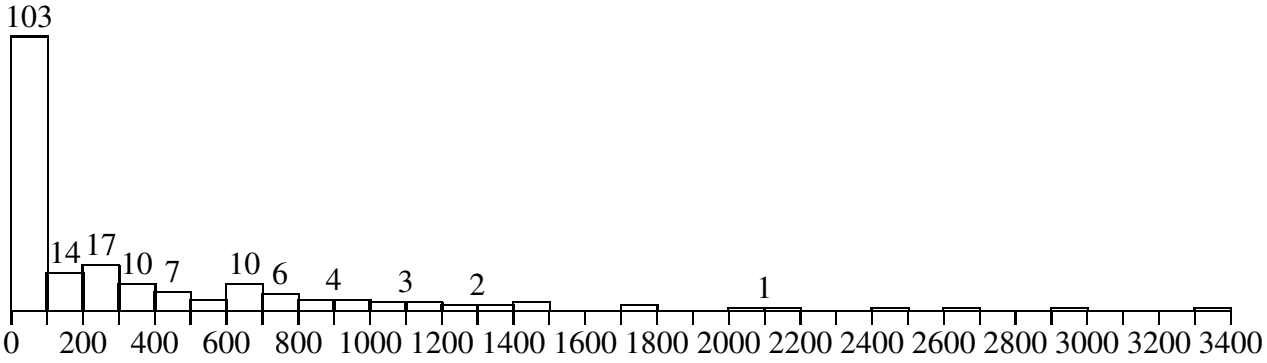


Figure 2.8: Variable-width density histogram of the number of games played by $n_1 = 200$ major league baseball players.

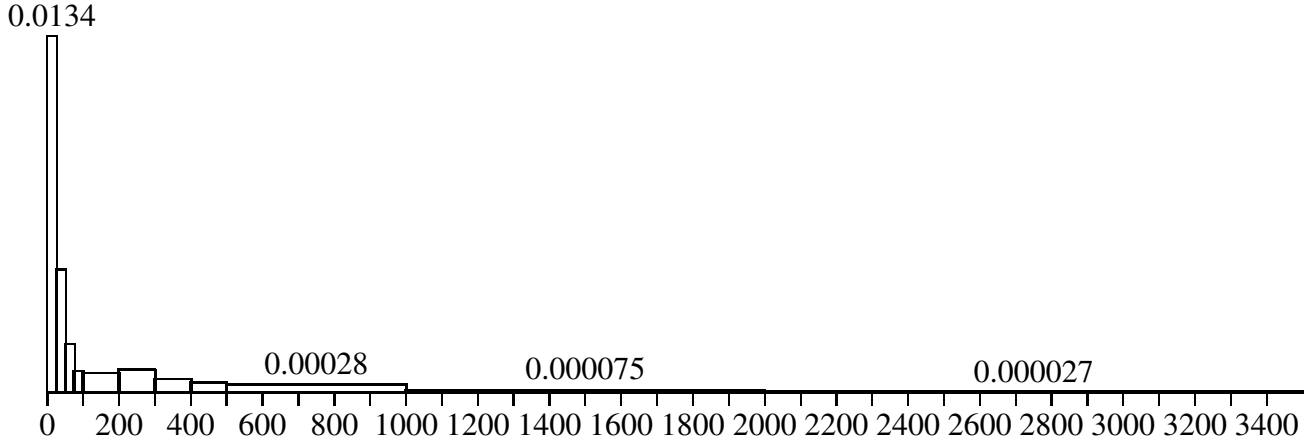


Figure 2.9: Misleading frequency histogram of the number of games played by $n_1 = 200$ major league baseball players.

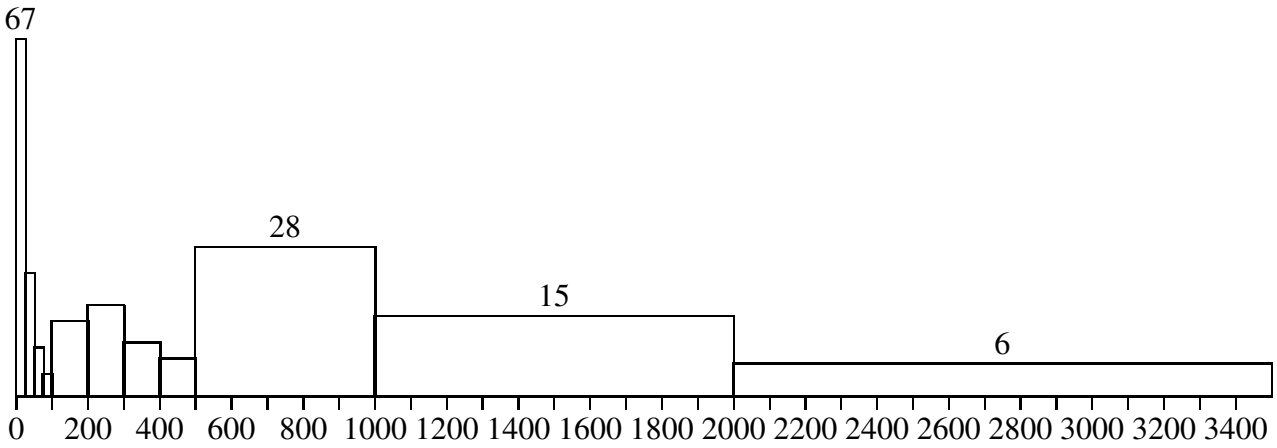


Table 2.7: An examination of the performance of the Empirical Rule for the three artificial data sets in Practice Problem 7 and the real baseball career data in Practice Problem 6.

				Actual Counts:		
Data Set	Shape	Min.	Max.	$\bar{w} \pm s$	$\bar{w} \pm 2s$	$\bar{w} \pm 3s$
1	Symmetric, bell	177	823	682	954	998
2	Symmetric, rectangular	327	673	578	1,000	1,000
3	Skewed to the right	402	989	868	941	979
Empirical Rule Approximation:				680	950	997
Baseball Careers:		0	3,308	175	192	194
Empirical Rule Approximation:				136	190	199

2.8 Solutions to Practice Problems

1. The frequency tables are in Table 2.8.

2. The histograms are in Figure 2.10.

3. These pictures are smooth, of which I approve! The 6:00 PM kernel density is skewed to the right with a peak at about 29 MPH, in agreement with our earlier pictures. The 11:00 PM kernel density has a single peak at about 33 MPH. Its two tails have approximately the same length, but the right tail is heavier.

   Here is a feature that *I do not like* about the 11:00 PM kernel density: To me, one of the most interesting features in the data set is the fact that while five cars were traveling 40 MPH, none was going faster. This feature is obliterated in the kernel density.

4. The first interval for the 6:00 PM data is:

$$\bar{x} \pm s_1 = 29.68 \pm 2.81 = [26.87, 32.49].$$

   From the table, we see that two observations are smaller than 26.87 and seven observation are larger than 32.49. Thus, this interval encompasses $50 - (2 + 7) = 41$ observations, which is 82% of the total of 50 observations. The Empirical Rule approximation of 68% is not good.

   The second interval for the 6:00 PM data is:

$$\bar{x} \pm 2s_1 = 29.68 \pm 5.62 = [24.06, 35.30].$$

   This interval encompasses $50 - 1 = 49$ observations, which is 98% of the total of 50 observations. The Empirical Rule approximation of 95% is a bit small.

   Finally, the third interval for the 6:00 PM data is:

$$\bar{x} \pm 3s_1 = 29.68 \pm 8.43 = [21.25, 38.11].$$

   This interval encompasses $50 - 1 = 49$ observations, which is 98% of the total of 50 observations. The Empirical Rule approximation of 99.7% is a bit large.

   The first interval for the 11:00 PM data is:

$$\bar{y} \pm s_2 = 34.42 \pm 3.25 = [31.17, 37.67].$$

   From the table, we see that eight observations are smaller than 31.17 and nine observation are larger than 37.67. Thus, this interval encompasses $50 - (8 + 9) = 33$ observations, which is 66% of the total of 50 observations. The Empirical Rule approximation of 68% is quite good.

   The second interval for the 11:00 PM data is:

$$\bar{y} \pm 2s_2 = 34.42 \pm 6.50 = [27.92, 40.92].$$

This interval encompasses $50 - 1 = 49$ observations, which is 98% of the total of 50 observations. The Empirical Rule approximation of 95% is a bit small.

Finally, the third interval for the 11:00 PM data is:

$$\bar{y} \pm 3s_2 = 34.42 \pm 9.75 = [24.67, 44.17].$$

This interval encompasses all 50 observations, The Empirical Rule approximation of 99.7% is quite good.

5. (a) Given that the height is 0.03:
   i. This is impossible. For a frequency histogram the height of each rectangle must be an integer.
   ii. Three percent of the 1000 observations are in this interval:
   $$0.03(1000) = 30 \text{ observations.}$$
   iii. The area of the rectangle is $3(0.03) = 0.09$. Thus, 9% of the 1000 observations are in this interval:
   $$0.09(1000) = 90 \text{ observations.}$$

   (b) Given that the height is 3:
   i. 3. For a frequency histogram the height of a rectangle tells us how many observations are in its class interval.
   ii. This is impossible. For a relative frequency histogram the height of a rectangle cannot exceed one.
   iii. The area of the rectangle is $0.05(3) = 0.15$. Thus, 15% of the 500 observations are in this interval:
   $$0.15(500) = 75 \text{ observations.}$$

   (c) We are given that the relative frequency of the interval is 0.10, which makes its frequency $0.10(600) = 60$.
   i. 60. For a frequency histogram the height of a rectangle equals the number of observations in the class interval.
   ii. 0.10. For a relative frequency histogram the height of a rectangle equals the relative frequency of observations in it.
   iii. The width of the class interval is 2. Thus, the density of the interval is $0.10/2 = 0.05$, which is also the height of its rectangle.

Table 2.8: Frequency tables for Kenny's data.

		6:00 PM			11:00 PM		
Class Interval	Width	Freq.	Rel. Freq.	Density	Freq.	Rel. Freq.	Density
26–29	3	19	0.38	0.127	2	0.04	0.013
29–32	3	23	0.46	0.153	6	0.12	0.040
32–35	3	6	0.12	0.040	20	0.40	0.133
35–38	3	1	0.02	0.007	13	0.26	0.087
38–41	3	0	0.00	0.000	9	0.18	0.060
41–44	3	1	0.02	0.007	0	0.00	0.000
Total		50	1.00		50	1.00	

Figure 2.10: Frequency histograms of car speeds, by time.

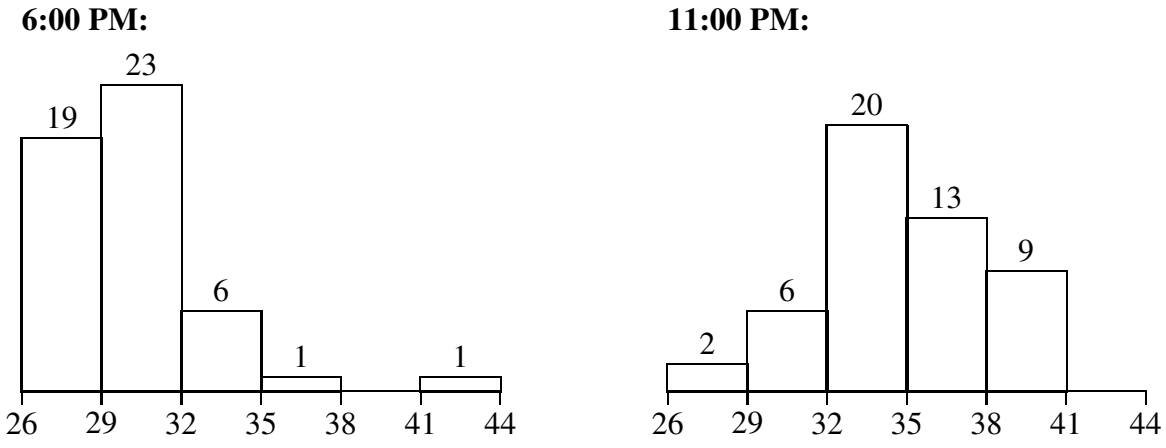
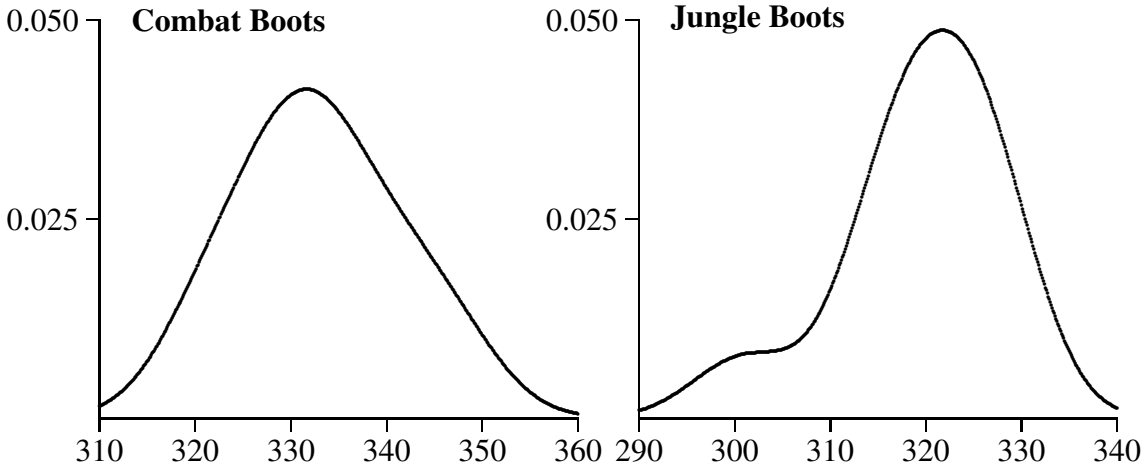


Figure 2.11: Kernel densities for Brian's combat boots and jungle boots data.



2.9 Homework Problems

In the Chapter 1 Homework Problems we learned about Brian's study of his running times. (See page 24.) Use Brian's data, reproduced below, to solve problems 1–5. Wearing combat boots, Brian's sorted times were:

321 323 329 330 331 332 337 337 343 347

Wearing jungle boots, Brian's sorted times were:

301 315 316 317 321 321 323 327 327 327

1. Create the frequency table and draw the frequency histogram for Brian's combat boot data, using the class intervals: 320–330, 330–340 and 340–350. Briefly describe the shape of the histogram.

2. Create the frequency table and draw the frequency histogram for Brian's combat boot data, using the class intervals: 321–333, 333–345 and 345–357. Briefly describe the shape of the histogram.

3. Compare your answers to problems 1 and 2.

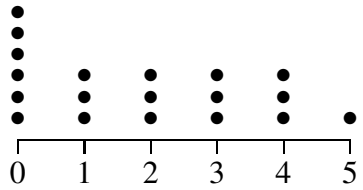
4. Create the frequency table and draw the frequency histogram for Brian's jungle boot data, using the class intervals: 300–310, 310–320 and 320–330. Briefly describe the shape of the histogram.

5. Figure 2.11 present kernel densities for Brian's two data sets. Briefly describe what these pictures reveal about the data. Compare these kernel densities to the three frequency histograms from problems 1, 2 and 4.

6. Refer to Sara's golfing data in Table 2.2. *For this problem only* I am going to combine Sara's two sets of data to obtain one set of data with 80 observations. The mean of these 80 numbers is 102.52 and the standard deviation is 29.25.

   How do the Empirical Rule (Result 2.2) approximations perform for these data?

7. Below is a dot plot of $m = 19$ observations.



   (a) How would you label the shape of this dot plot? Approximately symmetric? Skewed? Other?

   (b) Calculate the mean and median of these data.

   (c) Given your answers to (a) and (b), comment on Result 2.1.

Chapter 3

Randomization, Probability and Sampling Distributions

3.1 Assignments and Randomization

Recall that Dawn's study of her cat Bob was presented in Chapter 1. Table 3.1 presents her data. Reading from this table, we see that in Dawn's study, the chicken-flavored treats were presented to Bob on days (trials):

$$1, 5, 7, 8, 9, 11, 13, 15, 16 \text{ and } 18.$$

Why did she choose these days? How did she choose these days? It will be easier to begin with the 'How' question.

We have been looking at the data collected by Dawn. We have listed the observations; separated them by treatment; sorted them within treatment; and, within treatments, drawn dot plots and computed means, medians, variances and standard deviations. But now we need to get into our time machine and travel back in time to **before** Dawn collected her data. We go back to when Dawn had her study largely planned: treatments selected; trials defined; response specified; and the decision to have a balanced study with 20 trials. We are at the point where Dawn pondered, "Which 10 trials should have chicken-flavored treats assigned to them? How should I decide?"

Table 3.1: Dawn's data on Bob's consumption of cat treats. 'C' ['T'] is for chicken [tuna] flavored.

Day:	1	2	3	4	5	6	7	8	9	10
Flavor:	C	T	T	T	C	T	C	C	C	T
Number Consumed:	4	3	5	0	5	4	5	6	1	7

Day:	11	12	13	14	15	16	17	18	19	20
Flavor:	C	T	C	T	C	C	T	C	T	T
Number Consumed:	6	3	7	1	3	6	3	8	1	2

The answer is that Dawn did this by using a process called **randomization**. I will explain *what* randomization is by showing you three equivalent ways to randomize.

First, some terminology. We call the list of 10 trials above an **assignment** of treatments to trials. It tells us which trials were assigned to the first treatment (chicken). It also implies which trials were assigned to the second treatment (tuna); namely, all of the trials *not* listed above. If we are going to study assignments—and we are—it is easier if we make our assignments as simple to display as possible. **Thus, an assignment will be presented by listing the trials that it assigns to treatment 1.**

A natural question is: How many different assignments were possible for Dawn's study? The answer is 184,756. I will give a brief digression into how I obtained this number.

You might recall from math the expression $m!$, which is read *em-factorial*. If m is a positive integer, then this expression is defined as:

$$m! = m(m-1)(m-2)\cdots 1 \tag{3.1}$$

Thus, for example,

$$1! = 1; 2! = 2(1) = 2; 3! = 3(2)(1) = 6; \text{ and so on.}$$

By special definition (which will allow us to write more easily certain formulas that will arise later in these notes), $0! = 1$. Finally, for any other value of m (negatives, non-integers), the expression $m!$ is not defined.

We have the following result. You don't need to worry about proving it; it is a given in these notes.

Result 3.1 (The number of possible assignments.) For a total of $n = n_1 + n_2$ units, the number of possible assignments of two treatments to the units, with n_1 units assigned to treatment 1 and the remaining n_2 units assigned to treatment 2, is

$$\frac{n!}{n_1!n_2!} \tag{3.2}$$

I will evaluate Equation 3.2 for three of the studies presented in Chapters 1 and 2.

- For Cathy's study, $n = 6$ and $n_1 = n_2 = 3$. Thus, the number of possible assignments is

$$\frac{6!}{3!3!} = \frac{6(5)(4)}{3(2)(1)} = 20.$$

  Notice that it is **always** possible to reduce the amount of arithmetic we do by canceling some terms in the numerator and denominator. In particular, the $6!$ in the numerator can be written as

$$6(5)(4)3!$$

  and its $3!$ cancels a $3!$ in the denominator.

- For Kymn's study, $n = 10$ and $n_1 = n_2 = 5$. Thus, the number of possible assignments is

$$\frac{10!}{5!5!} = \frac{10(9)(8)(7)(6)}{5(4)(3)(2)(1)} = 252.$$

- For Dawn's study, $n = 20$ and $n_1 = n_2 = 10$. Thus, the number of possible assignments is

$$\frac{20!}{10!10!} = \text{ 184,756.}$$

Notice that for Cathy's and Kymn's study, I determined the answer *by hand* because the numbers are small enough to handle easily. Dawn's study is trickier. Many of you, perhaps most, perhaps all, will consider it *easy* to determine the answer: 184,756. But I will not require you to do so. As a guide, I will **never** have you evaluate $m!$ for any $m > 10$.

Sara's study is a real challenge. The number of possible assignments is

$$\frac{80!}{40!40!}.$$

This answer, to four significant digits, is

$$1.075 \times 10^{23}.$$

Don't worry about how I obtained this answer. If this issue, however, keeps you awake at night, then send me an email and I will tell you. If enough people email me, then I will put a brief explanation in the next version of these *Course Notes*.

I now will describe three ways—two physical and one electronic—that Dawn could have performed her randomization.

1. **A box with 20 cards.** Take 20 cards of the same size, shape, texture, etc. and number them 1, 2, ..., 20, with one number to each card. Place the cards in a box; mix the cards thoroughly and select 10 cards at random without replacement. The numbers on the cards selected denoted the trials that will be assigned treatment 1.

2. **A deck of 20 ordinary playing cards.** (This method is especially suited for units that are trials.) We need to have 10 black cards (spades or clubs) and 10 red cards (diamonds or hearts). We don't care about the rank (ace, king, 3, etc.) of the cards. The cards are thoroughly shuffled and placed in a pile, face down. Before each trial select the top card from the pile; if it is a black card, then treatment 1 is assigned to the trial; if it is a red card, then treatment 2 is assigned to the trial. The selected card is set aside and the above process is repeated for the remaining trials.

3. **Using a website.** This method will be explained in Section 3.5 later in this chapter.

Are you familiar with the term **black box**? I like the definition in Wikipedia

`http://en.wikipedia.org/wiki/Black_box`

which is:

> In science and engineering, a black box is a device, system or object which can be viewed solely in terms of its input, output and transfer characteristics without any knowledge of its internal workings, that is, its implementation is "opaque" (black). Almost anything might be referred to as a black box: a transistor, an algorithm, or the human mind.

Our website for randomization is a black box. It executes a computer program that supposedly is mathematically equivalent to my two methods of randomization that involve using cards. I say it's a black box because we aren't really interested in the computer code of the program.

Thus, if you want to think about how randomization works, I recommend you think of the cards. For my purposes of instruction the most convenient method for me is to use the website to obtain examples for you. If I were to replicate Dawn's study on my cat **Buddy** I would use the second card method above. But that's me. If you perform a project for this class, you may randomize however you please, as long as you use one of the three methods above.

Before I get back to Dawn's study, I want to deal with an extremely common misconception about randomization. Randomization is a **process** or a **method** that is **fair** in the sense that every possible assignment has the same chance of being selected. Randomization **does not guarantee** that the assignment it yields **will look random**. (Among the many issues involved is how one decides what it means for an assignment to *look random*.) Later in these *Course Notes*, we will discuss designs that involve **restricted randomization**; in particular, we will learn about the **Randomized Pairs Design**.

I used the website and obtained the following assignment for Dawn's study:

> 1, 2, 4, 7, 9, 10, 11, 14, 15, 18.

Note that this assignment is different from the one that Dawn used. Given that there are 184,756 possible assignments I would have been *very surprised* if I had obtained the same assignment as Dawn!

Here is the commonality of our three methods of randomization: **Before** we select an assignment, all 184,756 possible assignments for Dawn's study are **equally likely** to be selected. For the first method of randomizing, this fact is conveyed by saying that the cards are indistinguishable; they are thoroughly mixed; and 10 cards are selected at random. For the second method of randomizing, this fact is conveyed by saying that the cards are shuffled thoroughly. Finally, whereas the electronic methods' operations are a total mystery to us, the programmer claims that it makes all assignments equally likely. In this class we will use the website randomizer for a variety of purposes (not just randomization) and we will **accept without question** its claim of making every assignment equally likely.

The most important thing to remember about probability is that it is **always** about a **future uncertainty**.

Now, back to Dawn's study of Bob. Remember, we are at a place in time *before* Dawn selected an assignment. Thus, it makes sense to talk about probabilities. Because all assignments are

equally likely to be selected, we assign each of them the same probability of being selected, as given below.

$$P(\text{Any particular assignment}) = \frac{1}{\text{Total number of possible assignments}} \tag{3.3}$$

For Dawn's study, the probability of any particular assignment is $1/184{,}756$. For Cathy's study, the probability of any particular assignment is $1/20 = 0.05$ because, recall (Equation 3.2) that for $n_1 = n_2 = 3$, there are 20 possible assignments of treatments to units.

We will be interested in combining assignments (perhaps with some common feature) into a collection of assignments. Let A denote any collection of assignments. Such a collection is called an **event**. **Before** we select an assignment via randomization, it makes sense to ask whether or not the assignment that *will be obtained* (note the future tense) is one of the assignments belonging to the event A. If it is, we will say that the event A has occurred. We have the following definition of the probability that an event A will occur.

$$P(A) = \frac{\text{Number of assignments in } A}{\text{Total number of possible assignments}} \tag{3.4}$$

An important example of a collection of assignments is the collection of **all** possible assignments; it is called the **Sample Space** and is denoted by $\mathcal{S}$ (an upper case ess with an attitude). From Equation 3.4 we see that the probability of the sample space occurring is 1 or 100%.

The sample space is called the certain event. Why? Well, by definition it contains all possible assignments; thus, it is certain to occur. Thus, we see that certainty corresponds to a probability of 1. An empty collection of assignments is impossible to occur and has probability equal to 0. Thus, probability is a quantification of uncertainty for which 0 corresponds to impossible and 1 corresponds to certainty.

3.2 The Skeptic's Argument

Recall my earlier discussion of within- and between-variation. Let's return to within variation. Consider, for example, Dawn's sorted chicken data: 1, 3, 4, 5, 5, 6, 6, 6, 7 and 8. These numbers vary. This fact cannot be debated. In words, some days Bob consumed a large number of chicken treats and other days he consumed very few. Bob exhibited a large amount of day-to-day variation in his eating of chicken treats. In somewhat more picturesque language, some days Bob was very hungry; other days hungry; other days not so hungry and other days barely hungry at all.

Now consider between-variation. I originally stated that the chicken data were, as a group, larger than the tuna data. I need to be more precise; I need to replace the vague expression *as a group* by something else. Well, I have already talked about this; I will summarize a data set by calculating its mean (or median, but let's stick to the mean now for simplicity). Thus, the between-variation is reflected in the mean for chicken being 2.2 treats larger than the mean for tuna.

This leads to a very important realization. Whereas the within variation is definitely real, I cannot be sure that the between variation is real. Indeed, for dramatic purposes I am going to invent a person I call the Skeptic. The Skeptic is the originator of what we will call the **Skeptic's Argument**, stated below:

The flavor of the treat is irrelevant. The number of treats that Bob consumed on any given day was determined by how hungry Bob was on that day.

The Skeptic's Argument is so central to what we do in this course that it behooves us to spend some time thinking about it.

Consider the day when Bob was offered chicken and he consumed eight treats. According to the Skeptic, the flavor was irrelevant. Bob consumed eight treats because he was very hungry that day. **If** Bob had been offered tuna that day, the Skeptic believes that Bob would have consumed eight treats. Similarly, there was a day that Bob was offered tuna and he refused to eat. To the Skeptic, this means that on that day Bob was not hungry at all and, if he had been offered chicken treats, he would have eaten none of them.

Please note the following two items:

1. As mere mortals, it is impossible for us to determine **with certainty** whether or not the Skeptic's Argument is correct.

2. You are certainly allowed to have your own opinion as to whether the Skeptic is correct. I am not going to tell you what to believe, just as I would not tell you whether you like chocolate. You need to learn, however, how statisticians evaluate the Skeptic's Argument.

Regarding the second item above, there are many details to learn about how statisticians work. The method we advocate has both strengths and weaknesses and we will learn about both categories of features.

It will be useful to invent an adversary for the Skeptic; I will call it the Advocate. I want to avoid making this current chapter too long; thus, I will delay our consideration of the Advocate's Argument until Chapter 5.

Please note the following. For ease of exposition, on occasion I will refer to the Skeptic and Advocate as real people. I want these characters to be gender-free; hence, when I refer to either of them with a pronoun—as I do the Advocate in the previous paragraph—I will use the pronoun *it* rather than *he* or *she*.

In the remainder of this chapter we will combine the Skeptic's Argument with our notions of probability to obtain **the sampling distribution of the test statistic**. The importance of this sampling distribution will be presented in Chapter 5.

3.3 The Sampling Distribution of the Test Statistic for Cathy's Study

Recall that Cathy's running study was introduced on page 40.

For convenience, I have presented her data again in Table 3.2. Also, her treatment means are:

$$\bar{x} = (530 + 521 + 539)/3 = 530 \text{ and } \bar{y} = (528 + 520 + 527)/3 = 525.$$

We see that Cathy's randomization assigned trials 1, 2 and 5 to the high school and the remaining

Table 3.2: Cathy's times, in seconds, to run one mile. HS means she ran at the high school and P means she ran through the park.

Trial:	1	2	3	4	5	6
Location:	HS	HS	P	P	HS	P
Time:	530	521	528	520	539	527

Table 3.3: The 20 possible assignments for Cathy's CRD.

1,2,3	1,2,4	1,2,5	1,2,6	1,3,4	1,3,5	1,3,6	1,4,5	1,4,6	1,5,6
2,3,4	2,3,5	2,3,6	2,4,5	2,4,6	2,5,6	3,4,5	3,4,6	3,5,6	4,5,6

trials to the park. In order to save space below, I will refer to this as assignment 1,2,5.

The number of possible assignments for Cathy's study is quite small; using Equation 3.2 on page 56, we obtain:

$$\frac{6!}{3!3!} = \frac{6(5)(4)}{3(2)} = 20 \text{ possible assignments.}$$

Thus, it is relatively easy to list all possible assignments; they are presented in Table 3.3.

As I show above, Cathy's mean time on the high school route is 5 seconds larger than her mean time on the park route, because, $\bar{x} - \bar{y} = 530 - 525 = 5$ seconds. It is helpful now, and necessary in Chapter 5, for me to introduce additional notation and terminology.

Because I am comparing treatments by subtracting means, I will give the difference a symbol; let $u = \bar{x} - \bar{y}$. For Cathy's data, $u = 5$. We call u the **observed value** of the **test statistic** U. Admittedly, this language and notation is confusing. But it is standard and I am unable to change it because, frankly, I am not the tsar of the Statistics world!

I suggest that you think of these ideas in the following way. Before anyone collects data, the test statistic, U, is a *rule* or *plan* or *protocol* for what will be done with the data. In this chapter (and two more to follow) the rule is: we will collect data, separate data by treatments, compute means and compare means by subtracting. When the rule is applied, the result of all that work will be a number, which we denote by u.

The reason we need the above is because of the interesting way (peculiar way?) statisticians analyze data. Above we see that Cathy's study led to $u = 5$ as the observed value of the test statistic. The question we face is: How do we interpret $u = 5$? Should we be impressed? Should we be unimpressed? Is it real? (Whatever that means!)

This leads us to one of the big ideas in Statistics:

> We evaluate what **actually happens** in a study by comparing it to **everything that could have happened** in the study.

So, the first thing I do now is I start adding the adjective **actual** in places that were previously modifier free. In particular, Cathy's actual assignment was 1,2,5 which led to her actual x's of 530,

521 and 539 and her actual y's of 528, 520 and 527, and her actual means of 530 and 525 and, most importantly, her actual u of 5.

Next, let's look at the statement, "Everything that could have happened." Do **not** take an expansive view of this statement! It *could have happened* that Cathy chose a different topic; or that Cathy did not take my class; or that I never became a statistician; or that the dinosaurs were never wiped out by an asteroid.

Statisticians take an extremely narrow view of the statement, "Everything that could have happened." To us, what happened is that Cathy obtained assignment 1,2,5; everything that could have happened refers to the 19 other possible assignments.

Therefore, the fundamental quest of a statistician looking at Cathy's data—and indeed most studies we ever consider— is to determine what u would have been obtained for each of the other (19, in Cathy's case) possible assignments. Nineteen is a pretty small number; thus, examining 19 assignments seems manageable. Let's give it a try!

Let's begin with assignment 1,2,3 which means, of course, that Cathy would run at the high school on trials 1, 2 and 3, and through the park on trials 4, 5 and 6. What would have happened?

Well, I can give you a partial answer quite easily. In fact, I can list several things that I *know* to be true:

- Trial 1: In the actual study, trial 1 was assigned to the high school and Cathy obtained a response of 530. Thus, because assignment 1,2,3 assigns trial 1 to the high school, the response would have been 530.

- Trial 2: For the same reason I give above for trial 1, the response would have been 521.

- Trial 4: In the actual study, trial 4 was assigned to the park and Cathy obtained a response of 520. Thus, because assignment 1,2,3 assigns trial 4 to the park, the response would have been 520.

- Trial 6: For the same reason I give above for trial 4, the response would have been 527.

- Trial 3: In the actual study, trial 3 was assigned to the park and Cathy obtained a response of 528. Assignment 1,2,3 assigns trial 3 to the high school. As a result, *nobody can say what the response would have been!*

- Trial 5: As with trial 3, nobody can say what the response would have been.

It seems that we are at an impasse. I cannot say, with certainty, what would have happened if Cathy had used assignment 1,2,3. In fact, a variation of the above argument can be made for *every one* of the 19 assignments that were not used in the actual study.

So, what do we do? We add an assumption. We add the assumption that the Skeptic's Argument is correct. Now, let's revisit my argument above for trial 3.

- Trial 3: In the actual study, trial 3 was assigned to the park and Cathy obtained a response of 528. Assignment 1,2,3 assigns trial 3 to the high school.

Table 3.4: The values of u for all possible assignments for Cathy's CRD.

Assignment	x values	y values	u
$1, 2, 3$	$530, 521, 528$	$520, 539, 527$	-2.33
$1, 2, 4$	$530, 521, 520$	$528, 539, 527$	-7.67
$1, 2, 5$	$530, 521, 539$	$528, 520, 527$	$+5.00$ (Actual)
$1, 2, 6$	$530, 521, 527$	$528, 520, 539$	-3.00
$1, 3, 4$	$530, 528, 520$	$521, 539, 527$	-3.00
$1, 3, 5$	$530, 528, 539$	$521, 520, 527$	$+9.67$
$1, 3, 6$	$530, 528, 527$	$521, 520, 539$	$+1.67$
$1, 4, 5$	$530, 520, 539$	$521, 528, 527$	$+4.33$
$1, 4, 6$	$530, 520, 527$	$521, 528, 539$	-3.67
$1, 5, 6$	$530, 539, 527$	$521, 528, 520$	$+9.00$
$2, 3, 4$	$521, 528, 520$	$530, 539, 527$	-9.00
$2, 3, 5$	$521, 528, 539$	$530, 520, 527$	$+3.67$
$2, 3, 6$	$521, 528, 527$	$530, 520, 539$	-4.33
$2, 4, 5$	$521, 520, 539$	$530, 528, 527$	-1.67
$2, 4, 6$	$521, 520, 527$	$530, 528, 539$	-9.67
$2, 5, 6$	$521, 539, 527$	$530, 528, 520$	$+3.00$
$3, 4, 5$	$528, 520, 539$	$530, 521, 527$	$+3.00$
$3, 4, 6$	$528, 520, 527$	$530, 521, 539$	-5.00
$3, 5, 6$	$528, 539, 527$	$530, 521, 520$	$+7.67$
$4, 5, 6$	$520, 539, 527$	$530, 521, 528$	$+2.33$

> Because the Skeptic is correct, the treatment does not matter. Cathy obtained a response of 528 on trial 3 because that reflected her energy, enthusiasm, the weather, whatever, on trial 3. If she had run at the high school, her response would have been 528.

With a similar argument, we see that by adding the assumption that the Skeptic is correct, we *know* that Cathy's time on trial 5 would have been 539 whether she ran at the high school (as she actually did) or through the park, as assignment 1,2,3 would have told her to do.

To summarize: for assignment 1,2,3, Cathy's x's would have been: 530, 521 and 528; her y's would have been 520, 539 and 527. (You can find these easily by looking at Table 3.2 and simply ignoring the listed treatments; after all, according to the Skeptic, the treatments don't matter.) Continuing, for assignment 1,2,3, Cathy's $\bar{x} = 526.33$, $\bar{y} = 528.66$ and $u = -2.33$. I can repeat this analysis for the remaining 18 possible assignments. My work is summarized in Table 3.4. You don't need to verify *all* of the entries in this table, but you should verify enough to convince yourself that you understand the method.

There is no nice way to say this: The results in Table 3.4 are a mess! The 20 possible as-

Table 3.5: The sampling distribution of U for Cathy's CRD.

u	$P(U=u)$	u	$P(U=u)$	u	$P(U=u)$
-9.67	0.05	-3.00	0.10	3.67	0.05
-9.00	0.05	-2.33	0.05	4.33	0.05
-7.67	0.05	-1.67	0.05	5.00	0.05
-5.00	0.05	1.67	0.05	7.67	0.05
-4.33	0.05	2.33	0.05	9.00	0.05
-3.67	0.05	3.00	0.10	9.67	0.05

Table 3.6: Kymn's times, in seconds, to row 2000 meters on an ergometer. Treatment 1 is the small gear with the vent closed; and treatment 2 is the large gear with the vent open.

Trial:	1	2	3	4	5	6	7	8	9	10
Treatment:	2	1	1	1	2	2	1	2	2	1
Response:	485	493	489	492	483	488	490	479	486	493

signments lead to 18 different values of u! Before I give in to despair, I will summarize this information in Table 3.5. In this new table, I have taken the 18 different values of u and sorted them from smallest to largest. I then divided their frequencies of occurrence (16 of which are one and two of which are two) by the total number of assignments, 20, to obtain the probabilities given in the table. This table is a **representation** of the **sampling distribution** of the test statistic U. Sometimes, the sampling distribution is represented by an equation; in any event it contains all possible values of U, each matched with its probability of occurring. As we will see in Chapter 5, the sampling distribution is a critical component of how we learn from a CRD. Before our next example, the last one of this chapter, I will make an obvious comment: Even for a CRD with a very small number of possible assignments—20 in this case—it is *extremely tedious* and *messy* to determine the sampling distribution of the test statistic U.

3.4 The Sampling Distribution of U for Kymn's CRD

Kymn's study and data were presented in Chapter 2. Her data are reproduced in Table 3.6.

Kymn's study will be manageable for me because, as shown earlier, there are only 252 possible assignments of trials to treatments. Note from Table 3.6 that her actual assignment was 2, 3, 4, 7 and 10, which yielded

$$\bar{x} = 491.4, \bar{y} = 484.2 \text{ and } u = 491.4 - 484.2 = 7.2.$$

I used the randomizer to obtain a second possible assignment: I obtained 4, 5, 7, 9 and 10.

Table 3.7: Frequency table for u for the 252 possible assignments for Kymn's study.

u	Freq.	u	Freq.	u	Freq.	u	Freq.	u	Freq.	u	Freq.
-7.2	1	-4.8	3	-2.4	10	0.4	12	2.8	10	5.2	4
-6.8	1	-4.4	5	-2.0	8	0.8	10	3.2	8	5.6	3
-6.4	1	-4.0	8	-1.6	14	1.2	13	3.6	6	6.0	1
-6.0	1	-3.6	6	-1.2	13	1.6	14	4.0	8	6.4	1
-5.6	3	-3.2	8	-0.8	10	2.0	8	4.4	5	6.8	1
-5.2	4	-2.8	10	-0.4	12	2.4	10	4.8	3	7.2	1
				0.0	16					Total	252

Referring to Table 3.6, you can verify the following:

$$\bar{x} = (492 + 483 + 490 + 486 + 493)/5 = 488.8, \bar{y} = (485 + 493 + 489 + 488 + 479)/5 = 486.8$$

and $u = 488.8 - 486.8 = 2.0$. I continued the above process, examining all 252 possible assignments and determining the value of $u = \bar{x} - \bar{y}$ for each of them. (Warning: Do not attempt this! It was no fun and quite tedious; although, in fairness, I should acknowledge that many teachers are actually quite good at doing this. I am not.) The 252 assignments result in 37 possible values of u; my results are presented in Table 3.7. If we divide these frequencies by 252 we have the sampling distribution for U. (Admittedly, we would need to change the column headings.) But I do **not** enjoy dividing by 252 and such divisions do not give pleasant decimals, so I won't do it. You get the idea without this extra pain. We will return to Kymn's sampling distribution in Chapter 5.

I will note that Kymn's actual $u = 7.2$ was the largest possible value of u; that is, every one of the other 251 possible assignments gives a value for u that is smaller than 7.2. This is no great surprise; Kymn's actual data had the five largest values on treatment 1 and the five smallest values on treatment 2. Thus, not a surprise, but as we will see, very important.

In Chapter 5 we will learn why the sampling distribution of a test statistic is important. But first we must deal with the following issue. When there are trillions or billions or even thousands of possible assignments, the method I used for Cathy's and Dawn's studies—basically, a tedious enumeration of every possibility—simply is not practical, even with the help of a clever computer program. Therefore, in Chapter 4 we will learn how to *approximate* a sampling distribution.

I end this section with the following useful result on symmetry.

Result 3.2 (Balance and symmetry.) *For a balanced CRD, the sampling distribution of U is symmetric around 0.*

Proof: *You don't need to read this proof, but it is so elegant that math-ophiles might enjoy it.*

Let a_1 denote any assignment. Let $-a_1$ denote its mirror image; i.e., the units assigned to treatment 1 [2] by a_1 are assigned to treatment 2 [1] by $-a_1$. Thus, for example, if $n = 6$, the mirror image of assignment 1,2,3 is assignment 4,5,6.

Let b be any positive number. Consider the collection, A, of assignments, if any, that give $u = b$. Let B denote the collection of assignments that are mirror images of the assignments in A.

Clearly, every assignment in B gives $u = -b$. As a result, $P(U = b) = P(U = -b)$ for every $b > 0$. Hence, the result follows.

The symmetry promised by Result 3.2 can be seen in Tables 3.5 and 3.7.

3.5 Computing

In this section you will learn how to use a website to perform randomization. Please click onto the following site:

`http://www.randomizer.org/form.htm`

If all is working well, you will still be able to see this page plus another window that has opened up and is at the above site. Let's look at your new web-window.

You will see that there are eight rectangular boxes. When you use this site to obtain an assignment via randomization you will be able to enter your *values of interest* in the top seven of these boxes. Currently, these seven boxes contain the *default values* provided by the site's programmer. After you have changed any, all or none of the entries in these seven boxes you should click on the bottom box, which contains the words:

Randomize Now!

My guess is that the exclamation point is to remind you how exciting it is to be doing Statistics! Personally, I am so excited to see how this works, I will click on the bottom box without any changes to the default settings. I did so and obtained:

11, 49, 18, 27, 50

You try it. I bet that you did not obtain the same numbers I obtained above.

Let's now look at the seven boxes and discover why they exist. The first box sits to the right of the following question:

How many sets of numbers do you want to generate?

As we shall see below, the "sets of numbers" are assignments. Thus, the question is: How many assignments do you want to generate? If I am a researcher seeking an assignment for my study, I would want only one assignment and I would not change the default value, which is 1. If a *Homework* question asks you to generate, say, five assignments, then you would place the number 5 in the box. Well, you could leave the 1 in the box and run the site five times, but that would be tedious and waste time.

The second box sits to the to the right of the question:

How many numbers per set?

For our purpose—obtaining an assignment for CRD—enter the value n_1 in the box.

The third and fourth boxes sit to the right of a directive to specify the range of numbers you desire. Calculate the value of $n = n_1 + n_2$, the total number of units in the CRD, and enter:

From: 1 (i.e., you may leave the default value)
To: n

The fifth box sits to the to the right of the question:

Do you wish each number in a set to remain unique?

For randomization we need n_1 *different* units selected for assignment to treatment 1. Thus, we opt for the default answer: Yes.

The sixth box sits to the to the right of the question:

Do you wish to sort the numbers that are generated?

For randomization we don't care the order in which the n_1 units are selected. Thus—and because it is easier to work with a sorted list—you should change the default to

Yes: Least to Greatest.

Finally, the seventh box sits to the to the right of the question:

How do you wish to view your random numbers?

I find the proffered *place markers* to be very annoying. Thus, I recommend using the default: Place Markers Off.

Problem 1 in the *Practice Problems* will provide you with practice using this site.

3.6 Summary

In a CRD, the researcher has the *authority* to assign units to levels of the study factor and *exercises this authority* by assigning units using the process of **randomization**. The three equivalent methods of **randomizing** are listed on page 57. The first two of these involve using a collection of *cards*. If you ever need to randomize (for example, in your own research or for a project in this class) it is fine to use cards, but most people find it more convenient to use an electronic method. In other words, the cards give us a way to *visualize*—if not perform—randomization. The electronic method was detailed in Section 3.5.

For a CRD with $n = n_1 + n_2$ units, of which n_1 are assigned to treatment 1, the number of possible assignments equals

$$\frac{n!}{n_1!n_2!},$$

where factorials are defined in Equation3.1 on page 56.

Probability is a quantification of uncertainty about the future. At this stage in these notes, we are interested in applying the idea of probability to the process of randomization. In particular, **before** we randomize, we believe that every possible assignment is equally likely to be the assignment chosen by randomization. Thus, we assign the same probability to every possible assignment:

$$P(\text{Any particular assignment}) = \frac{1}{\text{Total number of possible assignments}}.$$

An event is any specified collection of assignments. For any event A we calculate its probability as:

$$P(A) = \frac{\text{Number of assignments in } A}{\text{Total number of possible assignments}}.$$

Typically—but not always—in a CRD the two treatment means calculated from the data will be different numbers. Statisticians want to decide whether this observed difference between treatments is **real**. This is **not** an easy question to formulate or answer. The first step is to consider the Skeptic's Argument which states that the two treatments are the same in the following sense:

> For any given unit, if it were possible to assign the unit to both treatments, the two responses obtained would be identical.

In a CRD, because it is *impossible* to assign a unit to both treatments, a researcher cannot say, with certainty, whether the Skeptic's Argument is correct. In Chapter 5 we will learn how to examine the validity of the Skeptic's Argument using the ideas of a *statistical test of hypotheses*. Learning all the facets of a statistical test of hypotheses is a lengthy and difficult endeavor; as a result, we will break the process into manageable sized pieces. The first piece is the specification of the test statistic and the derivation of its sampling distribution.

We summarize the data from a CRD by computing the mean response for each treatment: $\bar{x}$ [$\bar{y}$] for treatment 1 [2]. Define $u = \bar{x} - \bar{y}$; u is the difference of the treatment means. The number u is called the **observed value of the test statistic** U.

One of the *big ideas* in Statistics is:

> We evaluate what **actually happens** in a study by comparing it to **everything that could have happened** in the study.

Statisticians take a rather modest view of the notion of **everything that could have happened**; this means the following. On the one hand we have the evidence from the actual data: the actual value of u. On the other hand, we have the distribution of the values of u. The distribution is obtained by looking at every possible assignment—including the assignment that was actually used—and for each assignment determining the value of u. As illustrated in the text, such a distribution cannot be obtained without an additional assumption. The assumption we will explore in the next few chapters is the assumption that the Skeptic is correct. With this assumption, the distribution is called the **sampling distribution of the test statistic** U.

In this chapter we have seen via two examples that if the number of possible assignments is *small*—an admittedly vague term—then the **exact sampling distribution** can be determined. In the next chapter, we will learn how a computer can be used to obtain an **approximate sampling distribution**. These ideas will come together in Chapter 5 when we learn how scientists and statisticians use either the exact or an approximate sampling distribution to investigate the issue of whether the observed treatment difference is real.

Much later in this Part I we will see that it is possible to obtain a distribution of values of u when the Skeptic is incorrect. This will allow us to investigate what is called the **power** of a test of hypotheses and we will learn why the power is very important to a scientist.

Result 3.2 on page 65 states that for a balanced CRD, the sampling distribution of U is symmetric around 0.

3.7 Practice Problems

1. Use the website discussed in Section 3.5 to obtain **five** assignments for Kymn's rowing study using the process of randomization. Recall that Kymn performed a balanced CRD with a total of 10 trials.

2. Table 3.8 presents artificial data for a balanced CRD with a total of $n = 6$ units.

   (a) What is the actual assignment that the researcher used?

   (b) Calculate the actual values of $\bar{x}$, $\bar{y}$ and $u = \bar{x} - \bar{y}$.

3. Refer to the previous question. It is too tedious for you to determine the entire sampling distribution for U, but I want to make sure you understand the steps. Thus, please answer the questions below.

   (a) Assuming that the Skeptic is correct, calculate the values of $\bar{x}$, $\bar{y}$ and $u = \bar{x} - \bar{y}$ for assignment 1,3,6.

   (b) Assuming that the Skeptic is correct, calculate the values of $\bar{x}$, $\bar{y}$ and $u = \bar{x} - \bar{y}$ for assignment 1,4,5.

   (c) Assuming that the Skeptic is correct, calculate the largest possible value of u; which assignment gives this value?

4. Table 3.9 presents an artificial data set with $n_1 = 2$ and $n_2 = 4$.

   (a) How many possible assignments are there?

   (b) List all possible assignments. Identify the actual assignment.

   (c) Calculate the actual values of $\bar{x}$, $\bar{y}$ and u.

   (d) Assuming that the Skeptic is correct, calculate the values of $\bar{x}$, $\bar{y}$ and $u = \bar{x} - \bar{y}$ for assignment 5,6.

   (e) Assuming that the Skeptic is correct, calculate the values of $\bar{x}$, $\bar{y}$ and $u = \bar{x} - \bar{y}$ for assignment 1,4.

   (f) The sampling distribution of U is given in Table 3.10.

      i. Is the sampling distribution of U symmetric? Explain your answer.

      ii. Obtain $P(U \geq 6)$; obtain $P(U \leq -6)$; obtain $P(|U| \geq 6)$.

Table 3.8: Artificial data for Practice Problems 2 and 3.

Unit:	1	2	3	4	5	6
Treatment:	2	1	1	2	1	2
Response:	9	6	30	3	12	15

Table 3.9: Artificial data for Practice Problem 4.

Unit:	1	2	3	4	5	6
Treatment:	2	2	1	2	1	2
Response:	8	4	4	0	20	12

3.8 Solutions to Practice Problems

1. My answer is below. You will no doubt obtain an answer different from mine.

   Below are my responses to the information requested by the website, some with a brief parenthetical explanation.

   - 5 (the number of assignments I requested); 5 (the value of n_1); From 1 To 10 (10 is the total number of trials in the study); Yes; Yes, Least to Greatest; and Place Markers Off.

   Below are the assignments I obtained:

   1, 3, 4, 8, 9
   3, 4, 5, 8, 9
   1, 4, 7, 9, 10
   1, 2, 3, 8, 9
   2, 3, 7, 8, 9

2. (a) From Table 3.8 we can see that units 2, 3 and 5 were assigned to treatment 1. Thus, the actual assignment is 2,3,5.

   (b) Again from Table 3.8,

   $$\bar{x} = (6 + 30 + 12)/3 = 16 \text{ and } \bar{y} = (9 + 3 + 15)/3 = 9.$$

Table 3.10: The sampling distribution of U for Practice Problem 4.

u	$P(U = u)$	u	$P(U = u)$	u	$P(U = u)$	u	$P(U = u)$
-9	$2/15$	-3	$3/15$	3	$2/15$	9	$1/15$
-6	$2/15$	0	$2/15$	6	$2/15$	12	$1/15$

Thus, $u = 16 - 9 = 7$.

3. (a) From Table 3.8,

$$\bar{x} = (9 + 30 + 15)/3 = 18 \text{ and } \bar{y} = (6 + 3 + 12)/3 = 7.$$

Thus, $u = 18 - 7 = 11$.

(b) Again from Table 3.8,

$$\bar{x} = (9 + 3 + 12)/3 = 8 \text{ and } \bar{y} = (6 + 30 + 15)/3 = 17.$$

Thus, $u = 8 - 17 = -9$.

(c) The six response values sum to 75. Thus, for any assignment

$$\bar{x} + \bar{y} = 75/3 = 25,$$

as illustrated above repeatedly. Thus,

$$u = \bar{x} - \bar{y} = \bar{x} - (25 - \bar{x}) = 2\bar{x} - 25.$$

We get the largest possible value of u by finding the largest possible value of $\bar{x}$. The largest possible value of $\bar{x}$ is $(30 + 12 + 15)/3 = 19$ which gives $\bar{y} = 25 - 19 = 6$. Thus, $19 - 6 = 13$ is the largest possible value of u. From Table 3.8, we see that this largest value of u is achieved by assignment 3,5,6.

4. (a) Using Equation 3.2 we obtain:

$$6!/[2!4!] = [6(5)]/2 = 15.$$

(b) As usual, I identify an assignment by listing the units that are assigned to the first treatment. I obtain:

1,2; 1,3; 1,4; 1,5; 1,6; 2,3; 2,4; 2,5; 2,6; 3,4; 3,5; 3,6; 4,5; 4,6; and 5,6.

From Table 3.9, the actual assignment is 3,5.

(c) From Table 3.9,

$$\bar{x} = (4 + 20)/2 = 12, \bar{y} = (8 + 4 + 0 + 12)/4 = 6 \text{ and } u = 12 - 6 = 6.$$

(d) From Table 3.9,

$$\bar{x} = (20 + 12)/2 = 16, \bar{y} = (8 + 4 + 4 + 0)/4 = 4 \text{ and } u = 16 - 4 = 12.$$

(e) From Table 3.9,

$$\bar{x} = (8 + 0)/2 = 4, \bar{y} = (4 + 4 + 20 + 12)/4 = 10 \text{ and } u = 4 - 10 = -6.$$

(f) i. No. There are many ways to justify this answer; here is mine. For symmetry (around any value, not just 0) the largest possible value of U, 12 in this case, must have the same probability as the smallest possible value of U, -9 in this case. They don't.

ii. The computations are below.

$$P(U \geq 6) = P(U = 6) + P(U = 9) + P(U = 12) = 2/15 + 1/15 + 1/15 = 4/15.$$

$$P(U \leq -6) = P(U = -6) + P(U = -9) = 2/15 + 2/15 = 4/15.$$

$$P(|U| \geq 6) = P(U \geq 6) + P(U \leq -6) = 4/15 + 4/15 = 8/15.$$

Table 3.11: Artificial data for Homework Problem 2.

Unit:	1	2	3	4	5
Treatment:	2	1	2	1	2
Response:	12	6	24	0	12

3.9 Homework Problems

1. Use randomization to obtain one assignment for each of the following situations.

   (a) $n_1 = n_2 = 12$.

   (b) $n_1 = 8$; $n_2 = 15$.

   (c) $n_1 = 15$; $n_2 = 8$.

2. Table 3.11 presents artificial data from a CRD with $n_1 = 2$ and $n_2 = 3$. Use these data to answer the questions below.

   (a) How many possible assignments are there?

   (b) List all possible assignments. Identify the actual assignment.

   (c) Calculate the actual values of $\bar{x}$, $\bar{y}$ and u.

   (d) Assuming that the Skeptic is correct, calculate the values of $\bar{x}$, $\bar{y}$ and u for assignment 1,4.

   (e) Assuming that the Skeptic is correct, calculate the values of $\bar{x}$, $\bar{y}$ and u for assignment 3,5.

Chapter 4

Approximating a Sampling Distribution

At the end of the last chapter, we saw how tedious it is to find the sampling distribution of U even when there are only 20 possible assignments. We also experienced the limit of my comfort zone: 252 possible assignments. For studies like Dawn's (184,756 possible assignments) and especially Sara's (1.075×10^{23} possible assignments) there are way too many possible assignments to seek an exact answer. Fortunately, there is an extremely simple way to obtain a good approximation—subject to the caveats given below—to a sampling distribution regardless of how large the number of possible assignments.

4.1 Two Computer Simulation Experiments

Let's return to Dawn's study. Our goal is to create a table for Dawn that is analogous to Table 3.7 on page 65 for Kymn's study; i.e. we want to determine the value of $u = \bar{x} - \bar{y}$ for every one of the 184,756 possible assignments. This is too big of a job for me!

Instead of looking at all possible assignments, we look at some of them. We do this with a **computer simulation experiment**. I wrote a computer program that selected 10,000 assignments for Dawn's study. For each selection, the program selected one assignment at random from the collection of 184,756 possible assignments. (You can visualize this as using our randomizer website 10,000 times.) For each of the 10,000 simulated assignments, I determined its value of $u = \bar{x} - \bar{y}$. My results are summarized in Table 4.1.

Suppose that we want to know $P(U = 0)$. By definition, it is the proportion of the (184,756) possible assignments that would yield $u = 0$. We do not know this proportion because we have not looked at all possible assignments. But we have—with the help of my computer—looked at 10,000 assignments; of these 10,000 assignments, 732 gave $u = 0$ (see Table 4.1). The relative frequency of $u = 0$ in the assignments we have examined is an intuitively obvious approximation to the relative frequency of $u = 0$ among all possible assignments. The relative frequency of $u = 0$ among all possible assignments is, by definition, $P(U = 0)$. To summarize, our approximation of the unknown $P(U = 0)$ is the relative frequency of assignments that gave $u = 0$; which is, from our table, 0.0732.

The above argument for $P(U = 0)$ can be extended to $P(U = u)$ for any of the possible

Table 4.1: The results of a computer simulation experiment with 10,000 replications (reps) for Dawn's study.

u	Freq.	Relative Freq.	u	Freq.	Relative Freq.	u	Freq.	Relative Freq.
-3.6	1	0.0001	-1.2	419	0.0419	1.2	394	0.0394
-3.4	1	0.0001	-1.0	506	0.0506	1.4	315	0.0315
-3.2	3	0.0003	-0.8	552	0.0552	1.6	251	0.0251
-3.0	4	0.0004	-0.6	662	0.0662	1.8	150	0.0150
-2.8	16	0.0016	-0.4	717	0.0717	2.0	108	0.0108
-2.6	25	0.0025	-0.2	729	0.0729	2.2	93	0.0093
-2.4	45	0.0045	0.0	732	0.0732	2.4	54	0.0054
-2.2	78	0.0078	0.2	765	0.0765	2.6	23	0.0023
-2.0	113	0.0113	0.4	716	0.0716	2.8	17	0.0017
-1.8	191	0.0191	0.6	674	0.0674	3.0	8	0.0008
-1.6	240	0.0240	0.8	553	0.0553	3.2	3	0.0003
-1.4	335	0.0335	1.0	507	0.0507			

Table 4.2: Selected probabilities of interest for Dawn's CRD and their approximations.

Probability of Interest	Its Approximation
$P(U \geq 2.2)$	r.f. $(U \geq 2.2) = 0.0198$
$P(U \leq 2.2)$	r.f. $(U \leq 2.2) = 1 - 0.0105 = 0.9895$
$P(\|U\| \geq 2.2)$	r.f. $(\|U\| \geq 2.2) = 0.0198 + 0.0173 = 0.0371$

values u. But it will actually be more interesting to us to approximate probabilities of more complicated events than $P(U = u)$. In particular, recall that Dawn's actual u was 2.2. As we will see in Chapter 5, we will be interested in one or more of the probabilities given in Table 4.2. Let me give you some details on how the answers in this table were obtained.

To obtain r.f. $(U \geq 2.2)$ we must sum the relative frequencies for the values 2.2, 2.4, 2.6, 2.8, 3.0 and 3.2. From Table 4.1, we obtain

$$0.0093 + 0.0054 + 0.0023 + 0.0017 + 0.0008 + 0.0003 = 0.0198.$$

For r.f. $(U \leq 2.2)$ note that this value is $1-$ r.f. $(U > 2.2) = 1 - 0.0105$. Finally, r.f. $(|U| \geq 2.2)$ is the sum of two relative frequencies: $(U \geq 2.2)$ and $(U \leq -2.2)$. The first of these has been found to equal 0.0198. The second of these is

$$0.0078 + 0.0045 + 0.0025 + 0.0016 + 0.0004 + 0.0003 + 0.0001 + 0.0001 = 0.0173.$$

Table 4.3: Selected probabilities of interest for Sara's CRD and their approximations.

Probability of Interest	Its Approximation
$P(U \geq 8.700)$	r.f. $(U \geq 8.700) = 0.0903$
$P(U \leq 8.700)$	r.f. $(U \leq 8.700) = 0.9107$
$P(\|U\| \geq 8.700)$	r.f. $(\|U\| \geq 8.700) = 0.0903 + 0.0921 = 0.1824$

Adding these we get,

$$\text{r.f. } (U \geq 2.2) + \text{r.f. } (U \leq -2.2) = 0.0198 + 0.0173 = 0.0371.$$

Next, I performed a computer simulation experiment for Sara's CRD. As stated earlier, there are more than 10^{23} different assignments for a balanced study with $n = 80$ total trials. Trying to enumerate all of these would be ridiculous, so we will use a computer simulation with 10,000 reps.

My simulation study yielded 723 distinct values of u! This is way too many to present in a table as I did for Dawn's study. (The simulation study for Dawn, recall, yielded 35 distinct values for u and that was unwieldy.)

Recall that for Sara's data

$$\bar{x} = 106.875 \text{ and } \bar{y} = 98.175, \text{ giving } u = 8.700.$$

Table 4.3 presents information that will be needed in Chapter 5.

4.2 How Good are These Approximations?

Tables 4.2 and 4.3 present six unknown probabilities and their respective approximations based on simulation experiments with 10,000 reps. Lacking knowledge of the exact probabilities I cannot say exactly how good any of these approximations are. What I can say, however, is that each of them is *very likely* to be *very close* to the exact (unknown) probability it is approximating. How can I know this? Well, we will see *how* later in this course when we learn about **confidence intervals**, so you will need to be patient.

And, of course, the terms *very likely* and *very close* are quite vague. Here is what we will do for now. First, the expression *very close* will be replaced by a specific number, call it h, that is computed from our simulation results. In words, it is very likely that the simulation study approximation is within h of its exact probability. (If this is confusing, see the numerical examples below.)

Next, the expression *very likely* will be replaced by *nearly certain*. I know what you are thinking: *nearly certain* also is vague, but I hope that you feel that it is somehow *more encouraging* than *very likely*. As we will learn later (for those who can't stand the suspense!) *nearly certain*

corresponds to what we will call being 99.73% **confident**. (Alas, I really can't tell you exactly what this means until we study confidence intervals.)

Here is what we do. Let m denote the number of reps in our simulation experiment; recall that $m = 10{,}000$ for our two studies. Let r denote any (unknown) probability that interests us. Let $\hat{r}$ denote the relative frequency approximation of r. For example, in Dawn's study $\hat{r} = 0.0198$ is our approximation to the unknown $r = P(U \geq 2.2)$. The **nearly certain interval** for r is given by the following:

$$\hat{r} \pm 3\sqrt{\frac{\hat{r}(1-\hat{r})}{m}} \tag{4.1}$$

Another way to say this is that we are nearly certain that $\hat{r}$ is within h of r, where

$$h = 3\sqrt{\frac{\hat{r}(1-\hat{r})}{m}}.$$

We will evaluate this interval for each of our six approximations. You don't need to verify these computations, but you will be asked to do similar computations for homework.

- For $\hat{r} = 0.0198$, the nearly certain interval for r is

$$0.0198 \pm 3\sqrt{\frac{0.0198(0.9802)}{10000}} = 0.0198 \pm 0.0042 = [0.0156, 0.0240].$$

- For $\hat{r} = 0.9895$, the nearly certain interval for r is

$$0.9895 \pm 3\sqrt{\frac{0.9895(0.0105)}{10000}} = 0.9895 \pm 0.0031 = [0.9864, 0.9926].$$

- For $\hat{r} = 0.0371$, the nearly certain interval for r is

$$0.0371 \pm 3\sqrt{\frac{0.0371(0.9629)}{10000}} = 0.0371 \pm 0.0057 = [0.0314, 0.0428].$$

- For $\hat{r} = 0.0903$, the nearly certain interval for r is

$$0.0903 \pm 3\sqrt{\frac{0.0903(0.9097)}{10000}} = 0.0903 \pm 0.0086 = [0.0817, 0.0989].$$

- For $\hat{r} = 0.9107$, the nearly certain interval for r is

$$0.9107 \pm 3\sqrt{\frac{0.9107(0.0893)}{10000}} = 0.9107 \pm 0.0086 = [0.9021, 0.9193].$$

- For $\hat{r} = 0.1824$, the nearly certain interval for r is

$$0.1824 \pm 3\sqrt{\frac{0.1824(0.8176)}{10000}} = 0.1824 \pm 0.0116 = [0.1708, 0.1940].$$

There is a surprising feature to Formula 4.1. The feature is hidden, so it is easy to overlook. The surprising feature is that the total number of possible assignments (184,756 for Dawn, more than 10^{23} for Sara) does **not** appear in the formula! All that matters is the number of reps, m, in the simulation experiment. In my experience, many people believe that precision is a function of the *percentage* of objects examined. For our current problem, the percentage of assignments examined does not matter; it's the number of assignments examined that matters.

We can improve the precision of a nearly certain interval by making it narrower, which we can achieve by increasing the value of m. Personally, I think that the precision of the nearly certain interval with $m = 10{,}000$ reps is fine unless $\hat{r}$ is very close to 0. (See more on this below.) In order to show you why I feel this way, I did another simulation experiment for Dawn's study, this time with $m = 90{,}000$ reps. Combining my two experiments I have $m = 100{,}000$ reps and I will use the results of the 100,000 reps to recalculate two of my nearly certain intervals for Dawn's study.

As an example, suppose I am interested in $r = P(U \geq 2.2)$. Of the 100,000 reps, 1,879 assignments gave $u \geq 2.2$. Thus, the approximate probability is

$$\hat{r} = 1879/100000 = 0.01879.$$

The nearly certain interval for the exact probability is

$$0.01879 \pm 3\sqrt{\frac{0.01879(0.98121)}{100000}} = 0.01879 \pm 0.00129 = [0.01750, 0.2008].$$

Note that the value of h for this interval is 0.00129, compared to $h = 0.0042$ for our simulation with 10,000 reps. Also, of the 100,000 reps, 1,792 assignments gave $u \leq -2.2$. Thus, the approximation for $P(|U| \geq 2.2)$ is

$$0.01879 + 0.01792 = 0.03671.$$

The nearly certain interval for the exact probability is

$$0.03671 \pm 3\sqrt{\frac{0.03671(0.96329)}{100000}} = 0.03671 \pm 0.00178 = [0.03493, 0.03849].$$

In my experience, the increase in precision does not justify running my computer 10 times longer than required for 10,000 reps. (These reps do take time and consume electricity!)

4.3 A Warning about Simulation Experiments

The six nearly certain intervals above for the six probabilities of interest suggest that with a simulation experiment consisting of 10,000 reps we can obtain a very precise approximation to an exact probability. There is a caveat I need to mention.

With a 10,000 rep simulation experiment, if $\hat{r}$ is ≤ 0.0050 or ≥ 0.9950 then you should **not** compute the nearly certain interval. For other values of $m \geq 1000$, do not compute the nearly certain interval if $\hat{r}$ is $\leq 50/m$ or $\geq 1 - (50/m)$. If you ignore my directive and go ahead and

calculate the nearly certain interval, it might indeed contain the exact probability of interest. (And if it does, we won't know it.) The difficulty is that we can no longer trust the modifier *nearly certain*. As an extreme example, suppose that $\hat{r} = 0$, which means that the event of interest never occurred in the simulation experiment. If you plug this value into Formula 4.1 the nearly certain interval becomes the single number 0. Whereas the probability of an event A can be very very small, as long as A contains at least one of the possible assignments, it can never be 0.

As we will learn later in these notes, if $\hat{r}$ does equal 0, we can be nearly certain that $r \leq 5.92/m$. There are other results for very small non-zero values of $\hat{r}$ and we will learn how to find them later.

In many applications—though not all, as we shall see— if $\hat{r} \leq 0.0050$, then the researcher is happy with the conclusion that r is very small and is not very concerned with having a precise nearly certain interval.

4.4 Computing

I performed the simulation studies reported in this chapter by using the statistical software package **Minitab**. I could find no website that allows me to perform a **full-blown** simulation study, by which I mean a simulation that gives me all m (recall, usually m = 10,000) values of the target statistic, in the current situation, u. Later I will show you a website that does allow us to perform a simulation study, but it gives only two relative frequencies as output. It turns out that these are two *very useful* relative frequencies, so the website will be valuable to us.

I have one other comment about our simulation studies. Some of you may have been wondering about this issue. Note that I have **never claimed** that my simulation examines m **different** assignments. Indeed, each rep of my simulation selects an assignment at random without caring about which assignments have already been examined. I do this for the following reasons.

1. A program that keeps track of the assignments already examined would be much more difficult to write, it would require more computer storage space and would require much more computer time to execute. (If you have any experience writing programs, you probably agree that these three claims are believable.)

2. The **possibility** that my simulation experiment **might** examine some assignments more than once creates **no bias** in my answers; it simply makes my answers a little bit less efficient; i.e., the nearly certain interval would change ever so slightly—becoming narrower—if I guaranteed m different assignments are selected.

3. I don't want you to waste time worrying about the validity of item 2 immediately above; we will revisit this issue when we discuss population-based inference in *Part II* of these *Course Notes*.

4. In view of items 1 and 2 in this list, if you want a more precise approximation to an unknown probability, increasing the value of m is a much smarter choice than writing a new computer program.

4.5 Summary

In Chapter 3 we learned about the exact sampling distribution for the test statistic U. In Chapter 5 we will learn why we need this sampling distribution to analyze our data with a statistical test of hypotheses. In this current chapter we learned an important technique for approximating a sampling distribution: a computer simulation experiment.

The sampling distribution calculates the value of u for *all possible assignments*. A computer simulation experiment calculates the value of u for *some possible assignments*. The number of assignments examined by a computer simulation experiment is called its **number of reps** and is denoted by m. In these notes, m will usually be 10,000; on some occasions it will equal 100,000; and on a few rare occasions it will equal 1,000. The assignment for any rep is selected at random from the collection of all possible assignments, for example, by using the randomizer website. As will be discussed later in these notes, you should realize that a computer simulation experiment need not select m *distinct* assignments.

After performing a computer simulation experiment, the analyst creates a listing of all values of u that were obtained in the experiment and the relative frequency of occurrence of each value of u. An example of such a table is given in Table 4.1 on page 76 for Dawn's study of her cat. We may use this table to approximate the exact and unknown $P(U = u)$ for all possible values u of the test statistics. The approximation is simply the relative frequency of occurrence of $(U = u)$ in the computer simulation experiment.

In our applications of these ideas in Chapter 5 and later in these notes, we will usually be interested in events that are more complicated than simply $(U = u)$. To this end, let A denote any event which involves the value of the test statistic U. Let r denote the exact probability of the event A. Let $\hat{r}$ be the relative frequency of the event A in the computer simulation experiment. We view $\hat{r}$ as our approximation of r.

For $m \geq 1000$, if $\hat{r}$ satisfies the following inequality

$$50/m < \hat{r} < 1 - (50/m),$$

then one can compute the nearly certain interval for r, given below:

$$\hat{r} \pm 3\sqrt{\frac{\hat{r}(1-\hat{r})}{m}}.$$

We are *nearly certain* that this interval of numbers includes the true, unknown, value of r. An equivalent way to interpret this result is:

> We are *nearly certain* that $\hat{r}$ is within h of r, where
>
> $$h = 3\sqrt{\frac{\hat{r}(1-\hat{r})}{m}}.$$

The meaning of the modifier *nearly certain* will be explored later in these notes. Also, the situations in which $\hat{r} \leq 50/m$ or $\hat{r} \geq 1 - (50/m)$ will be considered later in these notes.

4.6 Practice Problem

1. In the dart game *301* the object is to be the first player to score *exactly* 301 points. In each round a player throws three darts and the total score of the three darts is added to the player's score at the end of the previous round. If the new total is greater than 301, the player's score reverts to the total at the end of the previous round. Thus, for example, if a player reaches a total of 300 at the end of a round, then the player will need exactly one point on the next round to win; any larger score will be ignored. Doug performed a balanced CRD with $n = 40$ trials to compare his *personal darts* (treatment 1) to *bar darts* (treatment 2). The response was the number of rounds Doug required to score exactly 301. The sorted responses with Doug's darts are:

12	13	14	14	15	15	17	18	18	19
19	19	20	20	21	21	22	23	25	27

The sorted responses with the bar darts are:

13	15	16	16	17	17	17	18	19	21
21	22	23	25	26	26	27	27	28	30

I obtained the following summary statistics for Doug's data:

$$\bar{x} = 18.60 \text{ and } \bar{y} = 21.20, \text{ giving } u = 18.60 - 21.20 = -2.60.$$

Note that smaller response values are preferred.

I performed a simulation experiment with 10,000 reps. Each rep yielded a possible observed value u of the test statistic U. I won't show you all of my results, but I will tell you the following two relative frequencies:

$$\text{r.f. } (U \leq -2.60) = 0.0426; \text{ and r.f. } (U \geq 2.60) = 0.0418.$$

Use these results to answer the questions below.

(a) What is our approximation for $r = P(U \leq -2.60)$?

(b) Compute the nearly certain interval for r in (a).

(c) What is our approximation for $r = P(|U| \geq 2.60)$?

(d) Compute the nearly certain interval for r in (c).

4.7 Solution to Practice Problem

1. (a) The approximation is $\hat{r} = 0.0426$.

   (b) The nearly certain interval is

$$0.0426 \pm 3\sqrt{\frac{0.0426(0.9574)}{10{,}000}} = 0.0426 \pm 0.0061 = [0.0365, 0.0487].$$

   (c) The approximation is

$$\hat{r} = 0.0426 + 0.0418 = 0.0844.$$

   (d) The nearly certain interval is

$$0.0844 \pm 3\sqrt{\frac{0.0844(0.9156)}{10{,}000}} = 0.0844 \pm 0.0083 = [0.0761, 0.0927].$$

4.8 Homework Problems

1. In the Chapter 1 Homework you learned about Reggie's study of darts. Reggie performed a balanced CRD with a total of $n = 30$ trials. Dot plots of his data are presented in Figure 1.4. It can be shown that the means for Reggie's data are:

$$\bar{x} = 201.533 \text{ and } \bar{y} = 188, \text{ giving } u = 201.533 - 188 = 13.533.$$

   I performed a simulation experiment with 10,000 reps. Each rep yielded a possible observed value u of the test statistic U. I won't show you all of my results, but I will tell you the following two relative frequencies:

$$\text{r.f. } (U \geq 13.533) = 0.0058; \text{ and r.f. } (U \leq -13.533) = 0.0039.$$

   Use these results to answer the questions below.

   (a) What is our approximation for $r = P(U \geq 13.533)$?

   (b) Compute the nearly certain interval for r in (a).

   (c) What is our approximation for $r = P(|U| \geq 13.533)$?

   (d) Compute the nearly certain interval for r in (c).

2. In the Chapter 1 Homework you learned about Brian's study of running. Brian performed a balanced CRD with a total of $n = 20$ trials. Dot plots of his data are presented in Figure 1.3. It can be shown that the means for Brian's data are:

$$\bar{x} = 333.0 \text{ and } \bar{y} = 319.5, \text{ giving } u = 333.0 - 319.5 = 13.5.$$

   I performed a simulation experiment with 100,000 reps. Each rep yielded a possible observed value u of the test statistic U. I won't show you all of my results, but I will tell you the following two frequencies:

$$\text{freq. } (U \geq 13.5) = 56; \text{ and freq. } (U \leq -13.5) = 45.$$

   Use these results to answer the questions below. Note that the number of reps is 100,000, not the usual 10,000.

   (a) What is our approximation for $r = P(U \geq 13.5)$?

   (b) Compute the nearly certain interval for r in (a).

   (c) What is our approximation for $r = P(|U| \geq 13.5)$?

   (d) Compute the nearly certain interval for r in (c).

Chapter 5

A Statistical Test of Hypotheses

Chapter 3 introduced the Skeptic's Argument. Also in Chapter 3, we learned that if we assume the Skeptic is correct, then there is a computable sampling distribution for a test statistic. *Computable* is a bit optimistic; as we have seen, I can obtain the exact sampling distribution only for very small studies. In Chapter 4 we learned how to interpret the results of a computer simulation experiment. In particular, we found that we can approximate exact probabilities by using relative frequencies. Furthermore, by calculating the nearly certain interval, we can get an idea of the precision of our approximations.

There are three main areas of statistical inference that scientists use: prediction, estimation and tests of hypotheses. In this chapter you will learn how to use the ideas of Chapters 3 and 4 to perform a test of hypotheses for the CRDs introduced in Chapters 1 and 2 .

5.1 Step 1: Choice of Hypotheses

I will introduce the ideas of this chapter first for Dawn's study of her cat Bob.

I introduced the Skeptic's Argument in Chapter 3; it is repeated below:

> The flavor of the treat is irrelevant. The number of treats that Bob consumed on any given day was determined by how hungry Bob was on that day.

It will be useful to invent an adversary for the Skeptic; I will call it the Advocate. The Advocate believes that the flavor of the treat *matters*; it is *not irrelevant*. It will, however, require some work to understand exactly what the Advocate believes.

After arguing back-and-forth for some time, the Skeptic and Advocate state their positions concisely:

- **The Skeptic:** Flavor does not matter. The difference in means, $\bar{x} - \bar{y} = 5.1 - 2.9 = 2.2$ treats, is meaningless. It was just **by chance** that Bob ate more treats on the chicken days than he did on the tuna days.
- **The Advocate:** Flavor matters. The difference in means, 2.2 treats, is **too large to reasonably be attributed to chance**.

For a statistician, the debate between the Skeptic and the Advocate is evaluated by performing a **statistical test of hypotheses.** We will spend some time putting Dawn's study into this context.

I conjecture that most of you are familiar with the word hypotheses from your work in science, but I will present these ideas as if they are totally new to you.

First, hypotheses is the plural of the noun hypothesis. A hypothesis is a conjecture about *the way things are*. There are always two hypotheses in our approach. (It is possible to have more than two hypotheses, but this occurs only rarely in practice.) The two hypotheses are:

- The **null hypothesis**, denoted by H_0; and
- The **alternative hypothesis**, denoted by H_1.

These hypotheses do not overlap; hence, they cannot both be true. It is almost always possible to argue that both hypotheses could be false, but—for the most part—scientists don't dwell on this fact.

Before we conduct a study, we don't know which of these hypotheses is correct. One approach for dealing with this uncertainty is to assign probabilities. For example, we might begin a study by saying, "I believe that the probability is 70% that the null hypothesis is true and 30% that the alternative is true." Then after we collect and analyze the data, we reassess these probabilities. This approach is called the **Bayesian approach to tests of hypotheses**. Historically, the Bayesian approach has not been very popular with statisticians, but it is becoming more and more popular.

The approach that we will follow can be described quite easily: We assume the null hypothesis is correct (or true). This will allow us to analyze our data. As a result of our analysis we will reach one of two possible decisions:

- We will decide that there is enough evidence in our data to switch our *allegiance* from the null to the alternative; this is referred to as **rejecting the null hypothesis**.
- We will decide that the evidence in our data is **not sufficiently strong** to warrant switching our allegiance from the null to the alternative; this is referred to as **failing to reject the null hypothesis**.

Stating the obvious, note that we do **not** treat the hypotheses in a symmetrical manner: We begin our analysis by assuming that the null hypothesis is correct.

For all the studies in this chapter, including Dawn's study, we take the null hypothesis to be that the Skeptic is correct. The alternative hypothesis is that the Advocate is correct, but, surprisingly, we consider three different ways the Advocate can be correct. Thus, it will require some care to specify the alternative hypothesis.

If you are a supporter of the Advocate, or if you just think that life should be fair whenever possible, the stipulation in the previous paragraph is strange. Why do we begin the analysis by assuming the Skeptic is correct? We are actually following a very popular *principle* of science, **Occam's Razor**. If you are interested in it, I encourage you to read about Occam's Razor on the web or elsewhere. For our purposes, Occam's Razor states that whenever there are two competing theories that are similar in their ability to explain a phenomenon, we should prefer the simpler theory. The Skeptic's Argument, that flavor does not matter, is considered to be simpler than the

notion that flavor matters. (If flavor matters we wonder: Why does it matter? *and* How much does it matter?) Thus, following Occam's Razor, we assume that flavor does not matter unless and until experimentation tells us otherwise.

Remember that we are still at the planning stage of our study. We have decided that our null hypothesis is that the Skeptic is correct.

As discussed in Chapter 3, we develop the sampling distribution of a test statistic by examining every possible assignment. In other words, we consider lots of studies that never were performed. We do something similar now.

Recall that in a CRD we use randomization to assign n_1 units to treatment 1 and n_2 units to treatment 2 and have a total of $n = n_1 + n_2$ units. Imagine the following alternative design for a study: Assign all n units to the first treatment. Call this the **All Treatment-1** (**AT-1**) study. Similarly, the **All Treatment-2** (**AT-2**) study would assign all n units to the second treatment. If the researcher wants to learn about a single treatment, then an AT- study on that treatment would be great. But if, as is our situation, the researcher wants to *compare* treatments, these AT- studies are very bad. Very bad, but useful for our immediate purposes.

Imagine that the researcher performs the AT-1 study; let μ_1 denote the mean of the responses from the n units. Similarly, imagine that the researcher performs the AT-2 study; let μ_2 denote the mean of the responses from the n units. Let's consider an example.

In Dawn's study, there were $n = 10 + 10 = 20$ units (or trials or days). If the AT-1 study had been performed, then Dawn would know the value of μ_1; it would just be the mean of her 20 observations. Similarly, if Dawn had performed the AT-2 study, then she would know the value of μ_2. If Dawn could have performed both the AT-1 and AT-2 studies then she would know the values of μ_1 and μ_2; this would allow her to determine definitively which flavor of treats yielded greater consumption by Bob.

Of course it was impossible for Dawn to have performed **both** the AT-1 and AT-2 studies. Allow me to be fanciful for a moment. Assume that on each day Dawn could immediately create a clone of Bob. She would then have two Bobs for the trial and could indeed assign each treatment to a Bob. If she did this **clone-enhanced** study every day, it would be equivalent to performing both of the AT- studies. Of course, I am being fanciful because cloning of the kind I am describing is not possible. (Even if it were possible, there would be major ethical issues of what to do with all the extra Bobs!)

Henceforth, when I say I have a clone-enhanced study, it means that both of the AT- studies are performed and, hence, that the values of both μ_1 and μ_2 are known to me. While it is impossible to have a clone-enhanced study, this bit of make-believe, as we will see, is quite useful for the development of our theory and methods.

5.1.1 An Artificial Study

Consider our studies of Chapters 1 and 2 again. In Sara's golf study, we will be learning how statisticians decide whether the Skeptic is correct or incorrect. With a total of $n = 80$ trials in the study, I find that the idea of the Skeptic being correct or incorrect is a bit overwhelming to visualize. Therefore, I want to further explore the topic of the Skeptic being correct or incorrect in a very small study. How small? I will create a balanced study with $n = 6$ treatments. I will refer

Table 5.1: Headache Study-1 (HS-1). The response is the time, in minutes, required for headache relief. Note that smaller values of the response are better.

Subject:	1	2	3	4	5	6
Treatment:	1	2	1	2	2	1
Response:	18	3	24	12	6	15

to the study as the Headache Study-1 (HS-1); its data are given in Table 5.1. Let me describe in detail HS-1.

The researcher wants to compare two active drugs that are thought to be effective for speedy relief of minor headache pain; call the drugs A and B. The study is on six subjects. Each subject is a person who chronically suffers mild tension headaches, not migraines. The six subjects are assigned labels: 1, 2, ..., 6. The subjects are assigned to drugs by randomization: the result being that subjects 1, 3 and 6 are assigned to drug A (treatment 1) and subjects 2, 4 and 5 are assigned to drug B (treatment 2). Each subject is given the following instructions:

> The next time you suffer from mild headache pain, immediately take the pill (drug) you have been given. Record the length of time, to the nearest minute, until your headache pain begins to diminish.

The response, of course, is the number of minutes reported by the subject. Note that the smaller the value of the response, the better.

Our earlier *named-studies*, performed by Dawn, Kymn, Sara and Cathy, were real-life studies and the data I use match the values the students reported to me. On occasion, it will be easier for me—yes, I admit it!—to use artificial data rather than search for real data that illustrate a particular point I want to make. Especially when I am trying to motivate how statisticians think, it is convenient to divorce my presentation from a real scientific problem. In the current situation I could use Cathy's data, but I prefer artificial data so that we don't get bogged down in the arithmetic. Also, I want to introduce the idea of studying headaches, because this topic will help motivate our later work with paired data. This latter reason explains why I have numbered this study '1;' we will have other artificial headache studies in this course.

5.1.2 What if the Skeptic is Correct?

Let's take a quick look at the data from HS-1 in Table 5.1. The sorted data on the first treatment are: 15, 18 and 24 which gives $\bar{x} = 19$. The sorted data on the second treatment are: 3, 6 and 12 which gives $\bar{y} = 7$. There is separation of the responses, just as we found in Kymn's study of rowing: all subjects on treatment 1 gave larger (worse) responses than all subjects on treatment 2. Table 5.2 rewrites the data in anticipation of the clone enhanced study. Look at this table. There are 12 spaces for the 12 responses that would be obtained in the clone-enhanced version of HS-1. Six of these responses are known to us, but each of the other six possible responses (denoted by ?'s in the table) could be any nonnegative integer.

Table 5.2: The clone-enhanced version of HS-1 without further assumptions.

Subject:	1	2	3	4	5	6	Means
Response on Treatment 1:	18	?	24	?	?	15	$\bar{x} = 19$
Response on Treatment 2:	?	3	?	12	6	?	$\bar{y} = 7$

Next, let us assume that the Skeptic is correct. On this assumption, the clone-enhanced study would yield the data given at the top of Table 5.3 as Case 1. (Note to the reader: Don't simply read the previous sentence and then proceed to the next sentence below! You should look at Case 1 carefully. Note that for each of the six subjects, the response on treatment 1 is *exactly the same as* the response on treatment 2. This is what it means for the Skeptic to be correct.) Let's look at the data in this table. We can see from this table that $\mu_1 = \mu_2$ and this common value is

$$(18 + 3 + 24 + 12 + 6 + 15)/6 = 78/6 = 13.$$

Later in these notes we will learn about **estimation**. In a CRD, we will view $\bar{x}$ as our **point estimate** of μ_1 and $\bar{y}$ as our point estimate of μ_2. (Here is the idea behind the term *point estimate*: we say point because it is a single number and we say estimate because, well, statisticians refer to this endeavor as *estimation*; more on this later in these notes.)

On the assumption that the Skeptic is correct in HS-1, we see that the point estimate ($\bar{x} = 19$) of $\mu_1 = 13$ is much too large and the point estimate ($\bar{y} = 7$) of $\mu_2 = 13$ is much too small. In other words, the actual study suggested a big difference in treatments ($19 - 7 = 12$) when, in fact, the treatments have an identical effect on the response.

In a real study, of course, a scientist would not know whether the Skeptic is correct; other possibilities are shown in Cases 2–5 in Table 5.3. These four possibilities are examined in detail in the following subsection.

5.1.3 Four Examples of the Skeptic Being Incorrect

In this subsection I will discuss many important features of Cases 2–5 in Table 5.3.

The first thing to note is that, indeed, as I claim, the Skeptic is incorrect for the scenarios in Cases 2–5. We can see this easily by looking at the responses for subject 1 in each case. For every case, the response of subject 1 differs under the two treatments; for example, in Case 2 subject 1 gives a response of 18 under treatment 1 and a response of 12 under treatment 2. Because the Skeptic's argument is that treatment does not matter for *any* subject, by finding even one subject for which treatment matters, we have established that the Skeptic is incorrect.

Next, note that there are many ways that the Skeptic can be incorrect. Why? We can imagine the six question marks in Table 5.2 being replaced by *any number*. Of all these possible replacements, only one coincides with the Skeptic being correct. All other combinations of choices makes the Skeptic incorrect.

Note: My use of *any number* in the previous paragraph is a bit inaccurate. For a bounded count response, as in Dawn's cat treat study, there are only a finite number of possibilities for the

Table 5.3: A number of possibilities for the responses to the clone-enhanced version of HS-1. In each case below, the actual data are in bold-faced type.

Case 1: The Skeptic is Correct.

Subject:	1	2	3	4	5	6	μ_i
Response on Treatment 1:	**18**	3	**24**	12	6	**15**	13
Response on Treatment 2:	18	**3**	24	**12**	**6**	15	13

Case 2: The clone-enhanced version of HS-1 with a constant treatment effect of $c = 6$.

Subject:	1	2	3	4	5	6	μ_i
Response on Treatment 1:	**18**	9	**24**	18	12	**15**	16
Response on Treatment 2:	12	**3**	18	**12**	**6**	9	10

Case 3: A possible clone-enhanced version of HS-1 under the assumption that the Skeptic is incorrect.

Subject:	1	2	3	4	5	6	μ_i
Response on Treatment 1:	**18**	24	**24**	9	6	**15**	16
Response on Treatment 2:	6	**3**	15	**12**	**6**	18	10

Case 4: The clone-enhanced version of HS-1 with a constant treatment effect of $c = -3$.

Subject:	1	2	3	4	5	6	μ_i
Response on Treatment 1:	**18**	0	**24**	9	3	**15**	11.5
Response on Treatment 2:	21	**3**	27	**12**	**6**	18	14.5

Case 5: A possible clone-enhanced version of HS-1 under the assumption that the Skeptic is incorrect.

Subject:	1	2	3	4	5	6	μ_i
Response on Treatment 1:	**18**	9	**24**	12	3	**15**	13.5
Response on Treatment 2:	21	**3**	21	**12**	**6**	18	13.5

numbers to use in replacing the question marks. For Dawn's study, there are 11 possible choices (the integers 0, 1, 2, ..., 10) for replacing each question mark. With 20 such substitutions—one for each trial—there are $11^{20} = 6.73 \times 10^{20}$ possible combinations of substitutions. And *exactly one of these* makes the Skeptic correct. Thus, I hope that you will concede the following point: For a CRD with a numerical response, there are a great many possible ways to substitute numbers for the question marks **and** only one of these substitutions makes the Skeptic correct.

If you concede my point at the end of the previous paragraph, it will be believable (and, in fact, it is true) that we cannot possibly hope to learn exactly *how* the Skeptic is incorrect, if indeed it is incorrect. This situation has two consequences of great importance to us:

- We will spend most of our effort examining the consequences of assuming that the Skeptic is correct. (But not *all* of our effort; see the later material on **the power of a test of hypotheses**.)
- We will focus on the values of μ_1 and μ_2. These values are important because they tell us (if we only knew what they were!) how each treatment would perform if it was assigned to every unit. Thus, for example, if smaller responses are preferred, then deciding, say, that $\mu_1 < \mu_2$ could be interpreted by saying that **overall treatment 1 is better than treatment 2.**

Let's look at Cases 2–5 now.

Case 2 is a pretty wonderful situation for a scientist. But it does require care to see why. If you look at the row means, you find $\mu_1 = 16$ and $\mu_2 = 10$ which tells you that, on average, headache relief on treatment 2 is six minutes faster than headache relief on treatment 1. But we can actually say something much stronger. If you look at Case 2 carefully, you will notice that *for every subject* the response on treatment 1 is *exactly* six minutes larger than the response on treatment 2. Thus, it is not simply the situation that treatment 2 is six minutes faster, on average, it is six minutes faster *for every subject*! Case 2 motivates the following definition.

Definition 5.1 (The constant treatment effect.) *In a clone-enhanced study, suppose that the response on treatment 1 minus the response on treatment 2 equals the nonzero number c* **for every unit**. *In this situation we say that the treatment has a* **constant treatment effect** *equal to c.*

Note the following about this definition. Case 2 has a constant treatment effect equal to 6 minutes because, for every subject, the subtraction yields $c = 6$ minutes. We require c to be nonzero because a constant treatment effect with $c = 0$ is just another way of saying that the Skeptic is correct.

Next, let's look at Case 3. Cases 2 and 3 have the same row means, which tells us that, on average, relief with treatment 2 is six minutes faster than with treatment 1 in both cases. But now look at the individual subjects in Case 3. For subjects 1–3, treatment 2 is better than treatment 1 and better by more than six minutes. For subjects 4 and 6, treatment 1 is actually better than treatment 2. Finally, for subject 5 the two treatments give the same response.

In short, Cases 2 and 3 are very different. In Case 2, a physician could tell the group of six subjects,

> All of you should take treatment 2. For every one of you relief will come six minutes faster with treatment 2 than with treatment 1.

Contrast this with what the physician would say in Case 3. I am assuming, of course, that the physician does not know which people are which subjects. The physician would say,

> I recommend that all of you take treatment 2. On average for the six of you, relief will come six minutes faster with treatment 2 than with treatment 1. To be honest, I must admit that for two of you, treatment 1 is actually better than treatment 2; and, for one of you, the treatments have exactly the same effect.

Now I must tell you the disappointing news. For a CRD, it is impossible to distinguish between Cases 2 and 3. This doesn't bother me a great deal because I know the following to be true.

> There is no single study that will answer every possible question of interest. And this remains true whether you are a good statistician or not.

Thus, you might ask, why do I mention this issue? Why mention the difference between Cases 2 and 3 if I cannot help you distinguish between them? Because as a scientist you should always be seeking a better understanding of the world. One study, CRD or otherwise, is not the ultimate goal in one's career.

Thus, I want to remind you that if you conclude that one treatment is better than the other, on average, it *does not* mean that said treatment is better *for all units*. Similarly, as we will see below when we look at Case 5, if we conclude that two treatments might be identical, on average, it does not mean that they are identical for all units. Thus, as a scientist, you should *think about* the possibility of Case 3 being the truth. If Case 3 is the truth, as a scientist you should explore *why* the units might respond so differently to the treatments. In our Case 3, perhaps subjects 1–3 belong to one genotype, subjects 4 and 6 belong to another and subject 5 belongs to a third. In this scenario, the physician could advise a patient on which treatment to use by determining the patient's genotype. Statistically, this issue can be explored by using a **randomized pair design**, which we will study in later chapters.

Case 4 shows us that even though treatment 2 performed better than treatment 1 in our data ($\bar{y} < \bar{x}$), it is possible that, either overall or for every subject, treatment 1 is actually better than treatment 2. In fact, if you examine Definition 5.1, you will see that Case 4 is the constant treatment effect with $c = -3$.

Finally, Case 5 shows that it is possible for the Skeptic to be wrong, yet, overall, the treatments perform the same. Sadly, a CRD is **not** able to detect the difference between Cases 1 and 5, regardless of how many subjects are in the study.

5.1.4 Finally! The Alternative Hypothesis is Specified

Recall that our null hypothesis is that the Skeptic is correct, which we write as follows:

> H_0 : The Skeptic is correct.

As illustrated in Case 1 in Table 5.3, if the Skeptic is correct, then $\mu_1 = \mu_2$. This equality is always a consequence of the Skeptic being correct. As illustrated in Case 5 in Table 5.3, however, the reverse implication is not true; it is possible to have equality of row means ($\mu_1 = \mu_2$) **without**

the Skeptic being correct. Thus, be careful! Our null hypothesis is that the Skeptic is correct, **not** that the row means are equal. The alternative hypothesis, however, is written in terms of the row means, as you will now see.

There are three options for the alternative hypothesis. The researcher selects from the following options.

- $H_1 : \mu_1 > \mu_2$.
- $H_1 : \mu_1 < \mu_2$.
- $H_1 : \mu_1 \neq \mu_2$.

Note that even if one chooses the last of these options, $\mu_1 \neq \mu_2$, it is scientifically possible that neither the null nor the alternative is correct; i.e., Case 5 in Table 5.3 could be correct. It is easier—as well as being the standard approach—for statisticians to ignore this possibility and talk of *either the null or alternative being true*. I will take this approach. Does it bother me to do so? No, because I want you to always interpret any statistical analysis with caution. Always be aware that we might be ignoring something important.

It is useful to think of the third option for the alternative, $\mu_1 \neq \mu_2$, as the *combination* of the $\mu_1 > \mu_2$ and $\mu_1 < \mu_2$ options; i.e., for the treatment means to be different (not equal) then either the mean of treatment 1 is larger than the mean of treatment 2 or the mean of treatment 1 is smaller than the mean of treatment 2.

Note that usually in these notes I will abbreviate: $\mu_1 > \mu_2$ by $>$; $\mu_1 < \mu_2$ by $<$; and $\mu_1 \neq \mu_2$ by $\neq$. This should cause no confusion provided you remember that μ_1 is to the left and μ_2 is to the right of the abbreviating symbol.

How does one decide between these three options for the alternative? It is my experience that many people, especially those new to tests of hypotheses, believe that the obvious choice is $\neq$. Some go so far as to say that the alternative **should** be $\neq$ and that the other two options should be forgotten. There is logic to this attitude: If the Skeptic is correct, then $\mu_1 = \mu_2$. The negation of this equality is that $\mu_1 \neq \mu_2$.

We could take a survey of all people in this class. Suppose that we **all agreed** that the alternative must be $\neq$. It wouldn't matter. One of my jobs in this class is to *transmit the culture to you*. Currently, the culture of science and statistics is that there are three options. I don't anticipate that this culture will change. Thus, we need to consider all three possible alternatives.

We find ourselves back at the question: How does one decide between these three options for the alternative? Well, again let me remind you that we make this decision before we collect data. It is considered to be serious cheating to look at one's data and then select the alternative. (Later in these notes we will see why this is so.)

Next, other than timing, there are no absolute rules on how to choose among the alternatives. Each researcher is allowed to make a personal choice and nobody can say that the researcher is wrong. You might disagree with the researcher or believe that the researcher is misguided, but you cannot say the researcher is wrong. (It's a bit like having a favorite color. I cannot say that your favorite color is wrong!)

While I am not going to give you a rule, I will give you guidance. I recommend that you use what I call the **Inconceivable Paradigm**. Here is how you proceed.

(Note: If you have seen the 1987 movie *The Princess Bride* you might recall that the hero of the movie repeatedly said that things that had occurred were *inconceivable*! This is the attitude I want to convey; we may think something is inconceivable and be wrong. Inconceivable is weaker than impossible. And despite what certain sports gear ads imply, impossible is, well, impossible. Oh, and by the way, if a major character in a movie is old, bald and dumpy, I label him the movie's hero. If you need a break from these notes, the following website presents a brief humorous video on the connection between the movie and the word *inconceivable*:

`https://www.youtube.com/watch?v=qhXjcZdk5QQ`.)

Definition 5.2 (The Inconceivable Paradigm.) *The Inconceivable Paradigm is defined to be the following procedure.*

- *Consider the alternative $\mu_1 > \mu_2$. Is this possibility—in the mind of the researcher—inconceivable? Answer yes or no.*
- *Consider the alternative $\mu_1 < \mu_2$. Is this possibility—in the mind of the researcher—inconceivable? Answer yes or no.*
- *If the answer to both questions is 'No, it's conceivable,' then use the alternative $\neq$. If the answer is 'Yes, it's inconceivable' to exactly one question then you throw away the corresponding alternative and use the other one.*

By the way, if one believes that both $>$ and $<$ are inconceivable, then there is, arguably, no point in doing the study. This person believes that $\mu_1 = \mu_2$ and anything else is inconceivable. It seems to me that a minimal requirement for being a scientist is the willingness to learn something from experimentation!

Let's see how the Inconceivable Paradigm works for Dawn's study. I will discuss three versions of Dawn.

1. Dawn G.T. (for $>$; i.e. greater than; motivated by a favorite comedian of mine, Louis C.K.) decides that it is inconceivable for Bob to consume less chicken than tuna (this is $<$). Her reasoning is, as she wrote in her report,

   > I noticed that the percentage of real chicken in the chicken-flavored treats was larger than the percentage of real tuna in the tuna-flavored treats.

   Because Dawn G.T. believes that $<$ is inconceivable, she discards it and uses $>$ for her alternative.

2. Dawn L.T. decides that it is inconceivable for Bob to consume more chicken than tuna. Her reasoning is, as she wrote in her report,

   > . . . Bob absolutely loves canned tuna with a passion.

Because Dawn L.T. believes that $>$ is inconceivable, she discards it and uses $<$ for her alternative.

3. Dawn N.E. acknowledges her conflicting thoughts: For 'real' food Bob seems to prefer tuna, but he might prefer the chicken version of the -flavored food. Thus, she is unwilling to label either $>$ or $<$ inconceivable; she refuses to discard either; and is left with $\neq$ as her alternative.

I suspect that you may have surmised my attitude on this issue. If I were the researcher planning the study of Bob, then I would have used the alternative $\neq$. My philosophy is that I will use $\neq$ **unless** I feel **very strongly** that one of the options $>$ or $<$ is clearly inconceivable. You are free to develop your own attitude on this issue, but I would hope that you remain open-minded until we have covered more material in these notes. I have met scientists who say that they **never** use $\neq$ and I have met scientists who say that they **always** use $\neq$. I recommend something less rigid than either of these extremes.

Aside: One of the reasons I chose Dawn's study to begin these notes is because it is a study in which one might reasonably choose any of the three options for the alternative. This fact makes the study particularly attractive for my purpose of introducing you to tests of hypotheses.

Before I continue, let's look at Kymn's, Sara's and Cathy's studies.

In Kymn's study, she wrote,

> My coxswain told me that rowers who are muscularly very strong perform better on the small gear setting (treatment 1) than they do on the large gear setting (treatment 2), while those who are aerobically very fit show the reverse tendency.

Kymn believed her coxswain *and* considered herself to be aerobically very fit rather than muscularly very strong. As a result, Kymn considered the alternative $<$ (faster times on small gear) to be inconceivable. Thus, she selected the alternative $>$: in the clone-enhanced study, her times on the first treatment, on average, would be larger (worse) than her times on the second treatment.

Sara chose the alternative $\neq$ because she was unwilling to label either $>$ or $<$ inconceivable. Sara acknowledged that she was a novice golfer and stated in her report that she was learning golf primarily because she thought it would be useful in her future career in business.

Cathy chose the alternative $\neq$. She reported that she enjoyed the run through the park more than the run at the high school, but—before collecting data—she was uncertain as to whether more enjoyment would lead to running faster or slower.

5.2 Step 2: The Test Statistic and Its Sampling Distribution

We studied the test statistic U and its sampling distribution in Chapters 3 and 4. Recall that we calculated the sampling distribution by assuming that the Skeptic is correct. In the language of this chapter, the sampling distribution is calculated on the assumption that the null hypothesis is true.

Recall that the observed value of the test statistic U is denoted by u and is obtained as follows:

$$u = \bar{x} - \bar{y}.$$

Because of this definition of u, we will call the test of this chapter a **test based on means**. In Chapter 6 we will learn about a competitor to the test of this chapter which is a **test based on ranks**.

5.3 Step 3: Calculating the P-value

In Step 2 we focus on the null hypothesis and totally ignore the alternative. In particular, it does not matter which of the three possible alternatives ($>$, $<$ or $\neq$) was chosen by the researcher. In Step 3, we focus on the alternative. As a result, we must work through the details of Step 3 three times, once for each possible alternative.

5.3.1 The P-value for the alternative $>$

Remember that I have invented three Dawn researchers: Dawn G.T. ; Dawn L.T. ; and Dawn N.E. In this subsection we consider Dawn G.T. ; the *version* of Dawn who chose the alternative $>$.

The null hypothesis (Skeptic's Argument) implies that every day the chicken response would equal the tuna response. Yes, day-to-day responses could vary. But if the null is correct then in the clone-enhanced study the total (or the mean) number of treats eaten at the end of 20 days would be exactly the same on the two treatments. If, however, the alternative $>$ is true, then in the clone-enhanced study at the end of the 20 days the number (mean) of chicken treats eaten would be larger than the number (mean) of tuna treats eaten.

Of course, Dawn could not perform the clone-enhanced study. What Dawn *could perform* was her actual study. In her actual study—I like to use the colorful language of sports here—chicken defeated tuna by 2.2 treats. Intuitively, the fact that chicken *won* its *actual* game with tuna provides **evidence** that chicken would have won the clone-enhanced study. In other words, intuitively, the actual study provides evidence in support of the $>$ alternative.

Now we get to the key question: How strong is this evidence in support of $>$? We answer this question in two parts.

Recall that one of the *big ideas* of Statistics is to evaluate what actually happened by considering everything that could have happened. Recall, also, that when I type *everything that could have happened* I am referring to all the possible assignments that could have arisen from the process of randomization. We looked at this issue in Chapters 3 and 4 by looking at the sampling distribution of the test statistic U.

So, the first thing (part) we do is list all the possible values of the test statistic U. It can be shown that the possible values of U are:

$$-3.8, -3.6, \ldots, -2, 4, -2.2, -2.0, \ldots, 0, \ldots, 2.0, 2.2, 2.4, \ldots, 3.8.$$

Before I continue let me make a few comments about this list.

1. The ability to determine all possible values of U will not be helpful in this course. Thus, I won't explain how I created this list of values.

2. Our simulation experiment, reported in Table 4.1 on page 76, obtained almost all of the values I list above. Indeed, it missed only the values: -3.8, 3.4, 3.6 and 3.8. Because the simulation experiment yielded the values -3.4 and -3.6, we could have inferred the existence of the values 3.4 and 3.6 because of symmetry.

3. As we will see below, our argument for finding the P-value does not require us to know all possible values of U. But knowing them all helps me to *motivate* the formula for finding the P-value.

As stated earlier, the actual $u = 2.2$ provides evidence in support of $>$. Think of my sport's language: In the actual study, chicken defeated tuna by 2.2 treats. *If chicken had won the game* by 2.4 or 2.6 or ... or 3.8 treats, then the evidence in support of $>$ would be stronger than the actual evidence. By similar reasoning, if chicken had won the game 2.0, 1.8, ... -3.8, then the evidence in support of $>$ would be weaker than the actual evidence. (Note the absurdity of language of saying that chicken won the game by a negative amount; winning by a negative amount is more commonly referred to as **losing**, but my point remains valid. The evidence for $>$ provided by any negative value of u would be weaker than the evidence for $>$ provided by the actual positive u.)

To summarize the above, the event $(U \geq 2.2)$ consists of all assignments that would give evidence in support of $>$ that **is equal to or stronger than** the evidence in support of $>$ provided by the actual study.

All that remains is to calculate the **proportion** of assignments that give $(U \geq 2.2)$. This, of course, is called the probability of $(U \geq 2.2)$: $P(U \geq 2.2)$. This probability is called the **P-value** for our test of hypotheses and the alternative $>$.

Let's pause for a moment before we proceed. You might be thinking,

> I have learned the formula for the P-value in one very specific situation: Dawn's data with the alternative $>$. (Big deal!)

In fact, you have learned much more, although it remains for me to tell you. What you have learned is the **meaning** of the P-value. **This meaning is true in every statistical analysis of data. Every analysis in this course and beyond.** So what is the meaning?

> The P-value measures the probability—calculated under the assumption the null hypothesis is true—of obtaining the **evidence actually obtained or stronger evidence in support of the alternative**.

Note that we always talk about evidence in support of the alternative; we never talk about evidence in support of the null. There is no need to evaluate evidence in support of the null because we begin our analysis assuming the null is true. Here is an analogy: In a felony case, there is no need to provide evidence of innocence, only of guilt, because the trial begins with the assumption that the defendant is innocent.

Let's consider some extreme possibilities for the P-value.

1. Suppose that the P-value is really close to zero, say, one in a billion. This means that the evidence in the data for the alternative is very strong in the sense that only one in a billion

assignments would yield the same or stronger evidence. In the face of such a small P-value, one can cling to the Skeptic being correct only if one believes that something incredibly unlikely occurred in the study. Personally, I would always reject the null hypothesis with such a small P-value. I have never met a statistician or a scientist who would admit to disagreeing with me.

2. Suppose that the P-value is close to one, say, 0.90. This means that the evidence in the data for the alternative is quite weak in the sense that 90% of the assignments would yield the same or stronger evidence. Given our preference—because of Occam's Razor—for the null hypothesis, nobody would ever suggest rejecting the null hypothesis for such a large P-value.

I hope that the above two examples are helpful, but I must admit—as you have no doubt noticed—there is a lot of territory between one in a billion and 0.90! Thus, we will need to consider in more detail how to interpret a P-value. This will come later.

We have spent a great deal of time finding the formula for the P-value for the alternative $>$ for Dawn's CRD. Fortunately, the ideas above can be generalized easily to any CRD with a numerical response. In the general case, the value 2.2 for u simply is replaced by the actual value of u. Thus, for example, in Kymn's study we replace 2.2 by her actual $u = 7.2$; in Sara's study we replace 2.2 by her actual $u = 8.700$; and in Cathy's study we replace 2.2 by her actual $u = 5.00$.

We have the following general rule for computing the P-value.

Result 5.1 *For the alternative $\mu_1 > \mu_2$, the P-value is equal to*

$$P(U \geq u) \tag{5.1}$$

In this equation, remember that u is the actual observed value of the test statistic.

Let's apply Equation 5.1 to our four CRDs. We do this even though only Dawn G.T. and Kymn chose the alternative $>$. All of the studies have the potential to provide us with practice at using Equation 5.1.

1. For Dawn's CRD, the P-value for $>$ is $P(U \geq 2.2)$. I don't know this number. Following our work in Chapter 4, we will approximate this by its relative frequency in our computer simulation experiment.

   - For my original 10,000 rep simulation experiment, the approximate P-value, 0.0198, is presented in Table 4.2 on page 76. In addition, in the first bullet in the list beginning on page 78, I show that we can be *nearly certain* that the exact P-value is between 0.0156 and 0.0240.
   - For my extended, 100,000 rep, simulation experiment referenced on page 79, I obtained the approximate P-value 0.01879. In addition, I am *nearly certain* that the exact P-value is between 0.01750 and 0.02008.

2. For Kymn's CRD, the P-value for $>$ is $P(U \geq 7.2)$. It is particularly easy to obtain this P-value because 7.2 is the largest possible value of U and it is obtained by only one assignment. Thus, the exact P-value is $1/252 = 0.0040$.

3. For Sara's CRD, the P-value for $>$ is $P(U \geq 8.7)$. I don't know this number. Following our work in Chapter 4, we will approximate this by its relative frequency in our computer simulation experiment. For my 10,000 rep simulation experiment the approximate P-value, 0.0903, is presented in Table 4.3 on page 77. In addition, in the fourth bullet in the list beginning on page 78, I show that we can be *nearly certain* that the exact P-value is between 0.0817 and 0.0989.

4. For Cathy's CRD, the P-value for $>$ is $P(U \geq 5.0)$. We can see from Table 3.5 on page 64 that

$$P(U \geq 5.0) = P(U = 5.0) + P(U = 7.67) + P(U = 9.0) + P(U = 9.67) = 0.20.$$

How should **you** interpret the P-value? The **classical approach** advocated by many statisticians is:

> Reject the null hypothesis in favor of the alternative if, and only if, the P-value is less than or equal to 0.05.

This viewpoint is reinforced by some technical language: Whenever a test yields a P-value that is less than or equal to 0.05 we say that the data are **statistically significant**. If the P-value is greater than 0.05, we say that the data are **not statistically significant**. We never say that the data are *statistically insignificant*; statisticians are quite picky about how we negate our expressions!

A variation on the classical approach is to replace the threshold value of 0.05 by some other value. The most popular other threshold values are 0.01 and 0.10. In fact, whenever a test yields a P-value that is less than or equal to 0.01 we say that the data are **highly statistically significant**. Statisticians' use of modifiers in this context is restricted to *highly*. Perhaps you can someday popularize some other modifiers; e.g., phatly or awesomely or Bieberly statistically significant.

For the four researchers above, we know the exact P-values for Kymn and Cathy; Kymn's data are highly statistically significant and Cathy's are not statistically significant. For Dawn and Sara, we don't know the exact P-values, but based on the nearly certain intervals, we say that Dawn's data are statistically significant and Sara's data are not statistically significant.

5.3.2 The P-value for the alternative $<$

There is no work to be done for the alternative $<$ because it is mathematically equivalent to the alternative $>$. Why? Well, suppose that you are using the alternative $<$; literally, that treatment 1 gives smaller response values than treatment 2. If you simply relabel the treatments: The old 1 [2] becomes the new 2 [1], the alternative becomes $>$ and we can use the results of the previous subsection. For completeness, the rule for finding the P-value is below, just in case you don't want to be bothered with renaming treatments.

Result 5.2 *For the alternative $\mu_1 < \mu_2$, the P-value is equal to*

$$P(U \leq u) \tag{5.2}$$

In this equation, remember that u is the actual observed value of the test statistic.

I will illustrate the use of this formula for our four CRDs, despite the fact that none of the researchers selected this alternative. I simply want to give you practice.

1. For Dawn's CRD, the P-value for $<$ is $P(U \leq 2.2)$. I don't know this number. Following our work in Chapter 4, we will approximate this by its relative frequency in our computer simulation experiment. For my original 10,000 rep simulation experiment, the approximate P-value, 0.9895, is presented in Table 4.2 on page 76. This P-value is *so huge* that I am not going to bother reporting the nearly certain interval. (I did, you might recall, find it in Chapter 4.) In addition, with such a huge approximate P-value I am not going to bother with looking at the simulation experiment with 100,000 reps.

2. For Kymn's CRD, the P-value for $<$ is $P(U \leq 7.2)$. It is particularly easy to obtain this P-value because 7.2 is the largest possible value of U. Thus, it is the weakest possible evidence for $<$, making the exact P-value equal to 1.

3. For Sara's CRD, the P-value for $<$ is $P(U \leq 8.7)$. I don't know this number. Following our work in Chapter 4, we will approximate the probability of this event by its relative frequency in our computer simulation experiment. For my 10,000 rep simulation experiment the approximate P-value, 0.9107, is presented in Table 4.3 on page 77. This P-value is *so huge* that I am not going to bother reporting the nearly certain interval. (I did, you might recall, find it in Chapter 4.)

4. For Cathy's CRD, the P-value for $<$ is $P(U \leq 5.0)$. We can see from Table 3.5 on page 64 that
$$P(U \leq 5.0) = 1 - P(U > 5.0) = 1 - 0.15 = 0.85.$$

Note that all four of these P-values are very large. This is no surprise because for every study, $u > 0$. Hence, intuitively, there is extremely weak evidence in support of $<$.

5.3.3 The P-value for the alternative $\neq$

I will motivate our method for Dawn N.E. The key to our argument is: The alternative $\neq$ is the combination of $>$ and $<$.

Recall that Dawn's actual u equals 2.2. This positive value reflects the fact that $\bar{x} > \bar{y}$. Intuitively, any $u > 0$ gives stronger evidence for the alternative $>$ than it does for the alternative $<$. We know how to calculate the P-value for $>$; it is $P(U \geq u)$. Thus, for Dawn N.E. *part* of her P-value is $P(U \geq 2.2)$.

But the idea of the P-value is to look at all possible assignments; many of these assignments would give a value of u that is negative. In particular, I am interested in all the assignments that give $u = -2.2$. Here are two facts to note about $u = -2.2$:

1. Because it is a negative number, the evidence it provides for the alternative $<$ is stronger than the evidence it provides for the alternative $>$.

2. The strength with which the value $u = -2.2$ supports $<$ is **exactly equal to** the strength with which the value $u = 2.2$ supports $>$.

I hope that the second of these facts is reasonable to you. (Actually proving it is beyond the scope of these notes; I just want it to make sense to you.) If not, consider the following argument.

> Dawn obtained $u = 2.2$ because $\bar{x} = 5.1$ and $\bar{y} = 2.9$. The assignments that give $u = -2.2$ all have $\bar{x} = 2.9$ and $\bar{y} = 5.1$. **Clearly** the latter supports $<$ with exactly the same strength that the former supports $>$.

We almost have the answer. The P-value for Dawn N.E. must include $P(U \geq 2.2)$ as discussed earlier. It must also include $P(U = -2.2)$ because -2.2 provides the same strength of evidence for $\neq$ as $u = 2.2$ does. Finally, the P-value for Dawn N.E. must include $P(U < -2.2)$ because any $u < -2.2$ provides stronger evidence than $u = -2.2$ for the alternative $<$ and, hence, $\neq$.

To summarize, for Dawn N.E. the P-value equals

$$P(U \geq 2.2) + P(U \leq -2.2).$$

We can approximate this sum by adding the relative frequencies of each term, which we already did in Table 4.2 on page 76. The result is that the approximate P-value equals 0.0371. As shown in the third bullet in the list beginning on page 78 in Chapter 4, the nearly certain interval for the exact P-value is 0.0314 to 0.0428.

We can abstract the above argument for Dawn N.E.

Result 5.3 *For the alternative $\neq$,*

- *If the actual $u = 0$ then the P-value equals 1.*
- *If the actual $u > 0$ then the P-value equals*

$$P(U \geq u) + P(U \leq -u).$$

- *If the actual $u < 0$ then the P-value equals*

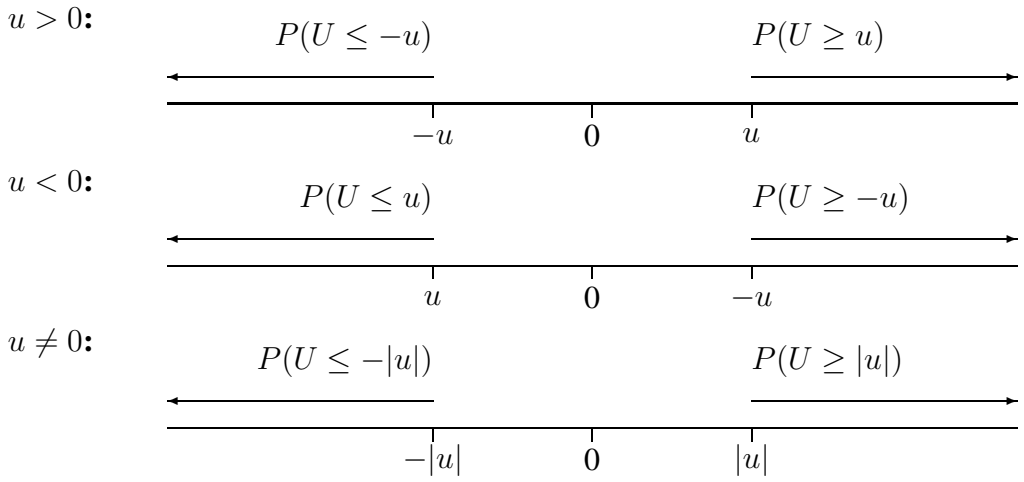
$$P(U \leq u) + P(U \geq -u).$$

- *We can combine the previous two items: If the actual $u \neq 0$, then the P-value equals*

$$P(U \geq |u|) + P(U \leq -|u|) \tag{5.3}$$

Let me make a few comments on the above. If the actual u equals 0, then this is the weakest possible evidence for the alternative of $\neq$. Remembering that the P-value is the proportion of assignments that yield evidence equal to or stronger than the actual evidence, the P-value must be 1 and no computations are required. Three versions of the situation for $u \neq 0$ are presented in Figure 5.1. I will end this subsection by finding the P-value for the alternative $\neq$ for our three remaining CRDs.

Figure 5.1: The P-value for the alternative $\neq$ for $u \neq 0$.



1. For Kymn's CRD, the P-value for $\neq$ is $P(U \geq 7.2) + P(U \leq -7.2)$. This value is $2/252 = 0.0079$. The data are highly statistically significant.

2. For Sara's CRD, the P-value for $\neq$ is $P(U \geq 8.7) + P(U \leq -8.7)$. We can approximate this sum by adding the relative frequencies of each term, which we already did in Table 4.3 on page 77. The result is that the approximate P-value equals 0.1824. As shown in the last bullet in the list beginning on page 78 in Chapter 4, the nearly certain interval for the exact P-value is 0.1708 to 0.1940. These data are not statistically significant.

3. For Cathy's CRD, the P-value for $\neq$ is $P(U \geq 5.0) + P(U \leq -5.0)$. We can see from Table 3.5 on page 64 that this probability equals $2(0.20) = 0.40$. These data are not statistically significant.

5.3.4 Some Relationships Between the Three P-values

Let me reiterate: The researcher selects **one** alternative before collecting data and, thus, obtains one P-value, be it approximate or exact. That being said, there is insight to be gained by considering the P-values for all three possible alternatives. I will do this in this subsection. Also, I have a few remarks about symmetry. Note that all of my comments and results below are for the **exact P-values**. This means that these comments and results are approximately true for approximate P-values.

Suppose that we add the P-value for $>$ to the P-value for $<$; from Results 5.1 and 5.2, we get:

$$P(U \geq u) + P(U \leq u) = 1 + P(U = u), \tag{5.4}$$

because the sum includes the assignments that give $(U = u)$ twice and all other assignments once. Equation 5.4 implies that at least one of these P-values must exceed 0.5000; in words, at least one

of these P-values must be very large. This is no surprise; if $u > 0$ [$u < 0$] then the P-value for $<$ [$>$] must be large.

Note that many books state that these two P-values add to one. Literally, this is an incorrect statement, but for studies with large n it usually is a good approximation.

I won't prove the facts below and I encourage you simply to accept them. If you want to prove them and have difficulty, contact your instructor or TA. I am **not** trying to make this into a **math class**; I simply want you to have a better understanding of P-values.

If the three P-values are different numbers, then the alternative $\neq$ **never** gives the smallest of the three. It might be the largest or the middle value, but is never the smallest. Curiously, the following is possible:

> If the P-values are two distinct numbers, say b, b and c, then it must be the case that $b < c$ and the P-value for $\neq$ could be b or c. (See the different result below if the distribution of U is symmetric.)

The effect of symmetry. We get *better* results (easier to interpret and understand) if the sampling distribution of U is symmetric around 0. We saw in Result 3.2 on page 65 that if a CRD is balanced, $n_1 = n_2$, then the sampling distribution of U is symmetric around 0. In math language, balance is sufficient for symmetry. It turns out that balance is not necessary for symmetry, but we won't try to characterize the situations that yield symmetry.

We need to separate results for the cases where $u = 0$ and $u \neq 0$.

- For $u = 0$, the P-values for $>$ and $<$ are equal and both are larger than 0.5000. Also both are smaller than 1, except in the trivial case when $P(U = 0) = 1$, which makes all three P-values equal to 1. (Think about what it means if $P(U = 0) = 1$.)

  We already know that the P-value for $\neq$ is 1.

- For $u \neq 0$, define:

$$b = P(U \geq u); c = P(U \leq u); \text{ and } d = P(U \geq |u|) + P(U \leq -|u|).$$

  We see that b, c and d are the P-values for $>$, $<$ and $\neq$, respectively.

  1. **If** $u > 0$: We have $b \leq 0.5 < c$ and, by symmetry, $d = 2b$.
  2. **If** $u < 0$: We have $c \leq 0.5 < b$ and, by symmetry, $d = 2c$.

In view of these last two items, you may have realized the following.

Let's look at the approximate P-value for Dawn N.E. based on my simulation experiment with 10,000 reps. As shown above, it is equal to 0.0371 with a nearly certain interval of $[0.0314, 0.0428]$. Now, let's look at the approximate P-value for Dawn G.T. based on the same simulation experiment. As shown above, it is equal to 0.0198 with a nearly certain interval of $[0.0156, 0.0240]$. But from the above, the exact P-value for $>$ is equal to one-half of the exact P-value for $\neq$. Thus, we

could simply halve our answer for $\neq$ to obtain an answer for $>$. If we do this, our approximation is $0.0371/2 = 0.01855$, *with nearly certain interval*

$$[0.0314/2, 0.0428/2] = [0.0157, 0.0214].$$

We see that by making use of the symmetry in the problem, we can obtain a more precise approximation for the P-value for $>$ (the value of h decreases from 0.0042 to 0.00285, a substantial change).

We have much material to cover in a short time. As a result, I have chosen not to make you responsible for using this improvement.

5.4 Computing

There is a website that will perform our simulation experiment, but it gives limited output. Begin by going to a familiar site:

`http://vassarstats.net`.

This is a good time for you to remember my suggestion that you use Firefox, Safari or Chrome as your web browser. (In truth, I can only testify that Firefox works; I have tried neither Safari nor Chrome.) I do know from personal experience that Windows Explorer did not work for me.

When you get to the *vassarstats* website you will see a list of options in a column on the left of the page. Click on **Miscellanea**, the penultimate item in the list. This action takes you to a new page which gives you a choice of three options; select the second one: **Resampling Probability Estimates for the Difference Between the Means of Two Independent Samples**. This selection takes you to another page which asks you for input.

I don't want this presentation to be too general, so let's proceed with a specific example: I will illustrate the use of this site using Dawn's study of her cat.

The page requesting input states:

> Please enter the size of sample A.

In the box below this request, enter the value of n_1, which is 10 for Dawn's study. After I type 10 I click on the box **OK**. This takes me to a new page, which states:

> Please enter the size of sample B.

In the box below this request, enter the value of n_2, which is 10 for Dawn's study. After I type 10 I click on the box **OK**.

Now that I have entered my two sample sizes, the site takes me to a page with lots of information, with the heading:

> Resampling Probability Estimates

followed by four bullets of information.

The first two bullets explain how to proceed; I will paraphrase these below. The third bullet presents a **Caveat** which you may ignore. (The programer's comments concern using the site for **population-based inference** and they are not relevant for our current **randomization-based inference.**) The fourth bullet provides us with a familiar equation that we first saw in Formula 3.2 in Chapter 3. This bullet tells us that the total number of possible assignments for Dawn's study is 184,756.

Now I have some bad news for you. The site requires you to enter the data by typing; i.e., cutting-and-pasting is not possible. Thus, I proceed to enter Dawn's 20 observations into the site. After entering the 20 numbers, I click on the box labeled **Calculate**. This action creates quite a lot of output which we will now examine.

- The site gives us the means of each data set: $\bar{x} = 5.1$ and $\bar{y} = 2.9$. This provides a partial check that I entered the data correctly.
- The site gives us $u = \bar{x} - \bar{y} = 2.2$, which is calls $M_a - M_b$. (Alas, my kingdom for universally accepted notation!)
- The site gives us several items that you should ignore: t, df and two proportions that are one-tailed and two-tailed P-values for something called a t-test. The df is short for degrees of freedom and if you recall our definition of degrees of freedom in Chapter 1, you know that each sample has 9 degrees of freedom. As we will learn later in these notes, degrees of freedom often add, as they do here to give $9 + 9 = 18$. The P-values are appropriate for population-based inference.

So far, the site has not done anything to make me excited. I already knew how to obtain $\bar{x}$, $\bar{y}$ and u. For excitement, scroll down to the box labeled:

> Resample $\times 1000$.

I click on this box **exactly ten times**. (Channeling the hand-grenade scene in *Monty Python and the Holy Grail*, not nine times; not 11 times; exactly 10 times.) After all this clicking, I obtain the following information:

> Under the heading **Probabilities estimated via resampling**, there are two P-values: 0.0173 for one one-tailed and 0.0372 for two-tailed. I also am told that the number of resamplings—as we would say, the number of reps—equals 10,000.

If you are mimicking my actions on your own computer—as I hope you are—then you likely obtained numbers different from my 0.0173 and 0.0372. For example, I repeated the above steps and obtained 0.0170 and 0.0372. A third 10,000 rep simulation yielded, for me, 0.0169 and 0.0366.

The one-tailed P-value, 0.0173 (let's focus on my first set of values), is the approximate P-value for $>$. The two-tailed P-value, 0.0372, is the approximate P-value for $\neq$. These are comparable to the values, 0.0198 and 0.0371, respectively, that I obtained using Minitab. I trust Minitab, whereas, to me, the *vassarstats* site is totally a black box. Thus, the fact that the latter's values are similar to Minitab's values makes me feel a bit better about recommending you use *vassarstats*.

Note that the *vassarstats* site does **not** give a P-value for the alternative $<$ for Dawn's study. I imagine that the reason for this programming decision is: because u is positive, we know that the P-value for $<$ will be very large. In fact, I have experimented with the *vassarstats* site and determined that its output obeys the following rule:

- If $u > 0$, the one-tailed approximate P-value is for the alternative $>$. No approximate P-value is given for the alternative $<$.
- If $u < 0$, the one-tailed approximate P-value is for the alternative $<$. No approximate P-value is given for the alternative $>$.

In my experimentation with this site, I turned to the issue of what happens when $u = 0$. I am very distressed with my findings. In particular, **if $u = 0$ the *vassarstats* website's one-tailed answer is wrong!** These are harsh words. I don't like to say that an approximation is wrong, but theirs is, as I will now demonstrate.

Basically, you may enter any data you want into the site, provided $u = 0$. I modified Dawn's data by switching the 7 and the 8 from treatment 1 to treatment 2 and a 1 and a 3 from treatment 2 to treatment 1. For these modified data, $\bar{x} = \bar{y} = 4.0$ and $u = 0$, values all verified by the *vassarstats* site. Let's summarize what we know must be true about the P-values for these modified data.

- Because $u = 0$ we know that the two-tailed P-value is exactly 1.
- The P-value for $>$ [$<$] is $P(U \geq 0)$ [$P(U \leq 0)$]. By symmetry (Result 3.2 on page 65), we know that these probabilities are identical and that both are larger than 0.5 (follows from Equation 5.4 on page 102).

Thus, as a practical matter, we don't need to approximate the P-value; we know it is large.

I used the *vassarstats* site to perform a 10,000 rep simulation experiment. The site gave me one for the approximate two-tailed P-value (correct answer), but it gave me 0.0761 for the one-tailed P-value (horribly wrong answer)!

5.4.1 A Final Comment

In the Practice Problems and Homework I want you to use the *vassarstats* site as detailed above to obtain approximate P-values for two of the three possible alternatives. I will **not** give you any data sets that have $u = 0$, first because we don't need to approximate the P-value in this situation and, second, because the *vassarstats* answer for one-tailed is wrong, as demonstrated above.

Not surprisingly, I have some misgivings about teaching you to use a site that, on occasion anyways, gives horribly wrong answers. Teachers are **not supposed to do this!** How can I justify my action? Below are my reasons.

1. I don't have time to teach you Minitab—or any other statistical software package—unless I delete several topics that I don't want to delete. I could let simulation experiments be something *magical* that only I can obtain, but I don't want to do that. I want you to be able to obtain some answers. As best I can tell, *vassarstats* fails only when $u = 0$. If I discover that the site is more flawed than that, I will reconsider my presentation of this material.

2. Stating the obvious, in the modern world we routinely obtain answers from computers and (supposedly) trust them. (Well, we don't trust what people *write* unless we are hopelessly naive. But, in my experience, people trust computations. After all, they are called *computers*.) I choose to believe that I am doing you a service by giving you a concrete example of an error on a site. Ideally, this will motivate you to challenge answers that you obtain. For example, once you have more experience with P-values, you should realize immediately that the answer 0.0761 above is ridiculous for $u = 0$.

3. I would appreciate your feedback, via email, on this issue.

5.5 Summary

There are three steps to any statistical test of hypotheses.

1. Step 1: Choice of hypotheses.
2. Step 2: The test statistic and its sampling distribution.
3. Step 3: Calculating the P-value.

These three steps will be revisited many times in these *Course Notes* for different *situations* in science and Statistics.

There are always two hypotheses: the null, H_0, and the alternative, H_1. These must be specified by the researcher **before** any data are collected. **Specified** is a bit misleading, because the researcher's options are usually limited. For example, for the test of this chapter, there is only one possibility for the null and only three possibilities for the alternative.

Compared to the alternative, the null hypothesis provides a simpler model for reality. Thus, following Occam's Razor, every statistical test of hypotheses begins with the assumption that the null is true. After data are collected and analyzed, the researcher will decide to either: **reject the null hypothesis** in favor of the alternative, or **fail to reject the null hypothesis** in favor of the alternative.

Throughout Part I of these *Course Notes* the null hypothesis is that the Skeptic is correct. Some foundation is required before I can specify the alternative.

As we know, in a CRD the goal is to compare how the two treatments influence the responses of the n units being studied. Most of the CRDs we have considered in exposition, Practice Problems or Homework have been balanced: equal numbers of units are assigned to each treatment. It seems reasonable for the purpose of *comparison*, that it is a good idea to give each treatment the same number of opportunities to *perform*. Indeed, there are various mathematical results that say, in certain circumstances, balance is the best way to compare two treatments. We are not, however, interested in exploring this topic mathematically.

It is useful to consider the most extreme departure from balance that is possible; namely, assigning all units to the same treatment. There are, of course, two ways to do this: AT-1 assigns all units to treatment 1 and AT-2 assigns all units to treatment 2. Suppose that all units are assigned to treatment 1; define μ_1 to equal the mean response over all n units. Similarly, suppose that all units are assigned to treatment 2; define μ_2 to equal the mean response over all n units.

We invent the fanciful *clone-enhanced* study in which each unit is cloned and one version of each unit is assigned to each treatment. *If we could perform the clone-enhanced study*, then we would be performing **both** AT-1 and AT-2 and, hence, we could compute the values of μ_1 and μ_2.

It is easy to see that if the Skeptic is correct, then $\mu_1 = \mu_2$. In fact, if the Skeptic is correct we can compute this common value—just calculate the mean response of all units ignoring the treatment—but we have no interest in this. If the Skeptic is incorrect, I demonstrate in Table 5.3 on page 90 that anything is possible: μ_1 can be smaller than, equal to or greater than μ_2. The situation in which the Skeptic is incorrect **and** $\mu_1 = \mu_2$ is vexing. Here is why.

The Skeptic being incorrect implies that for some units the treatment does influence the response. The equality of μ_1 and μ_2 implies that for all of the units studied, on average, the two

treatments give the same responses. In other words, it must be the case that treatment 1 is superior to treatment 2 for some units, while the reverse is true for other units. Sadly, there is no way to distinguish between these two situations with our CRDs.

Of particular interest to us are the alternatives with a constant treatment effect, as defined in Definition 5.1.

There are three choices for the alternative hypothesis: $\mu_1 > \mu_2$; $\mu_1 < \mu_2$; and $\mu_1 \neq \mu_2$. For brevity, these are referred to as the alternatives $>$, $<$ and $\neq$, respectively. These abbreviations should cause no difficulty, as long as you remember that μ_1 always appears to the left of μ_2.

After specifying the hypotheses; i.e., choosing the alternative from the options $>$, $<$ and $\neq$, it is time for Step 2: The Test Statistic and Its Sampling Distribution. Step 2 is big and we covered it in Chapters 3 and 4.

Finally, there is Step 3: Calculating the P-value. We derive three formulas for calculating the P-value, one for each choice of alternative. In the formulas below, remember that u denotes the observed value of the test statistic for the actual data obtained in the CRD.

1. For the alternative $\mu_1 > \mu_2$, the P-value is

$$P(U \geq u).$$

2. For the alternative $\mu_1 < \mu_2$, the P-value is

$$P(U \leq u).$$

3. For the alternative $\mu_1 \neq \mu_2$, obtaining the P-value is bit tricky.

   - If $u = 0$ then the P-value equals one.
   - If $u > 0$, then the P-value equals

$$P(U \geq u) + [P(U \leq -u).$$

   - If $u < 0$, then the P-value equals

$$P(U \leq u) + [P(U \geq -u).$$

   - These last two rules can be combined into one. For $u \neq 0$ the P-value equals

$$P(U \geq |u|) + P(U \leq -|u|).$$

     A nice feature of this last version of the rule is that it reminds us that for the alternative $\mu_1 \neq \mu_2$, the sign of u does not matter; only its magnitude.

Table 5.4: Frequency Table for u for the 252 possible assignments for Kymn's Study.

u	Freq.	u	Freq.	u	Freq.	u	Freq.	u	Freq.	u	Freq.
-7.2	1	-4.8	3	-2.4	10	0.4	12	2.8	10	5.2	4
-6.8	1	-4.4	5	-2.0	8	0.8	10	3.2	8	5.6	3
-6.4	1	-4.0	8	-1.6	14	1.2	13	3.6	6	6.0	1
-6.0	1	-3.6	6	-1.2	13	1.6	14	4.0	8	6.4	1
-5.6	3	-3.2	8	-0.8	10	2.0	8	4.4	5	6.8	1
-5.2	4	-2.8	10	-0.4	12	2.4	10	4.8	3	7.2	1
				0.0	16					Total	252

5.6 Practice Problems

1. The purpose of this question is to reinforce the ideas that were illustrated in Table 5.3. First, I present artificial data from a balanced CRD with a total of eight units:

Unit:	1	2	3	4	5	6	7	8
Treatment:	1	2	2	1	2	1	2	1
Response:	22	14	18	22	10	18	26	14

   Next, I present the (largely unknown) results that would be obtained for the clone-enhanced version of this study. Note that there are eight *question marks* in this table, one for each unit *because we don't know how the clones would respond*!

Unit:	1	2	3	4	5	6	7	8	Mean
Response on Treatment 1:	22	?	?	22	?	18	?	14	$\bar{x} = 19$
Response on Treatment 2:	?	14	18	?	10	?	26	?	$\bar{y} = 17$

   (a) Complete the data table above for the clone-enhanced study on the assumption that the Skeptic is correct.

   (b) Complete the data table above for the clone-enhanced study on the assumption that Skeptic is incorrect and there is a constant treatment effect that is equal to 10.

2. Refer to HS-1 in this chapter; its artificial data are presented in Table 5.1 on page 88. I want to use this study of headaches to explore choosing the alternative hypothesis by using the *Inconceivable Paradigm*, which was presented in Definition 5.2 on page 94.

   (a) Suppose that drug A (treatment 1) is actually an *extra-strength* version of drug B (treatment 2). In addition, suppose that the researcher believes it is inconceivable that the *extra-strength* version is inferior to the regular version.

   Given these two suppositions, determine the researcher's choice for the alternative hypothesis.

(b) Suppose that drug A is a placebo (which violates my earlier story; sorry) and that B is an active drug. In addition, suppose that the researcher believes it is inconceivable that a placebo is superior to drug B.

Given these two suppositions, determine the researcher's choice for the alternative hypothesis.

3. The purpose of this problem is to give you practice using the three rules for computing the P-value. In Chapter 3, I presented the distribution of the values of u for the 252 possible assignments for Kymn's study of rowing; for convenience, I have reproduced this distribution in Table 5.4. For parts (a)–(c) below, forget that Kymn's actual u equals 7.2. Instead, use the different values of u I specify and Table 5.4 to find the various P-values.

   (a) For the alternative $>$, find the exact P-value for each of the following three values of u: 6.8, 5.6 and 4.0.

   (b) For the alternative $<$, find the exact P-value for each of the following three values of u: -6.4, -5.2 and -2.8.

   (c) For the alternative $\neq$, find the exact P-value for each of the following three values of $|u|$: 5.6, 5.2 and 2.8.

4. In the lone Practice Problem of Chapter 4 (Section 4.6) I introduced you to Doug's study of the dart game 301. Please read about the study again now. Remember that smaller response values correspond to better outcomes for Doug. Doug's summary statistics included:

$$\bar{x} = 18.60 \text{ and } \bar{y} = 21.20, \text{ giving } u = 18.60 - 21.20 = -2.60.$$

   In Chapter 4, I reported the following results from a simulation with 10,000 reps that I performed using Minitab:

$$\text{r.f. } (U \leq -2.60) = 0.0426; \text{ and r.f. } (U \geq 2.60) = 0.0418.$$

   I entered the Doug's 40 responses into the *vassarstats* website and had it perform a simulation with 10,000 reps; I obtained the output below:

$$\text{one-tailed P-value } = 0.0433; \text{ two-tailed P-value } = 0.0836.$$

   (a) Imagine you are back in time with Doug before he collected his data. Doug says, "It is inconceivable that my throwing bar darts will yield better results than my throwing my personal darts."

   Which alternative should Doug choose?

   (b) Consider the alternative $<$.

      i. What is the approximate P-value based on my Minitab simulation?

      ii. What is the approximate P-value based on the *vassarstats* simulation?

iii. Compare these two approximations. You don't need to compute any nearly certain intervals; simply tell me what you think.

(c) Consider the alternative $>$.

i. What is the approximate P-value based on my Minitab simulation?

ii. What is the approximate P-value based on the *vassarstats* simulation?

iii. Compare these two approximations. You don't need to compute any nearly certain intervals; simply tell me what you think.

(d) Consider the alternative $\neq$.

i. What is the approximate P-value based on my Minitab simulation?

ii. What is the approximate P-value based on the *vassarstats* simulation?

iii. Compare these two approximations. You don't need to compute any nearly certain intervals; simply tell me what you think.

5.7 Solutions to Practice Problems

1. (a) There is only one possible way to make the Skeptic correct; it is:

Unit:	1	2	3	4	5	6	7	8	Mean
Response on Treatment 1:	22	14	18	22	10	18	26	14	$\mu_1 = 18$
Response on Treatment 2:	22	14	18	22	10	18	26	14	$\mu_2 = 18$

(b) There is one possible correct answer; it is:

Unit:	1	2	3	4	5	6	7	8	Mean
Response on Treatment 1:	22	24	28	22	20	18	36	14	$\mu_1 = 23$
Response on Treatment 2:	12	14	18	12	10	8	26	4	$\mu_2 = 13$

2. (a) The researcher believes it is inconceivable that the numbers on treatment 1 will be larger than those on treatment 2; thus, $>$ is inconceivable. The researcher's choice for the alternative is $<$.

(b) The researcher believes it is inconceivable that the numbers on treatment 1 will be smaller than those on treatment 2; thus, $<$ is inconceivable The researcher's choice for the alternative is $>$.

3. (a) For $u = 6.8$, the exact P-value is $P(U \geq 6.8)$. From Table 5.4 we find:

$$\text{Frequency } (U \geq 6.8) = 1 + 1 = 2.$$

Dividing by the total number of possible assignments, 252, we obtain

$$P(U \geq 6.8) = 2/252 = 0.0079.$$

For $u = 5.6$, the exact P-value is $P(U \geq 5.6)$. Mimicking what I did above,

$$\text{Frequency } (U \geq 5.6) = 3 + 1 + 1 + 1 + 1 = 7.$$

Dividing by the total number of possible assignments, 252, we obtain

$$P(U \geq 5.6) = 7/252 = 0.0278.$$

For $u = 4.0$, the exact P-value is $P(U \geq 4.0)$. Mimicking what I did above,

$$\text{Frequency } (U \geq 4.0) = 8 + 5 + 3 + 4 + 3 + 1 + 1 + 1 + 1 = 27.$$

Dividing by the total number of possible assignments, 252, we obtain

$$P(U \geq 4.0) = 27/252 = 0.1071.$$

(b) For $u = -6.4$, the exact P-value is $P(U \leq -6.4)$. From Table 5.4 we find:

$$\text{Frequency } (U \leq -6.4) = 1 + 1 + 1 = 3.$$

Dividing by the total number of possible assignments, 252, we obtain

$$P(U \leq -6.4) = 3/252 = 0.0119.$$

For $u = -5.2$, the exact P-value is $P(U \leq -5.2)$. Mimicking what I did above,

$$\text{Frequency } (U \leq -5.2) = 1 + 1 + 1 + 1 + 3 + 4 = 11.$$

Dividing by the total number of possible assignments, 252, we obtain

$$P(U \leq -5.2) = 11/252 = 0.0437.$$

For $u = -2.8$, the exact P-value is $P(U \leq -2.8)$. Mimicking what I did above,

$$\text{Frequency } (U \leq -2.8) = 1 + 1 + 1 + 1 + 3 + 4 + 3 + 5 + 8 + 6 + 8 + 10 = 51.$$

Dividing by the total number of possible assignments, 252, we obtain

$$P(U \leq -2.8) = 51/252 = 0.2024.$$

(c) The easy way to solve this is to remember that because the study is balanced, the sampling distribution of U is symmetric around 0. (Alternatively, symmetry can be seen in Table 5.4.) Thus, for example, for any $u > 0$,

$$P(U \geq u) = P(U \leq -u).$$

For $|u| = 5.6$ the P-value is

$$P(U \geq 5.6) + P(U \leq -5.6) = 2(7/252) = 7/126 = 0.0556,$$

from our result in (a) above.
For $|u| = 5.2$ the P-value is

$$P(U \geq 5.2) + P(U \leq -5.2) = 2(11/252) = 11/126 = 0.0873,$$

from our result in (b) above.
For $|u| = 2.8$ the P-value is

$$P(U \geq 2.8) + P(U \leq -2.8) = 2(51/252) = 51/126 = 0.4048,$$

from our result in (b) above.

4. (a) Remember that lower responses are preferred and that treatment 1 is using the personal darts. Doug believed it is inconceivable that the responses on treatment 1 would be larger than the responses on treatment 2. Thus, he believed that $>$ is inconceivable; his alternative is $<$.

   (b) For the alternative $<$ the P-value equals $P(U \leq -2.60)$.

      i. Using Minitab, the approximate P-value is 0.0426.
      ii. Using *vassarstats*, the approximate P-value is 0.0433.
      iii. These approximations are very close in value. Because I trust Minitab, I now feel better about the accuracy of *vassarstats*.

   (c) For the alternative $>$ the P-value equals $P(U \geq -2.60)$.

      i. **Trick question!** I do not give you the appropriate relative frequency from my Minitab simulation.
      ii. The *vassarstats* site does not give an approximate P-value for this alternative.
      iii. I cannot compare approximations that I don't know. I will note, however, that because $u < 0$, I know that the exact P-value is very large. In fact, it is likely greater than 0.95 because of Equation 5.4 and my answers to (a).

   (d) For the alternative $\neq$ the P-value equals

$$P(U \geq 2.60) + P(U \leq -2.60).$$

      i. Using Minitab, the approximate P-value is $0.0426 + 0.0418 = 0.0844$.
      ii. Using *vassarstats*, the approximate P-value is 0.0836.
      iii. These approximations are very close in value. Because I trust Minitab, I now feel better about the accuracy of *vassarstats*.

5.8 Homework Problems

1. The purpose of this question is to reinforce the ideas that were illustrated in Table 5.3. First, I present artificial data from an unbalanced CRD with a total of seven units:

Unit:	1	2	3	4	5	6	7
Treatment:	1	2	2	1	2	1	1
Response:	30	24	21	40	33	36	26

   Next, I present the (largely unknown) results that would be obtained for the clone-enhanced version of this study. Note that there are seven *question marks* in this table, one for each unit *because we don't know how the clones would respond*!

Unit:	1	2	3	4	5	6	7	Mean
Response on Treatment 1:	30	?	?	40	?	36	26	μ_1
Response on Treatment 2:	?	24	21	?	33	?	?	μ_2

   (a) Complete the data table above for the clone-enhanced study on the assumption that the Skeptic is correct.

   (b) Complete the data table above for the clone-enhanced study on the assumption that Skeptic is incorrect and there is a constant treatment effect that is equal to -12.

2. Suppose that **you** are going to perform the following CRD. Your units are trials with each trial being the toss of one dart at a standard dartboard. Your goal on each trial is to have the dart stick in the center (bull's eye) of the board. Your response is the distance, measured to the nearest centimeter, that your dart lands away from the center. If you miss the board entirely, make your best guess as to where the dart bounced off the wall—assuming you don't miss the wall too—and measure from your guess to the center. If your dart hits the board, but does not stick, use the radius of the board as your response; i.e., this is worse than any dart that sticks, but better than hitting the wall.

   Your treatments are: (1) using your right hand and (2) using your left hand.

   Use the Inconceivable Paradigm to obtain **your** choice for alternative. Briefly explain your answer.

3. The purpose of this problem is to give you practice using the three rules for computing the P-value. Table 4.1 in Chapter 4 presents the complete results of a simulation experiment with 10,000 reps for Dawn's study of her cat Bob. I have reproduced this table in Table 5.5.

   For parts (a)–(c) below, forget that Dawn's actual u equals 2.2. Instead, use the different values of u I specify and Table 5.5 to find the approximations to the various P-values.

   Use this table to obtain the approximate P-values requested below.

   (a) For the alternative $>$, find the approximate P-value for each of the following values of u: 2.8 and 2.4.

Table 5.5: The results of a simulation experiment with 10,000 reps for Dawn's study.

u	Freq.	Relative Freq.	u	Freq.	Relative Freq.	u	Freq.	Relative Freq.
-3.6	1	0.0001	-1.2	419	0.0419	1.2	394	0.0394
-3.4	1	0.0001	-1.0	506	0.0506	1.4	315	0.0315
-3.2	3	0.0003	-0.8	552	0.0552	1.6	251	0.0251
-3.0	4	0.0004	-0.6	662	0.0662	1.8	150	0.0150
-2.8	16	0.0016	-0.4	717	0.0717	2.0	108	0.0108
-2.6	25	0.0025	-0.2	729	0.0729	2.2	93	0.0093
-2.4	45	0.0045	0.0	732	0.0732	2.4	54	0.0054
-2.2	78	0.0078	0.2	765	0.0765	2.6	23	0.0023
-2.0	113	0.0113	0.4	716	0.0716	2.8	17	0.0017
-1.8	191	0.0191	0.6	674	0.0674	3.0	8	0.0008
-1.6	240	0.0240	0.8	553	0.0553	3.2	3	0.0003
-1.4	335	0.0335	1.0	507	0.0507			

(b) For the alternative $<$, find the approximate P-value for each of the following values of u: -2.6, and -1.8.

(c) For the alternative $\neq$, find the approximate P-value for each of the following values of $|u|$: 2.6 and 3.2.

4. Refer to Problem 1 of the Chapter 4 Homework (Section 4.8), a consideration of Reggie's dart data from Homework problems 5–7 in Chapter 1 (Section 1.8).

(a) Using the Inconceivable Paradigm, Reggie chose the alternative $>$. While you cannot read Reggie's mind, I am asking you to provide a *plausible* reason for his choice of $>$.

(b) Enter Reggie's data (available in Chapter 1 Homework) into the *vassarstats* website. Perform a simulation experiment with 10,000 reps.

i. The *vassarstats* site gives you two approximate P-values based on its simulation. Match these P-values to their alternatives.

ii. In Problem 1 of the Chapter 4 Homework I reported partial results from a simulation experiment with 10,000 reps that I obtained using Minitab. Use these partial results to obtain approximate P-values for the same alternatives you used in (i).

iii. Compare your P-values in (i) and (ii). Briefly comment.

Chapter 6

The Sum of Ranks Test

Thus far in these *Course Notes* we have considered CRDs with a numerical response. In Chapter 5 we learned how to perform a statistical test of hypotheses to investigate whether the Skeptic's Argument is correct. Every test of hypotheses has a test statistic; in Chapter 5 we chose the test statistic U which has observed value $u = \bar{x} - \bar{y}$. For rather obvious reasons, this test using U is referred to a test of means or a test of comparing means.

We learned in Chapter 1 that the mean is a popular way to summarize a list of numbers. Thus, it is not surprising to learn that comparing means, by subtraction, is a popular way to compare two treatments and, hence, the test of Chapter 5 seems sensible. But we also learned in Chapter 1 that the median is another popular way to summarize a list of numbers. Thus, you might guess that another popular choice for a test statistic would be the one whose observed value is $v = \tilde{x} - \tilde{y}$. If you make this guess, you would be wrong, but close to the truth.

Recall from Chapter 1 that the distinction between the mean and the median can be viewed as the distinction between focusing on *arithmetic* versus *position*. The median, recall, is the number at the center position of a sorted listed—for an odd sample size—or the average of the values at the two center positions—for an even sample size. Thus, the value v in the previous paragraph compares two sorted lists by comparing the numbers in their center positions. This comparison ignores a great deal of information! In those situations in which, for whatever reasons, we prefer to focus on positions rather than arithmetic, it turns out that using **ranks**, defined below, is superior to using medians in order to compare two sets of numbers.

In this chapter we will consider an option to using U: i.e., we will present a test that compares the two sets of data by comparing their ranks. When we study power in a later chapter, we will see that sometimes the test that compares ranks is better than the test that compares means. The last section of this chapter presents an additional advantage of using a test based on ranks; it can be used when the response is ordinal, but not numerical.

6.1 Ranks

We begin by doing something that seems quite odd: We combine the data from the two treatments into one set of data and then we sort the $n = n_1 + n_2$ response values. For example, for Dawn's

Table 6.1: Dawn's 20 sorted response values, with ranks.

Position:	1	2	3	4	5	6	7	8	9	10
Response:	0	1	1	1	2	3	3	3	3	4
Rank:	1	3	3	3	5	7.5	7.5	7.5	7.5	10.5

Position:	11	12	13	14	15	16	17	18	19	20
Response:	4	5	5	5	6	6	6	7	7	8
Rank:	10.5	13	13	13	16	16	16	18.5	18.5	20

study of her cat, the 20 sorted response values are given in Table 6.1. You can verify these numbers from the data presented in Table 1.3 on page 7, but I recommend that you just trust me on this. We note that Dawn's 20 numbers consist of nine distinct values. Her number of distinct values is smaller than 20 because several of the responses are **tied**; for example, four responses are tied with the value 3. Going back to Chapter 1, we talk about the 20 positions in the list in Table 6.1. As examples: position 1 has the response 0; position 20 has the response 8; and positions 6–9 all have the response 3.

If the n numbers in our list are all distinct, then the rank of each response is its position. This is referred to as the **no-ties situation** and it makes all of the computations below much simpler. Sadly, in practice, data with ties are commonplace. Whenever there are ties, all tied responses receive the same rank, which is equal to the mean of their positions. Thus, for example, all four of the responses equal to 3 receive the rank of 7.5 because they occupy positions 6 through 9 and the mean of 6, 7, 8 and 9 is 7.5. It is tedious to sum these four numbers to find their mean; here is a shortcut that always works: simply compute the mean of the smallest (first) and largest (last) positions in the list. For example, to find the mean of 6, 7, 8 and 9, simply calculate $(6+9)/2 = 7.5$. When we consider ordinal data in Section 6.5 we will have occasion to find the mean of

$$43, 44, \ldots, \text{ and } 75.$$

Summing these 33 numbers is much more tedious than simply computing $(43 + 75)/2 = 59$. Finally, for any responses in the list that is not tied with another—responses 0, 2 and 8 in Dawn's data—its rank equals its position.

The basic idea of our test based on ranks is that we analyze the ranks, not the responses. For example, I have retyped Table 6.1 in Table 6.2 (dropping the two *Position* rows) with the added feature that the responses from treatment 1 (chicken) and their ranks are in bold face type. For the test statistic U we performed arithmetic on the responses to obtain the means for each treatment and then we subtracted. We do the same arithmetic now, but we use the ranks instead of the responses. For example, let R_1 denote the sum of the ranks for treatment 1 and let r_1 denote its observed value. For Dawn's data we get:

$$r_1 = 3 + 7.5 + 10.5 + 13 + 13 + 16 + 16 + 16 + 18.5 + 20 = 133.5.$$

Table 6.2: Dawn's 20 sorted responses, with ranks. The responses from treatment 1, and their ranks, are in bold-faced type.

Response:	0	**1**	1	1	2	**3**	3	3	3	**4**
Rank:	1	**3**	3	3	5	**7.5**	7.5	7.5	7.5	**10.5**
Response:	4	**5**	**5**	5	**6**	**6**	**6**	**7**	7	**8**
Rank:	10.5	**13**	**13**	13	**16**	**16**	**16**	**18.5**	18.5	**20**

Similarly, let R_2 denote the sum of the ranks for treatment 2 and let r_2 denote its observed value. For Dawn's data we get:

$$r_2 = 1 + 3 + 3 + 5 + 7.5 + 7.5 + 7.5 + 10.5 + 13 + 18.5 = 76.5.$$

In order to compare the treatments' ranks descriptively, we calculate the mean of the ranks for each treatment:

$$\bar{r}_1 = r_1/n_1 = 133.5/10 = 13.35 \text{ and } \bar{r}_2 = r_2/n_2 = 76.5/10 = 7.65,$$

which show that, based on ranks, the responses on treatment 1 are larger than the responses on treatment 2.

The next obvious step is that we define

$$v = \bar{r}_1 - \bar{r}_2 = r_1/n_1 - r_2/n_2,$$

to be the observed value of the test statistic V. I say that this step is obvious because it is analogous to our definition of $u = \bar{x} - \bar{y}$; i.e., v is for ranks what u is for responses. **Except** we don't do the obvious. Here is why.

As the legend goes, as a child, Carl Friedrich Gauss (1777–1855), discovered that for any positive integer n:

$$1 + 2 + 3 + \ldots + n = n(n + 1)/2.$$

In our current chapter, this says that the sum of the positions for the combined set of n responses equals $n(n + 1)/2$. Because of the way ranks are defined above, it follows that the sum of the n ranks also equals $n(n + 1)/2$. If this is a bit abstract, note that for $n = 20$ Gauss showed that the sum of the ranks is $20(21)/2 = 210$, which agrees with our findings for Dawn's data:

$$r_1 + r_2 = 133.5 + 76.5 = 210.$$

As a result, given the values of n_1 and n_2—which the researcher will always know—knowledge of the value of r_1 immediately gives us the value of v. (If you like to see such things explicitly, for Dawn's study:

$$v = r_1/n_1 - (210 - r_1)/n_2.)$$

Now, I don't want to spend my time doing messy arithmetic, converting back-and-forth between v and r_1. Messy arithmetic is not my point! My point is that we are free to use either v or r_1 as the observed value of our test statistic. In these *Course Notes*, we will use r_1 as the observed value of the test statistic R_1. There are several advantages to using r_1 [R_1] instead of v [V]:

1. The value of r_1 is always a positive integer, whereas v will typically be a decimal and can be negative.

2. If computing by hand, one can obtain r_1 faster than one can obtain v.

3. For given values of n_1 and n_2, in the old days, it was more elegant to have tables of exact P-values as a function of the positive integer r_1 rather than the fraction v.

Admittedly, the last two of these *advantages* are less important in the computer age. On the other hand, positive integers have been our friends since early childhood and we all have early bad memories of decimals and negatives!

There are two main disadvantages of working with r_1 rather than v:

1. The value of r_1 alone does **not** tell us how the treatments compare descriptively.

2. Because r_1 is always positive, when we have symmetry (see below) it is no longer around 0, which makes our rule for the P-value for $\neq$ a bit more difficult to remember.

I want to introduce you to the **Mann Whitney (Wilcoxin) Rank Sum Test**. (The tribute to Wilcoxin, a chemist by training, is often suppressed because there is another test called the Wilcoxin Signed Rank Test.) We will call it the **sum of ranks test**, or, occasionally, the Mann Whitney test. The obvious reason for either name is: the observed value of the test statistic is obtained by summing ranks.

We will not spend a great deal of time in these *Course Notes* on procedures based on ranks. A big problem is interpretation. I can understand what $u = 2.2$ signifies: on average, Bob consumed 2.2 more chicken treats than tuna treats. I do not have a clear idea of how to interpret the difference in mean ranks for Dawn's data:

$$133.5/10 - 76.5/10 = 13.35 - 7.65 = 5.70.$$

For the—admittedly narrow—goal of deciding whether or not to reject the null hypothesis that the Skeptic is correct, the sum of ranks test can be useful.

6.2 The Hypotheses for the Sum of Ranks Test

The Skeptic's Argument is exactly the same as it was earlier for the difference of means test: The treatment is irrelevant; the response to any unit would have remained the same if the other treatment had been applied. The null hypothesis is, as before, that the Skeptic is correct.

In order to visualize the alternative, we need to remember the imaginary *clone-enhanced* study, first introduced on page 87. There is a total of $n = n_1 + n_2$ units being studied. With the clone-enhanced study, each of these units would yield two responses. Thus, the combined data for the clone-enhanced study would consist of $2n$ observations, with n observations from each treatment. Thus, note that even if the CRD is unbalanced the clone-enhanced study is balanced because each unit gives a response to both treatments.

Table 6.3: Case 2 of the clone-enhanced studies of Chapter 5: A constant treatment effect of $c = 6$.. The actual data are in bold-faced type.

Unit:	1	2	3	4	5	6	ρ_i
Response on Treatment 1:	**18**	9	**24**	18	12	**15**	
Ranks:	10	3.5	12	10	6	8	8.25
Response on Treatment 2:	12	**3**	18	**12**	**6**	9	
Ranks:	6	1	10	6	2	3.5	4.75

This presentation is getting quite abstract; thus, we will look at a numerical example. Table 6.3 reproduces Case 2 of Table 5.3 on page 90. In HS-1, $n_1 = n_2 = 3$ giving $n = 6$ and $2n = 12$ total responses in the clone-enhanced study. You may verify the ranks given in Table 6.3 for practice. Or not; your choice.

Define ρ_1 (pronounced 'roh') to be the mean of the ranks on treatment 1 in the clone-enhanced study. Similarly, let ρ_2 be the mean of the ranks on treatment 2 in the clone-enhanced study. If the Skeptic is correct then the two sets of data in the clone-enhanced study will be identical, which implies that $\rho_1 = \rho_2$. Although I won't give you details (and, thus, don't worry about it) it is possible for the Skeptic to be incorrect and yet $\rho_1 = \rho_2$.

For the clone-enhanced data in Table 6.3 we see that $\rho_1 = 8.25$ is larger than $\rho_2 = 4.75$ which is in the same direction as our earlier computation that for Case 2, $\mu_1 = 16$ is larger than $\mu_2 = 10$. Thus, *sometimes* (often?) looking at ranks gives a *similar* answer to looking at means, but, as we will see below, not always.

For the sum of ranks test, the three options for the alternative are given below:

- $H_1 : \rho_1 > \rho_2$.
- $H_1 : \rho_1 < \rho_2$.
- $H_1 : \rho_1 \neq \rho_2$.

As a practical matter, just remember that $>$ [$<$] means that—in terms of ranks—treatment 1 tends to give larger [smaller] responses than treatment 2; and that $\neq$ means that treatment 1 tends to give either larger or smaller responses than treatment 2.

Earlier in these notes I mentioned that $\bar{x}$ [$\bar{y}$] can be viewed as our point estimate of μ_1 [μ_2]. The relationships between r_1, r_2, ρ_1 and ρ_2 are problematic. All that I can safely say is that

$$r_1/n_1 > r_2/n_2 \text{ provides evidence that } \rho_1 > \rho_2.$$

I will defer further discussion of this issue until we examine population-based inference for the sum of ranks test. (For example, performing the AT-1 study would give us the value of μ_1. Thus, even though we need to imagine the fanciful clone-enhanced study to obtain both μ_1 and μ_2, it is always possible to determine one of these with a study. By contrast, with ranks, the AT-1 study tells us nothing about ρ_1; think about why this is true.)

Let's stop and take a breath. We have completed Step 1 our sum of ranks test; we have specified the null and alternative hypotheses. We can now move to Steps 2 and 3.

Table 6.4: Cathy's times, in seconds, to run one mile. HS means she ran at the high school and P means she ran through the park. The responses and ranks of the high school data are in bold-face.

Trial:	1	2	3	4	5	6
Location:	HS	HS	P	P	HS	P
Time:	**530**	**521**	528	520	**539**	527
Rank:	**5**	**2**	4	1	**6**	3

6.3 Step 2: The Test Statistic and Its Sampling Distribution

The test statistic will be R_1 with observed value r_1. We have three options for finding the sampling distribution and then using it to find our P-value. Two of these options are familiar and one is new. The familiar ones are:

1. Determine the exact sampling distribution of R_1. In these notes, this option is practical only for very small studies.

2. Use a computer simulation experiment to approximate the sampling distribution of R_1.

The third option is called **the Normal curve approximation** of the sampling distribution of R_1; it will be presented in Chapter 7.

I will begin by determining the exact sampling distribution of R_1 for Cathy's study of routes for running. Table 6.4 is a reproduction of our earlier table of Cathy's data, with ranks now added. Again, make sure you are able to assign ranks correctly. You will need this skill for exams and homework.

Note that if the Skeptic is correct, then trials 1–6 will always yield the responses (ranks) 5, 2, 4, 1, 6 and 3, respectively. We see that the actual observed value of R_1 for Cathy's study is $r_1 = 5 + 2 + 6 = 13$. Although we don't need it, I note that the actual observed value of R_2 is $r_2 = 4 + 1 + 3 = 8$. Thus, the mean of the ranks at the high school ($\bar{r}_1 = 13/3$) is larger than the mean of the ranks through the park ($\bar{r}_2 = 8/3$) because, as a group, Cathy's times were larger (worse) at the high school.

Table 6.5 is analogous to Table 3.4 on page 63. In the earlier table, we determined the values of $\bar{x}$, $\bar{y}$ and $u = \bar{x} - \bar{y}$ for each of the 20 possible assignments of trials to treatments. In the current chapter, our task is much easier; for each assignment we need calculate only the value of r_1. You should make sure that you follow the reasoning behind Table 6.5. (A similar problem is on the homework and might be on an exam.) For example, reading the first row of this table, we see that we are interested in assignment 1,2,3. Reading from Table 6.4 we see that the ranks are 5, 2 and 4, giving $r_1 = 5+2+4 = 11$. The information in Table 6.5 is summarized in Table 6.6, the sampling distribution of R_1 for Cathy's study. Again, make sure you can create this latter table from the former.

In fact, *any* balanced CRD with $n = 6$ units and no tied responses will yield the sampling distribution for R_1 given in Table 6.6. This is easy to see because, without ties, the ranks will be 1,

Table 6.5: The values of r_1 for all possible assignments for Cathy's CRD.

Assignment	Ranks for Treatment 1	r_1	Assignment	Ranks for Treatment 1	r_1
$1, 2, 3$	$5, 2, 4$	11	$2, 3, 4$	$2, 4, 1$	7
$1, 2, 4$	$5, 2, 1$	8	$2, 3, 5$	$2, 4, 6$	12
$1, 2, 5$	$5, 2, 6$	13	$2, 3, 6$	$2, 4, 3$	9
$1, 2, 6$	$5, 2, 3$	10	$2, 4, 5$	$2, 1, 6$	9
$1, 3, 4$	$5, 4, 1$	10	$2, 4, 6$	$2, 1, 3$	6
$1, 3, 5$	$5, 4, 6$	15	$2, 5, 6$	$2, 6, 3$	11
$1, 3, 6$	$5, 4, 3$	12	$3, 4, 5$	$4, 1, 6$	11
$1, 4, 5$	$5, 1, 6$	12	$3, 4, 6$	$4, 1, 3$	8
$1, 4, 6$	$5, 1, 3$	9	$3, 5, 6$	$4, 6, 3$	13
$1, 5, 6$	$5, 6, 3$	14	$4, 5, 6$	$1, 6, 3$	10

Table 6.6: The sampling distribution of R_1 for Cathy's CRD.

r_1	$P(R_1 = r_1)$	r_1	$P(R_1 = r_1)$
6	0.05	11	0.15
7	0.05	12	0.15
8	0.10	13	0.10
9	0.15	14	0.05
10	0.15	15	0.05

2, 3, 4, 5 and 6. Thus, we have the following **huge difference** between the test based on U and the test based on R_1:

> In the no-ties situation, for any given values of n_1 and n_2 there is only one sampling distribution for R_1, but, as we can imagine, there are an infinite number of sampling distributions for U. (Just change any response and you will likely obtain a different sampling distribution for U.)

I will follow my *pattern of presentation* that is becoming familiar to you. For Cathy's very small study—only 20 possible assignments—we are able to obtain the exact sampling distribution of our test statistic quite easily; tedious perhaps, but easy. Next, we move to a larger study: Kymn's study of rowing with its 252 possible assignments. This number of assignments is easily manageable for me (even though I am not particularly good at this), but I would never require you to examine so many assignments.

As I did in Chapter 3—because I hate looking at the results from dividing by 252—Table 6.7 is not quite the sampling distribution we want; to obtain the sampling distribution we need to divide each of the table's frequencies by 252.

Table 6.7: Frequency table for the observed values r_1 of R_1 for the 252 possible assignments for Kymn's Study.

u	Freq.	u	Freq.	u	Freq.	u	Freq.	u	Freq.	u	Freq.
15.0	1	20.5	2	24.0	6	28.0	5	31.5	10	35.0	5
16.0	1	21.0	6	24.5	10	28.5	14	32.0	6	35.5	2
17.0	2	21.5	4	25.0	6	29.0	5	32.5	6	36.0	4
18.0	3	22.0	6	25.5	14	29.5	14	33.0	6	37.0	3
19.0	4	22.5	6	26.0	5	30.0	6	33.5	4	38.0	2
19.5	2	23.0	6	26.5	14	30.5	10	34.0	6	39.0	1
20.0	5	23.5	10	27.0	5	31.0	6	34.5	2	40.0	1
				27.5	16						
										Total	252

Dawn's and Sara's studies are too large for me to obtain the exact sampling distribution of R_1. Instead, for each study I performed a simulation experiment with 10,000 reps. I will report on the results of these simulations in the next section.

6.4 Step 3: The Three Rules for Calculating the P-value

You will recall that in Chapter 5, I presented lengthy arguments in order to derive the three rules (one for each alternative) for computing the P-value. I *could* type similar arguments below for the sum of ranks test. I could, but I won't. Why not?

1. If we go to the earlier arguments and replace references to *the observed value of the test statistic* U by references to *the observed value of the test statistic* R_1 the arguments—with very minor modifications—remain valid.

2. In view of the item above, and the amount of material I want to cover in these notes, I have made the executive decision that while there is substantial educational benefit to your seeing the arguments once, seeing similar arguments repeatedly in these notes is not warranted.

Thus, without further ado, I give you the three rules for finding the P-value for our sum of ranks test.

Result 6.1 (The P-values for the sum of ranks test.) *In the rules below, remember that r_1 is the sum of the ranks of treatment 1 data for the actual data.*

- *For the alternative $>$, the P-value is equal to:*

$$P(R_1 \geq r_1) \tag{6.1}$$

- *For the alternative $<$, the P-value is equal to:*

$$P(R_1 \leq r_1) \tag{6.2}$$

- *For the alternative $\neq$, compute*

$$c = n_1(n+1)/2.$$

  - *If $r_1 = c$ then the P-value equals 1.*
  - *If $r_1 > c$ then the P-value equals:*

$$P(R_1 \geq r_1) + P(R_1 \leq 2c - r_1) \tag{6.3}$$

  - *If $r_1 < c$ then the P-value equals:*

$$P(R_1 \leq r_1) + P(R_1 \geq 2c - r_1) \tag{6.4}$$

Before I illustrate the use of these rules, let me make a brief comment about the value c (short for center) in the rule for the alternative $\neq$. The value c plays the same role that 0 played in our rule for the test statistic U. The appearance of c in the current rule is a direct result of my choosing to have the test statistic be R_1 rather than the difference of the mean ranks. If our test statistic was the difference of the mean ranks, then c would be replaced by 0 in our rule. I made the executive decision that it is better to make the test statistic simple and the rule—for the two-sided alternative only—complicated.

I will illustrate these rules with our four studies from Chapters 1 and 2.

Example 6.1 (Cathy's study.) In the answers below, please refer to the sampling distribution in Table 6.6. Recall from above that Cathy's actual r_1=13. For the alternative $\neq$ only, we also need the values:

$$c = n_1(n+1)/2 = 3(7)/2 = 10.5 \text{ and } 2c - r_1 = 2(10.5) - 13 = 8.$$

For the alternative $>$ her P-value is:

$$P(R_1 \geq 13) = 0.10 + 0.05 + 0.05 = 0.20.$$

For the alternative $<$ her P-value is:

$$P(R_1 \leq 13) = 1 - 0.05 - 0.05 = 0.90.$$

For the alternative $\neq$ her P-value is:

$$P(R_1 \geq 13) + P(R_1 \leq 8) = 0.20 + 0.20 = 0.40.$$

We found in Chapter 4 that with the test statistic equal to the difference of the means, Cathy's P-values are: 0.20 for $>$; 0.85 for $<$; and 0.40 for $\neq$. Thus, the two tests give the same P-values for $>$ and $\neq$.

Example 6.2 (Kymn's study.) In the answers below, please refer to Table 6.7. It can be shown that for Kymn's actual data, $r_1 = 40$. For the alternative $\neq$ only, we also need the values:

$$c = n_1(n+1)/2 = 5(11)/2 = 27.5 \text{ and } 2c - r_1 = 2(27.5) - 40 = 15.$$

For the alternative $>$ her P-value is:

$$P(R_1 \geq 40) = 1/252 = 0.0040.$$

For the alternative $<$ her P-value is:

$$P(R_1 \leq 40) = 1.$$

For the alternative $\neq$ her P-value is:

$$P(R_1 \geq 40) + P(R_1 \leq 15) = 2/252 = 0.0080.$$

All three of these P-values are exactly the same as the ones we found in Chapter 5 with the test statistic equal to the difference of the means.

Example 6.3 (Dawn's study.) For brevity, I will restrict attention to the alternative $>$. As shown earlier in this chapter, the observed value of R_1 is $r_1 = 133.5$. Also, recall that for Dawn's actual data, $u = \bar{x} - \bar{y} = 5.1 - 2.9 = 2.2$. Thus, there are two possibilities for the P-value:

- $P(R_1 \geq 133.5)$; and
- $P(U \geq 2.2)$.

With 184,756 possible assignments, I am not going to compute these exact probabilities! Instead, I performed a simulation experiment with 10,000 reps. Each rep, of course, selected an assignment at random from the possible assignments. Then, for the given assignment I computed both of the values r_1 and u. Below are the results I obtained:

- The relative frequency of $(R_1 \geq 133.5)$ was 0.0127; thus, the approximate P-value using ranks is 0.0127.
- The relative frequency of $(U \geq 2.2)$ was 0.0169; thus, the approximate P-value using the difference of means is 0.0169.

How should we interpret the fact that the two tests give different approximate P-values for Dawn's data? The two tests summarize the data differently, so it should be no surprise that they give somewhat different answers. Also, in my opinion, the difference between an approximation of 0.0127 and 0.0169 is not very important. The difference does suggest, however, that the test based on ranks is a bit better than the test based on comparing means. This is a tricky point, so don't worry if you don't believe me; we will see a better way to look at this issue when I introduce you to the technical concept of the power of a test.

I did something subtle in my computer simulation experiment. Did you spot it? I *could have* performed separate simulations for each test statistic. Indeed, I previously showed you the results from two different simulation experiments for Dawn's data (a simulation with 10,000 reps and then another with 100,000 reps). Thus, I could have used either (or both) of my earlier simulations for U and combined that with a new simulation for R_1. Instead, I performed **one** new simulation experiment and in this new experiment **for every assignment it selected** I evaluated both test statistics. As we will see later in these notes, the method I used gives much more precision for comparing P-values than using two separate simulations.

Why is one simulation better than two? We will see the details later, but here is the intuition. In CRDs we have been comparing treatments. It can be difficult to reach a conclusion because of variation from unit-to-unit. For example, it seems that the appetite of Bob the cat varied a great deal from day-to-day. This variation makes it difficult to see which treatment is preferred by Bob. A common theme in science and Statistics is that we learn better (more validly and more efficiently) if we can reduce variation.

In the context of computer simulations, the role of unit-to-unit variation in a CRD is played by assignment-to-assignment variation. Thus, as we will see later in these notes, by comparing U to R_1 **on the same assignments** we obtain a much more precise comparison of the two tests. In the vernacular, we avoid the possibility that one test gets *lucky* and is evaluated on *better* assignments.

Example 6.4 (Sara's study.) For brevity, I will restrict attention to the alternative $>$. It can be shown that for Sara's actual data, the observed value of R_1 is $r_1 = 1816$. Also, recall that for Sara's actual data, $u = \bar{x} - \bar{y} = 106.875 - 98.175 = 8.700$. Thus, there are two possibilities for the P-value:

- $P(R_1 \geq 1816)$; and
- $P(U \geq 8.700)$.

I performed a simulation experiment with 10,000 reps. Each rep, of course, selected an assignment at random from the possible assignments. Then, for the given assignment I computed both of the values r_1 and u. Below are the results I obtained:

- The relative frequency of $(R_1 \geq 1816)$ was 0.0293; thus, the approximate P-value using ranks is 0.0293.
- The relative frequency of $(U \geq 8.700)$ was 0.0960; thus, the approximate P-value using the difference of means is 0.0960.

Our P-values for Sara's study are dramatically different than what we found earlier. For Cathy's study, U and R_1 give exactly the same P-values for all but the alternative not supported by the data. For Kymn's study, the two tests give exactly the same P-values for all three possible alternatives. For Dawn's study the tests give different P-values for the alternative supported by the data (and also for $\neq$, although I did not show you this), but the difference is not dramatic. For Sara's study, however, the two P-values for $>$ are dramatically different. All scientists would agree that a P-value of 0.0960 is importantly different than a P-value of 0.0293.

Why are the P-values so different for Sara's data? Sadly, I must leave this question unanswered until we learn about the power of a test.

6.5 Ordinal Data

Thus far in this chapter I have presented the sum of ranks test as an alternative to the test of means. For example, for the studies of Dawn, Kymn, Sara, Cathy and others mentioned in the homework, one could use either of these tests to investigate the Skeptic's Argument. Later in these *Course Notes* when we study power we will investigate the issue of when each test is better than the other—alas, neither is universally better for **all** scientific problems—as well as why the idea of doing both tests is tricky. In this section, I introduce you to a class of scientific problems for which the sum of ranks test can be used, but the test of means should **not** be used.

I will introduce the ideas with a **type** of medical study. The data are artificial.

Example 6.5 (Artificial study of a serious disease.) A collection of hospitals serves many patients with a particular serious disease. There are two competing methods of treating these patients, call them treatment 1 and treatment 2. One hundred patients are available for study and 50 are assigned to each treatment by randomization. Each patient is treated until a response is obtained and after all 100 responses are obtained the data will be analyzed. The response is **categorical** with three possibilities:

1. The patient is cured after a short period of treatment.
2. The patient is cured after a long period of treatment.
3. The patient dies.

The (artificial) data are presented in Table 6.8.

Before I proceed, I want to acknowledge that this example is a simplification that ignores some serious issues of medical ethics. I am not qualified to discuss—or even identify—these issues, so I won't try. For the sake of this presentation let's all agree to the following ideas.

1. The three response categories are naturally **ordered**: a fast cure is preferred to a slow cure and any cure is preferred to death.
2. **For convenience** I will assign numbers to these categories: 1 for fast cure, 2 for slow cure and 3 for death.

Table 6.8: Data from an artificial study of a serious disease.

	Response			
Treatment	Fast Cure	Slow Cure	Death	Total
1	24	16	10	50
2	18	17	15	50
Total	42	33	25	100

3. **We should never** compute a mean for these responses. It would be outrageous to say that a fast cure coupled with a death is the same as two slow cures. (The mean of 1 and 3 is 2.)

4. It is fine to compute medians, but with so few possible responses, medians are not very helpful. For example, in our data, the median for each treatment is 2, which suggests they are equivalent. but clearly treatment 1 is performing better than treatment 2 in these data.

5. As we will see below, positions and ranks make sense for these data, but there are a lot of ties! (Forty-two patients respond 1; and so on.)

I will now proceed to perform the sum of ranks test on these data. Recall that first we combine the 50 responses from treatment 1 with the 50 responses from treatment 2 to obtain a total of 100 responses. Next, we sort these 100 numbers and assign a rank to each response. It is quite easy to assign these ranks because of the many ties necessitated by the presence of only three possible responses. In particular, reading from the *Total row* of Table 6.8, we find:

- The same number, 1, appears in positions 1–42. Thus, each of these 42 numbers is assigned a rank equal to the mean of these positions: $(1 + 42)/2 = 21.5$.

- The same number, 2, appears in positions 43–75. Thus, each of these 33 numbers is assigned a rank equal to the mean of these positions: $(43 + 75)/2 = 59$.

- The same number, 3, appears in positions 76–100. Thus, each of these 25 numbers is assigned a rank equal to the mean of these positions: $(76 + 100)/2 = 88$.

From the above three bullets and reading from the treatment 1 row of Table 6.8, we find:

$$r_1 = 24(21.5) + 16(59) + 10(88) = 2340.$$

We may now use Formulas 6.1–6.4 on page 125 to obtain the expressions for the P-values for the three possible alternatives. They are:

- For the alternative $>$, the P-value equals

$$P(R_1 \geq 2340).$$

- For the alternative $<$, the P-value equals

$$P(R_1 \leq 2340).$$

- For the alternative $\neq$, first we compute

  $c = n_1(n+1)/2 = 50(101)/2 = 2525$ and $2c - r_1 = 2(2525) - 2340 = 5050 - 2340 = 2710$.

  Thus, from Formula 6.4, the P-value equals

$$P(R_1 \leq 2340) + P(R_1 \geq 2710).$$

I do not know the exact sampling distribution for R_1. Therefore, I performed a computer simulation experiment with 10,000 reps on Minitab. The simulation gave me the following approximate P-values.

- For the alternative $>$, the approximate P-value equals

$$\text{Rel. Freq. } (R_1 \geq 2340) = 0.9196.$$

- For the alternative $<$, the approximate P-value equals

$$\text{Rel. Freq. } (R_1 \leq 2340) = 0.0946.$$

- For the alternative $\neq$, the approximate P-value equals

$$\text{Rel. Freq. } (R_1 \leq 2340) + \text{ Rel. Freq. } (R_1 \geq 2710) = 0.0946 + 0.0919 = 0.1865.$$

For this study, smaller responses are better. Thus, the smallest P-value is obtained for the alternative $<$, which is the alternative supported by the data. (In the data treatment 1 performed better than treatment 2.)

6.6 Computing

In Section 5.4, beginning on page 104, we learned how to use the *vassarstats* website,

`http://vassarstats.net`,

to obtain a simulation study for our test of means with test statistic given by U. I have very good news for you! If we enter the ranks into the *vassarstats* website for comparing means, we can obtain a valid simulation for the sum of ranks test, following the same steps you learned in Chapter 5. Here is why, but you don't need to know this: As I argued earlier in this chapter, a test based on R_1 is the same as a test based on the difference of the mean ranks; hence, if we enter ranks into the *vassarstats* website, it works!

Below I give you examples of using the site for three of our studies. Note that if you replicate what I do below, you will likely obtain different, but similar, approximate P-values. Also, be aware that some of the examples below are fairly tedious because they involve typing lots of data values into the site.

1. I performed a 10,000 rep simulation of Dawn's data on *vassarstats* and obtained an approximate P-value of 0.0130 for the alternative $>$, which agrees quite well with the 0.0127 I obtained with Minitab.

2. I performed a 10,000 rep simulation of Sara's data on *vassarstats* and obtained an approximate P-value of 0.0253 for the alternative $>$, which agrees reasonably well with the 0.0293 I obtained with Minitab.

3. Finally, for the artificial ordinal data in Table 6.8, I performed a 10,000 rep simulation on *vassarstats* and obtained an approximate P-value of 0.0927 for the alternative $<$, which agrees quite well with the 0.0946 I obtained with Minitab.

Table 6.9: An Example of how to obtain r_1.

Data, by treatment:				
Treatment 1:	11	14	11	
Treatment 2:	14	15	12	14

Data combined, sorted, assigned ranks;
observations from treatment 1 are bold-faced:

Position:	1	2	3	4	5	6	7
Observation:	**11**	**11**	12	**14**	14	14	15
Rank:	**1.5**	**1.5**	3	**5**	5	5	7

$r_1 = 1.5 + 1.5 + 5 = 8$

6.7 Summary

In Chapter 3 we learned about the Skeptic's Argument which states that the treatment level in a CRD is irrelevant. In Chapter 5 we learned how to investigate the validity of the Skeptic's Argument by using a statistical test of hypotheses. The test statistic in Chapter 5 is U, which tells us to compare the two sets of data by comparing their means. In this chapter we propose an alternative to test statistic U: the test statistic based on summing ranks, R_1.

The obvious question is: What are ranks? In a CRD, combine the data from the two treatments into one list. Next, sort the data (from smallest to largest, as we always do in Statistics). An example of these ideas is presented in Table 6.9, artificial data for a CRD with $n_1 = 3$ and $n_2 = 4$. In this table we have the sorted combined data, which consists of the seven numbers: 11, 11, 12, 14, 14, 14 and 15. We assign ranks to these seven numbers as follows:

- When a value is repeated (or tied; 11's and 14's in our list) all occurrences of the value receive the same rank. This common rank is the mean of the positions of these values. Thus, the 11's reside in positions 1 and 2; hence, their ranks are both $(1+2)/2 = 1.5$. Also, the 14's reside in positions 4, 5 and 6; hence, their common rank is $(4 + 5 + 6)/3 = (4 + 6)/2 = 5$.

- For each non-repeated (non-tied) value, its rank equals its position; hence, rank 3 [7] for the observation 12 [15].

The observed value, r_1 of the test statistic R_1 is obtained by summing the ranks of the data that came from treatment 1. For our current table, this means that we sum the ranks in bold-faced type:

$$r_1 = 1.5 + 1.5 + 5 = 8.$$

We can also obtain the sum of the ranks of the data from treatment 2:

$$r_2 = 3 + 5 + 5 + 7 = 20.$$

We can reduce the time we spend summing ranks if we remember that for a total of n units in a CRD, the sum of all ranks is $n(n+1)/2$. For our artificial data in Table 6.9, $n = 7$; thus, the sum of all ranks is $7(8)/2 = 28$ which agrees with our earlier $r_1 + r_2 = 8 + 20 = 28$.

For descriptive purposes, we should compare the mean ranks, which for our artificial data are:

$$\bar{r}_1 = r_1/n_1 = 8/3 = 2.67 \text{ and } \bar{r}_2 = r_2/n_2 = 20/4 = 5.$$

In words, the data from treatment 2 tend to be larger than the data from treatment 1.

We are now ready to consider the sum of ranks test. The null hypothesis is that the Skeptic is correct. There are three possible choices for the alternative: abbreviated by $>$, $<$ and $\neq$. As a practical matter, just remember that $>$ [$<$] means that—in terms of ranks—treatment 1 tends to give larger [smaller] responses than treatment 2; and that $\neq$ means that treatment 1 tends to give either larger or smaller responses than treatment 2.

In principle, the sampling distribution of R_1 is obtained exactly like the sampling distribution of U: for every possible assignment we calculate its value of r_1, on the assumption, of course, that the Skeptic is correct. For studies with a small number of possible assignments, we can obtain the exact sampling distribution of R_1. For studies with a large number of possible assignments, we can use a computer simulation experiment to obtain an approximation to the sampling distribution of R_1. In addition, in Chapter 7 we will obtain a **fancy math approximation** to the sampling distribution of R_1. By **fancy math** I mean a result based on clever theorems that have been proven by professional mathematicians.

For Sara's data we found that the P-value for the test statistic U is very different from the P-value for the test statistic R_1. This issue will be explored later when we consider the power of a test.

Finally, the sum of ranks test can be used for ordinal data. The test of means should **not** be used for ordinal data because a mean is not an appropriate summary of ordinal data.

6.8 Practice Problems

1. An unbalanced CRD with $n_1 = 4$ and $n_2 = 5$ yields the following data:

Treatment 1:	4	12	9	12	
Treatment 2:	6	3	25	12	3

   Use these data to answer the following questions.

   (a) Assign ranks to these observations.

   (b) Calculate the values of r_1, r_2 and the mean rank for each treatment. Briefly interpret the two mean ranks.

   (c) Calculate $\bar{x}$ and $\bar{y}$ for these data. Briefly interpret these means. Compare this interpretation to your interpretation of the mean ranks in part(b). Comment.

2. The purpose of this problem is to give you practice at computing exact P-values. Suppose that you conduct a balanced CRD with a total of $n = 10$ units **and** that your sampling distribution is given by the frequencies (divided by 252) in Table 6.7.

   (a) Find the exact P-values for the alternatives $>$ and $\neq$ for each of the following actual values of r_1: $r_1 = 37.0$; $r_1 = 33.5$; and $r_1 = 31.5$.

   (b) Find the exact P-values for the alternatives $<$ and $\neq$ for each of the following actual values of r_1: $r_1 = 19.0$; $r_1 = 20.5$; and $r_1 = 22.0$.

3. A CRD is performed with an ordinal categorical response. The data are below.

	Response				
Treatment	Low	Middle Low	Middle High	High	Total
1	10	6	5	4	25
2	4	7	7	7	25
Total	14	13	12	11	50

   Assign numbers 1 (Low), 2 (Middle Low), 3 (Middle High) and 4 (High) to these categories.

   (a) Assign ranks to the 50 observations.

   (b) Calculate r_1, r_2 and the two mean ranks. Comment.

   (c) Use the *vassarstats* website to obtain two approximate P-values based on a simulation with 10,000 reps. Identify the alternative for each approximate P-value.

4. I reminded you of Doug's study of the dart game 301 in Practice Problem 4 in Chapter 5, in Section 5.6 on page 110.

   Below are the ranks for Doug's data on treatment 1 (personal darts):

1.0	2.5	4.5	4.5	7.0	7.0	12.5	16.0	16.0	19.5
19.5	19.5	22.5	22.5	25.5	25.5	28.5	30.5	32.5	37.0

Below are the ranks for Doug's data on treatment 2 (bar darts):

2.5	7.0	9.5	9.5	12.5	12.5	12.5	16.0	19.5	25.5
25.5	28.5	30.5	32.5	34.5	34.5	37.0	37.0	39.0	40.0

I entered these ranks into the *vassarstats* website and obtained the following output:

> The mean for the first [second] set of ranks is 17.7 [23.3]. The approximate P-values based on 10,000 reps are: 0.0623 for one-tailed and 0.1277 for two-tailed.

(a) We saw earlier that for Doug's data, $\bar{x} = 18.6$ and $\bar{y} = 21.2$. Do the means of the ranks tell a similar or different story than the means of the data? Explain.

(b) Match the two P-values given by *vassarstats* to their alternatives. Compare these two P-values to the approximate P-values I presented in Practice Problem 4 in Chapter 5. Comment.

6.9 Solutions to Practice Problems

1. (a) I combine the data into one set of $n = 9$ numbers and sort them:

Position:	1	2	3	4	5	6	7	8	9
Data:	3	3	4	6	9	12	12	12	25
Ranks:	1.5	1.5	3	4	5	7	7	7	9

The two observations equal to 3 reside in positions 1 and 2; hence, they are both assigned the rank of $(1 + 2)/2 = 1.5$. The three observations equal to 12 reside in positions 6–8; hence, they are all assigned the rank of $(6 + 8)/2 = 7$. There are no other tied values. Thus, the rank of each remaining observation equals its position.

(b) The observations on the first treatment, 4, 12, 9 and 12, have ranks 3, 7, 5 and 7; thus, $r_1 = 3 + 7 + 5 + 7 = 22$. The sum of all nine ranks is $9(10)/2 = 45$. Thus,

$$r_2 = 45 - r_1 = 45 - 22 = 23.$$

Note that we also could obtain r_2 by summing ranks:

$$r_2 = 1.5 + 1.5 + 4 + 7 + 9 = 23.$$

The mean ranks are:

$$r_1/n_1 = 22/4 = 5.5 \text{ and } r_2/n_2 = 23/5 = 4.6.$$

The mean of the treatment 1 ranks is larger than the mean of the treatment 2 ranks. This means, in terms of ranks, that the observations on treatment 1 tend to be larger than the observations on treatment 2.

(c) The means are

$$\bar{x} = (4+9+12+12)/4 = 37/4 = 9.25 \text{ and } \bar{y} = (3+3+6+12+25)/5 = 49/5 = 9.8.$$

In terms of the means, the observations on treatment 2 are larger than the observations on treatment 1. This interpretation is the reverse of what we found for ranks.

Comment: The one unusually large observation, 25, has a pronounced effect on $\bar{y}$. In terms of ranks, 25 has the same impact that it would if it were replaced by 13; it would still be the largest observation and have rank of 9.

2. (a) First, note that $c = n_1(n + 1)/2 = 5(11)/2 = 27.5$ is smaller than all of the values of r_1. Thus, in addition to using Formula 6.1 on page 125 for the alternative $>$, we use Formula 6.3 for the alternative $\neq$. You should verify the numbers in the following display. (I will verify one of them for you, immediately below this display.)

r_1	$P(R_1 \geq r_1)$	$2c - r_1$	$P(R_1 \leq 2c - r_1)$
37.0	$7/252 = 0.0278$	18.0	$7/252 = 0.0278$
33.5	$30/252 = 0.1190$	21.5	$30/252 = 0.1190$
31.5	$58/252 = 0.2302$	23.5	$58/252 = 0.2302$

For the first entry in the '$r_1 = 37.0$' row: from Table 6.7 we find that

$$\text{Frequency } (R_1 \geq 37.0) = 3 + 2 + 1 + 1 = 7.$$

We now have the following P-values:

	Alternative	
r_1	$>$	$\neq$
37.0	0.0278	$2(0.0278) = 0.0556$
33.5	0.1190	$2(0.1190) = 0.2380$
31.5	0.2302	$2(0.2302) = 0.4604$

(b) As above, $c = 27.5$, which is larger than all of the values of r_1. Thus, in addition to using Formula 6.2 for the alternative $<$, we use Formula 6.4 for the alternative $\neq$. You should verify the numbers in the following display:

r_1	$P(R_1 \leq r_1)$	$2c - r_1$	$P(R_1 \geq 2c - r_1)$
19.0	$11/252 = 0.0437$	36.0	$11/252 = 0.0437$
20.5	$20/252 = 0.0794$	34.5	$20/252 = 0.0794$
22.0	$36/252 = 0.1429$	33.0	$36/252 = 0.1429$

For the first entry in the '$r_1 = 19.0$' row: from Table 6.7 we find that

$$\text{Frequency } (R_1 \leq 19.0) = 4 + 3 + 2 + 1 + 1 = 11.$$

We now have the following P-values:

	Alternative	
r_1	$<$	$\neq$
19.0	0.0437	$2(0.0437) = 0.0874$
20.5	0.0794	$2(0.0794) = 0.1588$
22.0	0.1429	$2(0.1429) = 0.2858$

3. (a) In the combined list of 50 observations, positions 1–14 contain the response '1;' hence, each '1' is assigned the rank of $(1+14)/2 = 7.5$. Positions 15–27 contain the response '2;' hence, each '2' is assigned the rank of $(15 + 27)/2 = 21$. Positions 28–39 contain the response '3;' hence, each '3' is assigned the rank of $(28 + 39)/2 = 33.5$. Finally, positions 40–50 contain the response '4;' hence, each '4' is assigned the rank of $(40 + 50)/2 = 45$.

(b) From part(a),

$$r_1 = 10(7.5) + 6(21) + 5(33.5) + 4(45) = 548.5 \text{ and .}$$

$$r_2 = 4(7.5) + 7(21) + 7(33.5) + 7(45) = 726.5.$$

As a partial check, the sum of all ranks must be

$$50(51)/2 = 1275 \text{ and } r_1 + r_2 = 548.5 + 726.5 = 1275.$$

The mean ranks are

$$r_1/n_1 = 548.5/25 = 21.94 \text{ and } r_2/n_2 = 726.5/25 = 29.06.$$

The observations on treatment 2 tend to be larger than the observations on treatment 1.

(c) I enter 25 and 25 for the sample sizes. Then I enter the ranks as data and click on *Calculate*. The site presents 21.94 and 29.06 as the means; this is a partial check that I entered the ranks correctly! I click on *Resample x 1000* ten times to obtain my 10,000 reps. The relative frequencies I obtained are 0.0354 for one-tailed and 0.0709 for two-tailed. Thus, the approximate P-value for $\neq$ is 0.0709. The mean ranks of the data agree with the alternative $<$; hence, the approximate P-value for $<$ is 0.0354. The site does not give an approximate P-value for $>$.

4. (a) The mean of the ranks on treatment 1 is smaller than the mean of the ranks on treatment 2; this means that, in terms of position, the data in treatment 1 tend to be smaller than the data in treatment 2. The data support the alternative $<$, not $>$. Similarly, the fact that $\bar{x} < \bar{y}$ supports the alternative $<$, not $>$. Thus, unlike in Practice Problem 1 of this section, looking at observations gives the same qualitative conclusion as looking at ranks.

(b) For the sum of ranks test, the approximate P-value for $<$ is 0.0623. In Chapter 4, I analyzed these data using the test that compares means. I reported on two simulation experiments, each with 10,000 reps (one using Minitab, one using *vassarstats*). My two approximate P-values for $<$ were 0.0426 and 0.0433. Both of these are considerably smaller than the value for the sum of ranks test. As we will learn when we study power, this suggests that comparing means is better than comparing mean ranks for Doug's data. Recall that we had found the opposite pattern for Sara's golfing data.

For the sum of ranks test, the approximate P-value for $\neq$ is 0.1277. In Chapter 4, I analyzed these data using the test that compares means. I reported on two simulation experiments, each with 10,000 reps (one using Minitab, one using *vassarstats*). My two approximate P-values for $\neq$ were 0.0844 and 0.0836. Both of these are considerably smaller than the value for the sum of ranks test.

Table 6.10: Artificial data for Homework Problem 1.

Treatment 1:	7	7	8	8	9	14	14	14
Treatment 2:	1	3	4	5	7	8	11	11

Table 6.11: Frequency table for the values r_1 of R_1 for the 252 possible assignments for a balanced CRD with $n = 10$ units and 10 distinct response values.

r_1	Freq.	r_1	Freq.	r_1	Freq.	r_1	Freq.	r_1	Freq.	r_1	Freq.
15	1	19	5	23	14	28	20	33	11	37	3
16	1	20	7	24	16	29	19	34	9	38	2
17	2	21	9	25	18	30	18	35	7	39	1
18	3	22	11	26	19	31	16	36	5	40	1
				27	20	32	14				
										Total	252

6.10 Homework Problems

1. Table 6.10 presents artificial data for a balanced CRD with a total of $n = 16$ units.

   (a) Calculate the values of r_1, r_2 and the two mean ranks. Comment.

   (b) Use the *vassarstats* website to perform a 10,000 rep simulation experiment. Use your results to obtain two approximate P-values and identify the alternative corresponding to each approximate P-value.

2. Table 6.11 presents frequencies for the sampling distribution of R_1 for **every** balanced CRD with $n = 10$ units and 10 distinct response values.

   (a) Find the exact P-values for the alternatives $>$ and $\neq$ for each of the following actual values of r_1: $r_1 = 37$; $r_1 = 33$; and $r_1 = 31$.

   (b) Find the exact P-values for the alternatives $<$ and $\neq$ for each of the following actual values of r_1: $r_1 = 17$; $r_1 = 21$; and $r_1 = 26$.

3. A CRD is performed with an ordinal categorical response. The data are below.

	Response			
Treatment	Disagree	Neutral	Agree	Total
1	8	7	5	20
2	3	5	7	15
Total	11	12	12	35

   Assign numbers 1 (Disagree), 2 (Neutral) and 3 (Agree) to these categories.

(a) Calculate r_1, r_2 and the two mean ranks. Comment.

(b) Use the *vassarstats* website to obtain two approximate P-values based on a simulation with 10,000 reps. Identify the alternative for each approximate P-value.

4. In Homework Problem 4 of Chapter 5 (Section 5.8 on page 115) you performed a test of means on Reggie's dart study introduced in Chapter 1. Below are the ranks for Reggie's data on treatment 1.

6.0	7.0	9.0	13.5	15.5	15.5	18.5	18.5	23.5	23.5
25.5	27.0	28.5	28.5	30.0					

Below are the ranks for Reggie's data on treatment 2.

1.0	2.0	3.0	4.0	5.0	8.0	10.0	11.5	11.5	13.5
18.5	18.5	21.0	22.0	25.5					

Enter these ranks into the *vassarstats* website and obtain two approximate P-values based on a simulation with 10,000 reps. Use the output to answer the following questions.

(a) What is the mean of the ranks on treatment 1? What is the mean of the ranks on treatment 2? Compare these means and comment.

(b) Match each approximate P-value for the sum of ranks test obtained from *vassarstats* to its alternative.

(c) Compare each sum of ranks test P-value to its corresponding P-value for a comparison of means that you found in doing the Chapter 5 Homework.

Chapter 7

Visualizing a Sampling Distribution

Let's review what we have learned about sampling distributions. We have considered sampling distributions for the test of means (test statistic is U) and the sum of ranks test (test statistic is R_1). We have learned, in principle, how to find an exact sampling distribution. I say *in principle* because if the number of possible assignments is large, then it is impractical to attempt to obtain an exact sampling distribution.

We have learned an excellent way to approximate a sampling distribution, namely a computer simulation experiment with $m = 10{,}000$ reps. We can calculate a nearly certain interval to assess the precision of any given approximation and, if we are not happy with the precision, we can obtain better precision simply by increasing the value of m. Computer simulations are a powerful tool and I am more than a bit sad that they were not easy to perform when I was a student many decades ago. (We had to walk uphill, through the snow, just to get to the large building that housed **the** computer and then we had to punch zillions of cards before we could submit our programs.)

Before computer simulations were practical, or even before computers existed, statisticians and scientists obtained approximations to sampling distributions by using what I will call **fancy math** techniques. We will be using several fancy math methods in these notes.

Fancy math methods have severe limitations. For many situations they give poor approximations and, unlike a computer simulation, you cannot improve a fancy math approximation simply by increasing the value of m; there is nothing that plays the role of m in a fancy math approximation. Also, there is nothing like the *nearly certain interval* that will tell us the likely precision of a fancy math approximation.

Nevertheless, fancy math approximations are very important and can be quite useful; here are two reasons why:

1. Do not think of *computer simulations* and *fancy math* as an *either/or* situation. We can, and often will, use them together in a problem. For example, a simple fancy math argument will often show that *one* computer simulation experiment can be applied to many—sometimes an infinite number of—situations. We will see many examples of this phenomenon later in these *Course Notes*.

2. Being educated is **not** about acquiring lots and lots of facts. It is more about seeing how lots and lots of facts *relate to each other* or *reveal an elegant structure in the world*. Computer

Table 7.1: The sampling distribution of R_1 for Cathy's CRD.

r_1	$P(R_1 = r_1)$	r_1	$P(R_1 = r_1)$
6	0.05	11	0.15
7	0.05	12	0.15
8	0.10	13	0.10
9	0.15	14	0.05
10	0.15	15	0.05

simulations are very good at helping us acquire *facts*, whereas fancy math helps us see how these facts fit together.

Fancy math results can be very difficult to prove and these proofs are not appropriate for this course. Many of these results, however, can be motivated with pictures. This begs the question: Which pictures? The answer: Pictures of sampling distributions.

Thus, our first goal in this chapter is to learn how to draw a particular picture, called the **probability histogram**, of a sampling distribution.

7.1 Probability Histograms

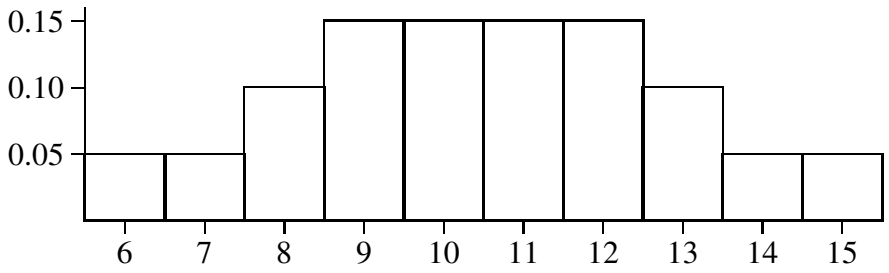
As the name suggest, a **probability histogram** is similar to the histograms we learned about in Chapter 2. For example, just like a histogram for data, a probability histogram is comprised of rectangles on the number line. There are some important differences, however. First, a motivation for our histograms in Chapter 2 was to group data values in order to obtain a better picture. By contrast, we **never** group values in a probability histogram. Second, without grouping, we don't need an endpoint convention for a probability histogram and, as a result, we will have a new way to place/locate its rectangles.

The total area of the rectangles in a probability histogram equals 1, which is a feature shared by density histograms of Chapter 2. The reason? Density histograms use area to represent relative frequencies of data; hence, their total area is one. Probability histograms use area to represent probabilities; hence, their total area equals the total probability, one.

Table 7.1 presents the sampling distribution of R_1 for Cathy's study of running. (Remember: There were no ties in Cathy's six response values.) This table was presented in Chapter 6. Its probability histogram is presented in Figure 7.1. Look at it briefly and then read my description below of how it was created.

First, some terminology. Thus far in these *Course Notes* our sampling distributions have been for test statistics, either U or R_1. In general, we talk about a sampling distribution for a **random variable** X, with observed value x. Here is the idea behind the term *random variable*. We say *variable* because we are interested in some feature that has the potential to vary. We say *random* because the values that the feature might yield are described by probabilities. Both of our test

Figure 7.1: The probability histogram for the sampling distribution in Table 7.1.



statistics are special cases of random variables and, hence, are covered by the method described below.

1. On a horizontal number line, mark all possible values, x, of the random variable X.

   For the sampling distribution in Table 7.1 these values of x (r_1) are 6, 7, 8, ... 15 and they are marked in Figure 7.1.

2. Determine the value of δ (lower case Greek delta) for the random variable of interest. The number δ is the smallest distance between any two consecutive values of the random variable.

   For the sampling distribution in Table 7.1, the distance between consecutive values is always 1; hence, $\delta = 1$.

3. Above each value of x, draw a rectangle, with its center at x, its base equal to δ and its height equal to $P(X = x)/\delta$.

   In the current example, $\delta = 1$, making the height of each rectangle equal to the probability of its center value.

For a probability histogram the area of a rectangle equals the probability of its center value, because:

$$\text{Area of rectangle centered at } x = \text{ Base } \times \text{ Height } = \delta \times \frac{P(X = x)}{\delta} = P(X = x).$$

In the previous chapter we studied the sum of ranks test with test statistic R_1. In all ways mathematical, this test statistic is much easier to study if there are no ties in the data. I will now show how the presence of one tie affects the probability histogram.

Example 7.1 (A small CRD with two values tied.) Table 7.2 presents the sampling distribution of R_1 for a balanced CRD with a total of $n = 6$ units with one particular pair of tied observations: the two smallest observations are tied and the other four observations are not. Thus, the ranks are: 1.5, 1.5, 3, 4, 5 and 6. If you feel a need to have more practice at determining sampling distributions, you may verify the entries in this table. Otherwise, I suggest you trust me on it. The probability histogram for this sampling distribution is presented in Figure 7.2. I will walk you through the three steps to create this picture.

Table 7.2: The sampling distribution of R_1 for a balanced CRD with a total of $n = 6$ units with ranks 1.5, 1.5, 3, 4, 5 and 6.

r_1	$P(R_1 = r_1)$	r_1	$P(R_1 = r_1)$	r_1	$P(R_1 = r_1)$
6.0	0.05	9.5	0.10	12.5	0.10
7.0	0.05	10.5	0.20	13.0	0.05
8.0	0.05	11.5	0.10	14.0	0.05
8.5	0.10	12.0	0.05	15.0	0.05
9.0	0.05				

1. On a horizontal number line, mark all possible values, x, of the random variable X.

   The 13 possible values of R_1 in Table 7.2 are marked, but not always labeled, in Figure 7.2.

2. Determine the value of δ for the random variable of interest. The number δ is the smallest distance between any two consecutive values of the random variable.

   This is trickier than it was in our first example. Sometimes the distance between consecutive values is 1 and sometimes it is 0.5. Thus, $\delta = 0.5$.

3. Above each value of x, draw a rectangle, with its center at x, its base equal to δ and its height equal to the $P(X = x)/\delta$.

   In the current example, $\delta = 0.5$, making the height of each rectangle equal to twice the probability of its center value.

Now that we have the **method** of constructing a probability histogram for a random variable, let's look at the two pictures we have created, Figures 7.1 and 7.2. Both probability histograms are symmetric with point of symmetry at 10.5. (It can be shown that the sampling distribution for R_1 is symmetric if the CRD is balanced and/or there are no ties in the data set.)

Figure 7.1 is, in my opinion, much more *well-behaved* than Figure 7.2. Admittedly, *well-behaved* sounds a bit subjective; here is what I mean. Figure 7.1 has one peak—four rectangles wide, but still only one peak—and no gaps. By contrast, Figure 7.2 has one dominant peak (at 10.5), eight lesser peaks (at 6, 7, 8.5, 9.5, 11.5, 12.5, 14 and 15) and six gaps (at 6.5, 7.5, 10.0, 11.0, 13.5 and 14.5).

In general, for the sum of ranks test, if there are no ties in the combined data set, the probability histogram for the sampling distribution of R_1 is symmetric, with one peak and no gaps. If there is as few as one pair of tied values in the combined data set, the probability histogram for the sampling distribution of R_1 might not be very nice! Or it might be, as our next example shows.

Example 7.2 (A small CRD with three values tied.) I have a balanced CRD with a total of six units. My six ranks are: 2, 2, 2, 4, 5 and 6. (An exercise for you: create data that would yield these ranks.) The sampling distribution for R_1 is given in Table 7.3 and its probability histogram is presented in Figure 7.3.

Figure 7.2: The probability histogram for sampling distribution in Table 7.2.

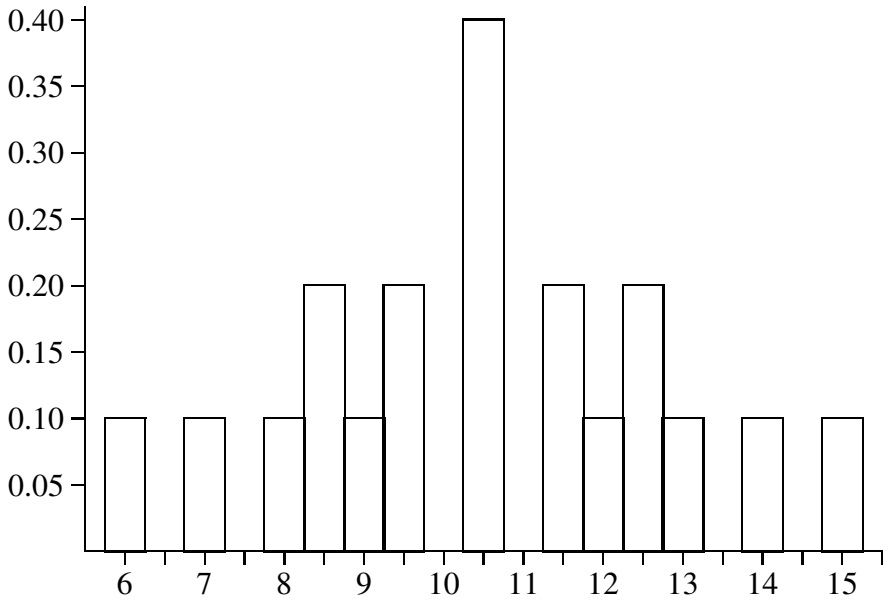


Table 7.3: The sampling distribution of R_1 for a balanced CRD with ranks 2, 2, 2, 4, 5 and 6.

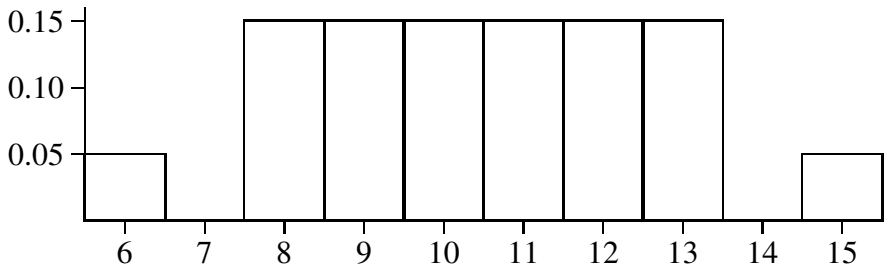
r_1	$P(R_1 = r_1)$	r_1	$P(R_1 = r_1)$
6	0.05	11	0.15
8	0.15	12	0.15
9	0.15	13	0.15
10	0.15	15	0.05

If you compare this table and figure to the table and figure for Cathy's *no-ties* data—Table 7.1 and Figure 7.1—you see that the effect of the three tied observations is quite minimal. Two of the 20 values of r_1 in Cathy's distribution, a 7 and a 14, are replaced by an 8 and a 13 in the current study.

To summarize, when there are no ties in the combined data set, our probability histograms are of a type: they are symmetric with one peak and no gaps. We have seen one such probability histogram in Figure 7.1 and will see another one later in this chapter in Figure 7.5.

We have seen two probability histograms for the situation in which there is at least one tie in the combined data set. The variety of such pictures (these two as well as other possible ones) is too great for us to devote more time to the topic. Thus, I propose the following agreement between you and me. I will focus on the no-ties pictures to motivate the fancy math approximate method I present later in this chapter. The fancy math method can be used when there are ties; in fact, there is an explicit *adjustment term* in the method that comes into play only if there are ties. As with all approximation methods in Statistics, the method's utility depends on how close the approximate

Figure 7.3: The probability histogram for sampling distribution in Table 7.3.



answers are to the exact answers; this utility does **not** depend on how cleverly—or clumsily—I motivate it.

Before we proceed further, let me acknowledge a point. There is **no unique way to draw a picture of a sampling distribution**. You might be able to create a picture that you prefer to the probability histogram. Thus, I do **not** present a probability histogram as **the** correct picture for a sampling distribution. You will see soon, however, that our probability histogram is very good at motivating the fancy math approximation of this chapter. Indeed, our definition of a probability histogram is very good at motivating many of the the fancy math approximations in these *Course Notes*.

Look at the rectangle centered at $r_1 = 10$ in Figure 7.1. Its area is its base times its height: $1 \times 0.15 = 0.15$. Thus, because area equals probability, $P(R_1 = 10) = 0.15$. Here is my point: the probability belongs to the value 10, but the probability histogram visually *spreads the probability* over the entire base of the rectangle, from 10.5 to 11.5. This *spreading* has no real meaning, but, as we will see, it does help motivate our fancy math approximation.

I have one more point to make before we move on to the next section. Knowing the sampling distribution of a test statistic (more generally, a random variable) is mathematically equivalent to knowing its probability histogram. Thus, approximating a sampling distribution is equivalent to approximating its probability histogram. Thus, below we will learn how to approximate a probability histogram.

7.2 The Mean and Standard Deviation of R_1

We learned in Chapter 1 that a set of numbers can be summarized—admittedly, sometimes poorly—by its mean. We also learned that the mean of a set of numbers can be visualized as the center of gravity of its dot plot. It is also possible to summarize the sampling distribution of a random variable by its mean. For example, consider the sampling distribution for R_1 given in Table 7.1. Recall that there are 20 possible assignments for the CRD in question and that each assignment yields a value of r_1. From the sampling distribution, we can infer that these 20 values are, after sorting:

$$6, 7, 8, 8, 9, 9, 9, 10, 10, 10, 11, 11, 11, 12, 12, 12, 13, 13, 14, 15.$$

(For example, $P(R_1 = 8) = 0.10$; thus, 10% of the 20 assignments—two— yield the value $r_1 = 8$.) If we sum these 20 values and divide by 20, we find that the mean of these 20 values is 10.5. For a sampling distribution, we refer to the mean by the Greek letter μ (pronounced 'mu' with a long 'u'). Thus, $\mu = 10.5$ for the sampling distribution in Table 7.1.

Technical Note: For a random variable, X, with observed value x, we have the following formula for μ:

$$\mu = \sum_x xP(X = x), \tag{7.1}$$

where the sum is taken over all possible values x of X. We won't ever use Equation 7.1; thus, you may safely ignore it. I have included it for completeness only.

For our method of creating a probability histogram, it can be shown that μ is equal to the center of gravity of the probability histogram. (Remember: our probability histogram, unlike the histograms for data introduced in Chapter 2, do not group values into categories.) Thus, we can see immediately from Figures 7.1—7.3 that all three probability histograms—and, hence, all three sampling distributions—have $\mu = 10.5$.

The following mathematical result is quite useful later in this chapter. I will not prove it and you need not be concerned with its proof.

Result 7.1 (Mean of R_1.) *Suppose that we have a CRD with n_1 units assigned to the first treatment and a total of n units. The mean, μ, of the sampling distribution of R_1 is given by the following equation.*

$$\mu = n_1(n + 1)/2 \tag{7.2}$$

Note that this result is true whether or not there are ties in the combined data set. For example, suppose that $n_1 = 3$ and $n = 6$ in Equation 7.2. We find that

$$\mu = 3(6 + 1)/2 = 10.5,$$

in agreement with what we could see in Figures 7.1—7.3.

Here is the important feature in Equation 7.2. Even though—as we have seen repeatedly—it can be difficult to obtain exact probabilities for R_1, it is very easy to calculate the mean of its sampling distribution.

In order to use our fancy math approximation, we also need to be able to calculate the variance of a sampling distribution. You recall that the word variance was introduced in Chapter 1. For a set of data, the variance measures the amount of spread, or variation, in the data. Let's review how we obtained the variance in Chapter 1.

In Chapter 1 we began by associating with each response value, e.g., x_i for data from treatment 1, its deviation: $(x_i - \bar{x})$. For a sampling distribution, the deviation associated with possible value x is $(x - \mu)$. Next, in Chapter 1 we squared each deviation; we do so again, obtaining the squared deviation $(x-\mu)^2$. Next, we sum the squared deviations over all possible assignments, and finally, for sampling distributions, we follow the method mathematicians use for data and divide the sum of squared deviations by the total number of possible assignments; i.e., unlike with data, we do not subtract one. (Trust me, the reason for this is not worth the time needed to explain it.) The result, the mean of these squared deviations, is called the variance of the sampling distribution

and is denoted by σ^2. (Note: σ is the lower case Greek letter sigma.) Following the ideas of Chapter 1, the positive square root of the variance, σ, is called the standard deviation of the sampling distribution.

Technical Note: For a random variable, X, with observed value x, we have the following formula for the variance σ^2:

$$\sigma^2 = \sum_x (x - \mu)^2 P(X = x), \tag{7.3}$$

where the sum is taken over all possible values x of X. We won't ever use Equation 7.3; thus, you may safely ignore it. I have included it for completeness only.

Statisticians prefer to focus on the formula for variance, as we do in Equation 7.3, rather than a formula for the standard deviation because the latter will have that nasty square root symbol. In our fancy math approximation, however, we will need to use the standard deviation. We always calculate the variance first; then take its square root to obtain the standard deviation.

I will present two formulas for the variance of the sampling distribution of R_1: the first is for the situation when there are no ties in the combined data set. If there are ties, then the second formula should be used.

Result 7.2 (Variance of R_1 when there are no ties.) *Suppose that we have a CRD with n_1 units assigned to the first treatment, n_2 units assigned to the second treatment and a total of n units. Suppose, in addition, that there are no tied values in our combined list of n observations. The variance of the sampling distribution of R_1 is given by the following equation.*

$$\sigma^2 = \frac{n_1 n_2 (n+1)}{12}. \tag{7.4}$$

For example, suppose we have a balanced CRD with $n = 6$ units and no ties; for example, Cathy's study. Using Equation 7.4 we see that the variance of R_1 is

$$\sigma^2 = \frac{3(3)(7)}{12} = 5.25 \text{ and } \sigma = \sqrt{5.25} = 2.291.$$

Next, suppose we have a balanced CRD with $n = 10$ units and no ties. Using Equation 7.4 we see that the variance of R_1 is

$$\sigma^2 = \frac{5(5)(11)}{12} = 22.917 \text{ and } \sigma = \sqrt{22.917} = 4.787.$$

When there are ties in the combined data set, then we need to use the following more complicated formula for the variance. Note that I will define the symbol t_i below the Result.

Result 7.3 (Variance of R_1 when there are ties.) *Suppose that we have a CRD with n_1 units assigned to the first treatment, n_2 units assigned to the second treatment and a total of n units. Suppose, in addition, that there is at least one pair of tied values in the combined list of n observations. The variance of the sampling distribution of R_1 is given by the following equation.*

$$\sigma^2 = \frac{n_1 n_2 (n+1)}{12} - \frac{n_1 n_2 \sum (t_i^3 - t_i)}{12n(n-1)}. \tag{7.5}$$

In this equation, the t_i's require a bit of explanation. As we will see, each t_i is a positive count. Whenever a $t_i = 1$ then the term in the sum, $(t_i^3 - t_i)$ will equal 0. This fact has two important consequences:

- All of the values of $t_i = 1$ can (and should) be ignored when evaluating Equation 7.5.
- If all $t_i = 1$ then Equation 7.5 is exactly the same as Equation 7.4. As we will see soon, saying that all $t_i = 1$ is equivalent to saying that there are no ties in the combined data set.

It will be easier for me to explain the t_i's with an example.

Recall Kymn's study. She performed a balanced CRD with $n = 10$ trials. In her combined data set she had nine distinct numbers: Eight of these numbers occurred once in her combined data set and the remaining number occurred twice. This means that Kymn's t_i's consisted of eight 1's and one 2. Thus, the variance of R_1 for Kymn's study is

$$\sigma^2 = \frac{5(5)(11)}{12} - \frac{5(5)(2^3 - 2)}{12(10)(9)} = 22.917 - \frac{150}{1080} = 22.778 \text{ and } \sigma = \sqrt{22.778} = 4.773.$$

For this example, the presence of one pair of tied values has very little impact on the value of the variance. Indeed, the standard deviation for Kymn's study is only 0.29% smaller than the standard deviation when there are no ties (which, recall, is 4.787). You can well understand why many statisticians ignore a small number of ties in a combined data set. When the response is ordinal and categorical, however, ties *can* have a notable impact on the standard deviation, as our next computation shows.

Recall the data in Table 6.8 on page 129. These artificial data come from a balanced CRD with $n = 100$ subjects and only three response values (categories). Recall that 42 subjects gave the first response, 33 subjects gave the second response and 25 subjects gave the third response. Thus, the t_i's are 42, 33 and 25. I will plug these values into Equation 7.5. First, I will calculate the value of the expression containing the t_i's:

$$\sum(t_i^3 - t_i) = (42^3 - 42) + (33^3 - 33) + (25^3 - 25) = \text{ 74,046 } + \text{ 35,904 } + \text{ 15,600 } = \text{ 125,550.}$$

Thus, the variance is:

$$\sigma^2 = \frac{50(50)(101)}{12} - \frac{50(50)(125{,}550)}{12(100)(99)} = \text{ 21,041.67 } - \text{ 2,642.05 } = \text{ 18,399.62.}$$

The correct standard deviation—taking into account ties—is $\sqrt{18{,}399.62} = 135.65$. It is smaller than the answer one would obtain by ignoring the effect of ties: $\sqrt{21{,}041.67} = 145.06$. The former is 6.49% smaller than the latter and this difference has an important impact on our fancy math approximation, as you will see later in this chapter.

7.3 The Family of Normal Curves

Do you recall π, the famous number from math? It is the ratio of the circumference to the diameter of a circle. Another famous number from math is e, which is the limit as n goes to infinity of

$(1 + 1/n)^n$. As decimals, $\pi = 3.1416$ and $e = 2.7183$, both approximations. If you want to learn more about π or e, go to Wikipedia. If you do not want to learn more about them, that is fine too.

Let μ denote any real number—positive, zero or negative. Let σ denote any positive real number. In order to avoid really small type, when t represents a complicated expression, we write e^t as $\exp(t)$. Consider the following function.

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp[\frac{-(x-\mu)^2}{2\sigma^2}], \text{ for all real numbers } x. \tag{7.6}$$

The graph of the function f is called the **Normal curve** with parameters μ and σ; it is pictured in Figure 7.4. By allowing μ and σ to vary, we generate the family of Normal curves. We use the terminology:

the N(μ, σ) curve

to designate the Normal curve with parameters μ and σ. Thus, for example, the N(25,8) curve is the Normal curve with mean $\mu = 25$ and standard deviation $\sigma = 8$.

Below is a list of important properties of Normal curves.

1. The total area under a Normal curve is one.

2. A Normal curve is symmetric about the number μ. Clearly, μ is the center of gravity of the curve, so we call it the mean of the Normal curve.

3. It is possible to talk about the spread in a Normal curve just as we talked about the spread in a sampling distribution or its probability histogram. In fact, one can define the standard deviation as a measure of spread for a curve and if one does, then the standard deviation for a Normal curve equals its σ.

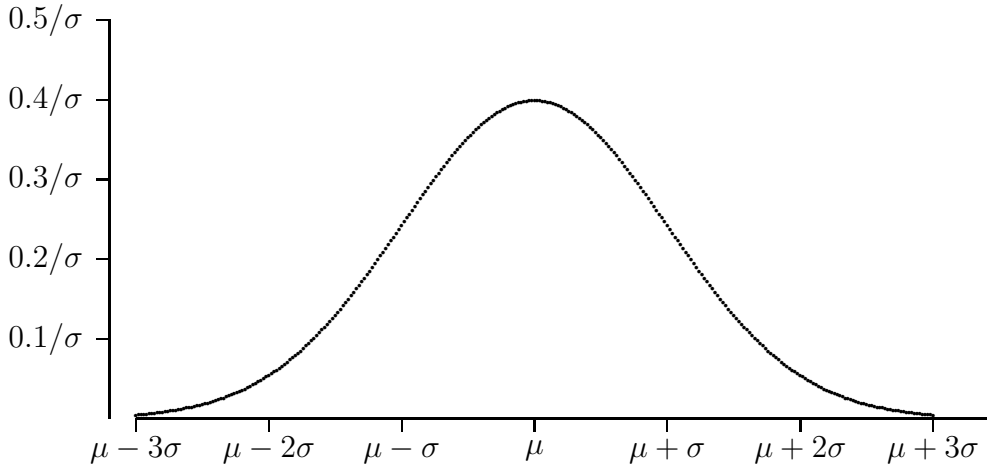
4. You can now see why we use the symbols μ and σ for the parameters of a Normal curve: μ is the mean of the curve and σ is its standard deviation.

5. A Normal curve has points of inflection at $\mu + \sigma$ and $\mu - \sigma$. If you don't know what a point of inflection is, here goes: it is a point where the curve changes from 'curving downward' to 'curving upward.' I only mention this because: If you see a picture of a Normal curve you can immediately see μ, its point of symmetry. You can also see its σ as the distance between μ and either point of inflection.

6. The Normal curve with $\mu = 0$ and $\sigma = 1$—that is, the N(0,1) curve—is called the **Standard Normal curve**.

Statisticians often want to calculate areas under a Normal curve. Fortunately, there exists a website that will calculate areas for us; I will present instructions on how to use this site in Section 7.4.

Figure 7.4: The Normal curve with parameters μ and σ; i.e., the N(μ, σ) curve.



7.3.1 Using a Normal Curve to obtain a fancy math approximation

Table 7.4 presents the exact sampling distribution for R_1 for a balanced CRD with $n = 10$ units and no tied values. (Trust me; it is no fun to verify this!) Figure 7.5 presents its probability histogram. In this figure, I have shaded the rectangles centered at 35, 36, ..., 40. If we remember that in a probability histogram, area equals probability, we see that the total area of these six shaded rectangles equals $P(R_1 \geq 35)$. If, for example, we performed a CRD and the sampling distribution is given by Table 7.4 and the actual value of r_1 is 35, then $P(R_1 \geq 35)$ would be the P-value for the alternative $>$; i.e., the area of the shaded region in Figure 7.5 would be the P-value for the alternative $>$.

Now look at Figures 7.5 and 7.4. Both pictures are symmetric with a single peak and are *bell-shaped*. These similarities *suggest* that using a Normal curve to approximate the probability histogram might work well. There is no reason to argue my visual interpretation; let's try it and see what happens.

The idea is to use one of the members of the family of Normal curves to approximate the probability histogram in Figure 7.5. Which one? Because the probability histogram is symmetric around 27.5, we know that its mean is 27.5. We could also obtain this answer by using Equation 7.2. As we found on page 148, the standard deviation of this probability histogram is 4.787. It seems sensible that we should use the Normal curve with $\mu = 27.5$ and $\sigma = 4.787$; i.e., the Normal curve with the same center and spread as the probability histogram.

I am almost ready to show you the fancy math approximation to $P(R_1 \geq 35)$. Look at the shaded rectangles in Figure 7.5. The left boundary of these rectangles is at 34.5 and the right boundary is at 40.5. The approximation I advocate is to find the area to the right of 34.5 under the Normal curve with $\mu = 27.5$ and $\sigma = 4.787$. There are two things to note about my approximation:

Table 7.4: The sampling distribution of R_1 for the 252 possible assignments of a balanced CRD with $n = 10$ units and 10 distinct response values.

r_1	$P(R_1 = r_1)$	r_1	$P(R_1 = r_1)$	r_1	$P(R_1 = r_1)$	r_1	$P(R_1 = r_1)$	r_1	$P(R_1 = r_1)$
15	1/252	20	7/252	25	18/252	31	16/252	36	5/252
16	1/252	21	9/252	26	19/252	32	14/252	37	3/252
17	2/252	22	11/252	27	20/252	33	11/252	38	2/252
18	3/252	23	14/252	28	20/252	34	9/252	39	1/252
19	5/252	24	16/252	29	19/252	35	7/252	40	1/252
				30	18/252			Total	1

- We begin at the value 34.5, not 35. This adjustment is called a **continuity correction**.
- Instead of ending the approximation at the value 40.5 (the right extreme of the rectangles), we take the approximation all the way to the right. As *Buzz Lightyear* from *Toy Story* would say, "To infinity and beyond." Well, statisticians don't attempt to go beyond infinity.

The area I seek: the area to the right of 34.5 under the Normal curve with $\mu = 27.5$ and $\sigma = 4.787$ is 0.0718. When you read Section 7.4 you will learn how to obtain this area. I don't show you now because I don't want to interrupt the flow of the narrative.

Is this approximation any good? Well, we can answer this question because it is possible to compute the exact $P(R_1 \geq 35)$ from Table 7.4. Reading from this table, we find:

$$P(R_1 \geq 35) = (7 + 5 + 3 + 2 + 1 + 1)/252 = 19/252 = 0.0754.$$

Let me make three comments about this approximation.

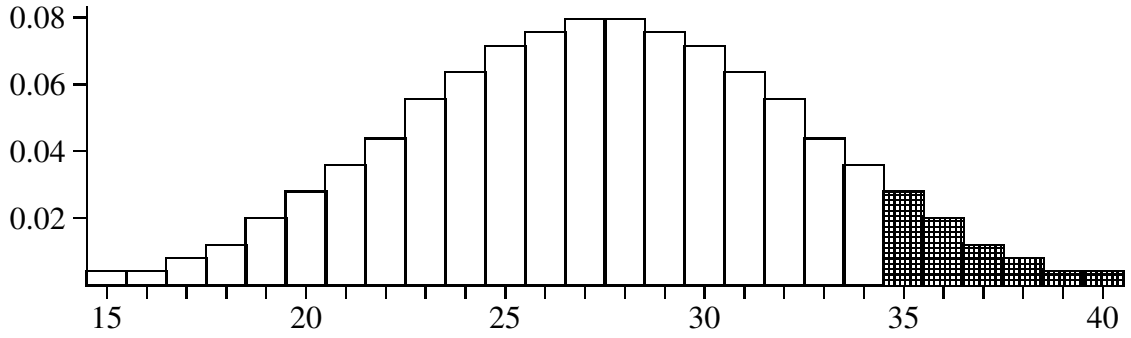
1. The approximation (0.0718) is pretty good; it is smaller than the exact probability (0.0754) by 0.0036. Math theory tells us that this approximation tends to get better as the total number of trials, n, becomes larger. This is a remarkably good approximation for an n that is so small that we don't really need an approximation (because it was 'easy' to find the exact probability).

2. This example shows that the continuity correction is very important. If we find the area under the Normal curve to the right of 35.0—i.e., if we do not adjust to 34.5— the area is 0.0586, a pretty bad approximation of 0.0754.

3. If we find the area between 34.5 and 40.5—the region suggested by the rectangles, instead of simply the area to the right of 34.5 which I advocate—the answer is 0.0685; which is a worse approximation than the one I advocate.

We have spent a great deal of effort looking at $P(R_1 \geq 35)$. Table 7.5 looks at several other examples. Again, after you read Section 7.4 you will be able to verify the approximations in this table, if you desire to do so.

Figure 7.5: The probability histogram for the sampling distribution in Table 7.4. The area of the shaded rectangles equals $P(R_1 \geq 35)$.



There is a huge amount of information in this table. I don't want you to stare at it for 15 minutes trying to absorb everything in it! Instead, please note the following features.

1. If you scan down the 'Exact' and 'Normal' columns, you see that the numbers side-by-side are reasonably close to each other. Thus, **the approximation is never terrible**.

2. Interestingly, the best approximation is the value 0.0473 which differs from the exact answer, 0.0476, by only 0.0003. We know from the definition of *statistical significance* in Chapter 5 that statisticians are particularly interested in P-values that are close to 0.05. Thus, in a sense, for this example the approximation is best when we especially want it to be good.

3. This table—and a bit additional work—illustrates how difficult it is to compare different methods of approximating a probability. I am a big fan of the continuity correction and want you to always use it. Intellectual honesty, however, requires me to note the following. From our table we see that for the event $(R_1 \geq 40)$ the exact probability is $1/252 = 0.0040$. This is considerably smaller (in terms of ratio) than the approximate probability of 0.0061. For this event, we actually get a better approximation if we don't use the continuity correction: the area under the Normal curve to the right of 40 is 0.0045. This is a much better answer than the one in the table.

   This result reflects a general pattern. In the extremes of sampling distributions, the Normal curve approximation often is better without the continuity correction. The continuity correction, however, tends to give much better approximations for probabilities in the neighborhood of 0.05. Because statisticians care so much about 0.05 and its neighbors, I advocate using the continuity correction. But I **cannot say** that it is always better to use the continuity correction, because sometimes it is not.

I have two final extended comments before I leave this section.

1. I have focused on approximating '$\geq$' probabilities, such as $P(R_1 \geq 35)$. In these situations we obtain the continuity correction by *subtracting 0.5* from the value of interest, in the previous sentence, 35. The thing to remember is that our motivation for subtracting 0.5

Table 7.5: Several Normal curve approximations for the sampling distribution in Table 7.4.

Event	Exact Prob.	Normal Approx.	Event	Exact Prob.	Normal Approx.	Event	Exact Prob.	Normal Approx.
$R_1 \geq 29$	0.4206	0.4173	$R_1 \geq 33$	0.1548	0.1481	$R_1 \geq 37$	0.0278	0.0300
$R_1 \geq 30$	0.3452	0.3380	$R_1 \geq 34$	0.1111	0.1050	$R_1 \geq 38$	0.0159	0.0184
$R_1 \geq 31$	0.2738	0.2654	$R_1 \geq 35$	0.0754	0.0718	$R_1 \geq 39$	0.0079	0.0108
$R_1 \geq 32$	0.2103	0.2017	$R_1 \geq 36$	0.0476	0.0473	$R_1 \geq 40$	0.0040	0.0061

comes from the probability histogram in Figure 7.5. We *see* that the rectangle centered at 35 has a left boundary of 34.5. Thus, we subtract 0.5 to move from 35 to 34.5 and include all of the rectangle. Note that we don't need to actually draw the probability histogram; from its definition we know that $\delta = 1$ is the base of the rectangle. Hence, to move from the center, 35, to its left boundary, we move one-half of δ to the left.

Suppose, however, that we want to find an approximate P-value for the alternative $<$ or $\neq$. In either of these cases we would need to deal with a '$\leq$' probability, say $P(R_1 \leq 21)$. Now we want the rectangle centered at 21 and all rectangles to its left. Thus, the continuity correction changes 21 to 21.5.

You have a choice as to how to remember these facts about continuity corrections: you may memorize that '$\geq$' means subtract and that '$\leq$' means add. Personally, I think it is better to think of the picture and deduce the direction by reasoning.

2. In many, but not all, of our Normal curve approximations in these *Course Notes*, $\delta = 1$ and the size of the continuity correction is 0.5. But as we saw above, if there are ties in our combined data set, then for R_1 we could have $\delta = 0.5$ or we could have $\delta = 1$. In addition, when there are ties, the probability histogram has gaps and it's not clear (trust me on this!) how, if at all, gaps should affect a continuity correction. **For simplicity,** I declare that whenever we want to use a Normal curve to approximate the sampling distribution of R_1, we will always adjust by 0.5, whether δ equals 1 or 0.5.

7.4 Computing

7.4.1 Areas Under any Normal Curve

My objective in this subsection is to introduce you to a website that can be used to obtain areas under any Normal curve.

Go the website:

`http://davidmlane.com/hyperstat/z_table.html`

Click on this site, please, and I will walk you through its use.

As you visually scan down the page, the first thing you will see is a graph of the N(0,1) curve. (I know this is the standard Normal curve—N(0,1)—because a bit farther down the page I see the entries 0 for *Mean* and 1 for *SD*.) For now, ignore the shaded region under the graph.

Immediately below the graph are two circles that let you choose the *type of problem* you wish to solve. The options are:

- Area from a value; and
- Value from an area.

The site's default is the first of these options. *Do not change the default.* (We will return to the second option in Chapter 9.)

Next, the site allows you to specify the Normal curve *of interest to you*. The default values are Mean (μ) equal to 0 and SD (σ) equal to 1. For now, let's leave the default values unchanged.

Immediately below *SD* is a vertical display of four circles, labeled: *Above, Below, Between* and *Outside*. To the right of each circle is a rectangle—or two—in which you may enter a number—or numbers. The default selects the circle *Above* with the numerical entry 1.96.

The website author chose the word *Above*, whereas I would prefer either *Greater than* or *To the right of*. Similarly, instead of *Below* I would prefer *Less than* or *To the left of*. Just so you don't think I am a Mr. Negative, I can't think of any way to improve upon either of the labels *Between* or *Outside*.

Now, move your eyes up a bit to the picture of the Standard Normal curve. You will note that the area *Above* 1.96 (when in Rome ...) is shaded black. Finally, below the display of circles and rectangles you will find the display: *Results: Area (probability) 0.025*. The norm in Statistics is to report this area rounded to four digits; for the current problem the area is 0.0250, and the site drops the trailing 0. (I would include it, but it's not my site!)

Let's do four quick examples of the use of this site.

1. In the rectangle next to *Above*, replace the value 1.96 by 2.34 and click on the *Recalculate* box, which is located immediately below the word *Area*. I did this and obtained 0.0096 as the answer from the site. Note that the site is behaving as I said; the answer is reported to four digits after the decimal point. Also, note that the area to the right of 2.34 is shaded in the site's graph.

2. I enter 3.89 in the rectangle next to *Above* and click on *Recalculate*; the answer is 0.0001.

3. I enter 3.90 in the rectangle next to *Above* and click on *Recalculate*; the answer is 0. I don't like this answer. This is not zero as in, "My Detroit Lions have won zero Super Bowls." This 0 means that the area, rounded to the fourth digit after the decimal point (nearest ten-thousandth, if you prefer), is 0.0000. If this were *my site*, it would definitely report the answer as 0.0000.

4. Finally, click on the circle next to *Below*; place -0.53 in its box and click on *Recalculate*. You should obtain the answer 0.2981.

Next, I will use this site to obtain the *Normal Approximation* entries in Table 7.5. As explained earlier, I want $\mu = 27.5$ and $\sigma = 4.787$; thus, I enter these numbers in *Mean* and *SD*, respectively.

Let's begin with $r_1 = 29$. I remember the continuity correction, which tells me to replace 29 by 28.5. I select the circle for *Above* and type 28.5 into its rectangle. I click on *Recalculate* and obtain the answer 0.4173, as reported in Table 7.5.

Here is one more example. For $r_1 = 33$ I entered 32.5 in the rectangle next to *Above*. The site responds with 0.1481, as reported in Table 7.5.

7.4.2 Using a Website to Perform the Sum of Ranks Test

In Chapter 6 you learned how to obtain the exact P-value for the sum of ranks test provided I give you the exact sampling distribution for R_1. In the absence of the exact sampling distribution, you learned how to use the *vassarstats* website to obtain an approximate P-value via a computer simulation. The computer simulation gives us two P-values: for the two-sided alternative ($\neq$) and for the one-sided alternative that is supported by the data: $>$ [$<$] if $r_1 > \mu$ [$r_1 < \mu$], with μ given by Equation 7.2. Also, if $r_1 = \mu$, then the website should not be used because: we know that the exact P-value for the two-sided alternative must be one and the website is wrong for the one-sided alternative.

The *vassarstats* site also gives two P-values based on the Normal curve approximation of this chapter. First, go to

`http://vassarstats.net`

About one-half way down the list of topics in the left margin, click on *Ordinal Data*. This action takes you to a page that offers you several options; click on the second option, the *Mann-Whitney Test*. (Recall that Mann Whitney (Wilcoxin) is the official name of our sum of ranks test.)

The site asks you to enter your values of n_1 and n_2, which the site call n_a and n_b, respectively. **Warning:** The site says that if $n_a \neq n_b$, then you must have $n_a > n_b$; thus, you might need to relabel your treatments in order to use the site. Remember: If you relabel your treatments, you need to reverse any one-sided alternative.

After entering your sample sizes, *vassarstats* is ready to receive your data. A nice feature of the *Mann-Whitney Test* site is that you can cut-and-paste your data. If you import your data—i.e., you cut-and-paste—then you must click on *Import data to data cells* or the site won't work. The effect of this clicking is that the site will move your data to the *Raw Data for* portion of the page.

Next, you click on *Calculate from Raw Data*. The site gives us some useful information and a lot of information that you should ignore. We are given the mean ranks, from which we could obtain the value r_1 for our data. The site provides something called U_A which is a function of r_1 and I recommend that you ignore it. The site gives two proportions, labeled $P_{(1)}$ and $P_{(2)}$. The first of these is the approximate P-value for the one-sided alternative that is supported by the data. The second of these is the approximate P-value for the two-sided alternative. Both approximations are based on using the Normal curve approximation of this chapter. As best I can tell, the site uses the continuity correction with an adjustment of 0.5.

Warning: Whether or not there are ties in the data set, this *Mann-Whitney Test* site calculates the variance of R_1 using the *no ties* formula, Equation 7.4. For numerical data with a few ties (I

know, this is vague), this variance oversight by *vassarstats* does **not** seriously affect the quality of the approximation. (More on this topic in the Practice Problems.) Note, however, that this site **should not be used** for ordinal categorical data. As we have seen, with ordinal categorical data the value of the variance is changed substantially by the presence of (many) ties.

7.4.3 Reading Minitab Output; Comparing Approximation Methods

This is a new topic for the *Computing* sections of these *Course Notes*. I have previously told you that I use Minitab to perform what I call **full-blown** simulation experiments. In those situations I took the Minitab output—which isn't pretty—and presented it to you in a more understandable format. This subsection, however, is different.

Several times in these notes I will show you the output of a Minitab statistical analysis without showing you how to create such output. In this subsection I will show you the output that Minitab creates for our sum of ranks test. Finally, this subsection ends with a comparison of methods for Dawn's study of her cat, Sara's study of golf and the *data from an artificial study of a serious disease* presented in Table 6.8 on page 129.

Example 7.3 (Dawn's study of her cat Bob.) Recall that Dawn's sorted data are presented in Table 1.3 on page 7. I will find the Normal curve approximate P-value for the sum of ranks test and the alternative $>$ for Dawn's data. I will do this three ways: using Minitab; using the *vassarstats* website; and by hand. These three methods better give the same answer!

I entered these data into Minitab and ran the *Mann* command on it. The edited output is below. (I eliminated output that we don't use.)

```
Mann-Whitney Test: C1, C2

C1          N =  10     Median =       5.500
C2          N =  10     Median =       3.000
W = 133.5
Test of  =   vs   >  is significant at 0.0171
The test is significant at 0.0163 (adjusted for ties)
```

Let me walk you through this output, although it is fairly self-explanatory.

- Minitab reminds me that I have placed my data into Minitab columns C1 and C2.
- Minitab reminds me that the number of trials on both treatments is 10; and that the median of the chicken [tuna] data is 5.5 [3.0].
- Minitab tells me that $r_1 = 133.5$, but calls it W.
- I told Minitab that I wanted the alternative $>$. Minitab gives me two Normal curve approximation P-values: the *better* one—adjusting for ties—0.0163; and the *inferior* one—no adjustment—0.0171. Even though it seems to me that there are a lot of ties in the data—the values of t_i's are three 1's, two 2's, three 3's and a 4—the presence of ties has very little influence on the P-value.

I ran these data through the *vassarstats* website and it gave me 0.017 for the approximate P-value based on the Normal curve; i.e., the same answer—to three digits—as the Minitab answer that does not adjust for ties.

Finally, by hand, I obtained $\mu = 105$ and $\sigma = 13.1089$, adjusted for ties. As noted above, Dawn's data give $r_1 = 133.5$. I went to

`http://davidmlane.com/hyperstat/z_table.html`

to find my approximate P-value. I entered the above values of μ and σ and asked for the area above $133.5 - 0.5 = 133$ (remember the continuity correction). The site gave me 0.0163, which agrees with Minitabs *adjusted* answer.

In Example 6.3 on page 126, I use a simulation experiment with 10,000 reps to obtain an approximate P-value for R_1 and the alternative $>$. I obtained the value 0.0127 which gives a nearly certain interval of $[0.0093, 0.0161]$. (Details not shown.) The Normal approximation—adjusted for ties—P-value, 0.0163, is very close to being in this interval. Thus, I conclude that the Normal approximation—adjusted for ties—is a pretty good approximation for Dawn's data, although it appears to be a bit too large.

Example 7.4 (Sara's study of golf.) Recall that Sara's sorted data are presented in Table 2.2 on page 29. I entered Sara's data into Minitab and ran the *Mann* command on it with the alternative $>$. The edited output is below.

```
Mann-Whitney Test: C1, C2

C1          N =  40     Median =      112.00
C2          N =  40     Median =       99.50
W = 1816.0
Test of = vs  > is significant at 0.0300
The test is significant at 0.0299 (adjusted for ties)
```

The sample sizes and treatment medians agree with our earlier work. We see that adjusting for ties has only an unimportant impact on our approximate P-value. Thus, the *vassarstats* site would have been fine for analyzing these data.

If you refer to Example 6.4 in Chapter 7, you will recall that our approximate P-value for $>$ based on a 10,000 rep simulation experiment is 0.0293, with nearly certain interval:

$$0.0293 \pm 3\sqrt{\frac{0.0293(0.9707)}{10{,}000}} = 0.0293 \pm 0.0051 = [0.0242, 0.0344].$$

I know that I can trust computer simulations. Thus, I conclude that the Normal curve approximation—0.0293—appears to be excellent for this data set.

Example 7.5 (An artificial study of a serious disease.) Example 6.5 on page 128 introduced an artificial study with an ordinal categorical response. Its data are in Table 6.8 on page 129.

I entered these data into Minitab and ran the *Mann* command on it with the alternative $<$. The edited output is below.

```
Mann-Whitney Test: C4, C5

C4          N =  50     Median =      2.0000
C5          N =  50     Median =      2.0000
W = 2340.0
Test of  =  vs  <  is significant at 0.1017
The test is significant at 0.0869 (adjusted for ties)
```

As discussed earlier in these notes, computing medians for ordinal categorical data often is nearly worthless, as it is for these data. We see that it is important to use the correct—adjusted for ties—formula for the variance because the two P-values are quite different.

Finally, Example 6.5 reported the results of a simulation experiment with 10,000 reps. It gave 0.0946 as the approximate P-value for the alternative $<$. Its nearly certain interval is

$$0.0946 \pm 3\sqrt{\frac{0.0946(0.9054)}{10{,}000}} = 0.0946 \pm 0.0088 = [0.0858, 0.1034].$$

The Normal approximation P-value, 0.0869, lies in this interval, but just barely.

In summary, the Normal curve approximation gives reasonably accurate P-values for each of these three studies. If one has access to Minitab, the analysis is quite easy to obtain. For all but ordinal categorical data, the *vassarstats* website usually will give an accurate approximate P-value.

7.5 Summary

A probability histogram is our picture of a sampling distribution. So far in these notes, we have had sampling distributions for our two test statistics: U and R_1. For more generality, we talk about a random variable X with observed value denoted by x. The random variable X has a sampling distribution and we obtain its probability histogram by following the three steps given on page 143.

We next turn to the question of finding an approximation of a probability histogram (which also provides an approximation of the sampling distribution represented by the probability histogram). Earlier in these *Course Notes* we learned how to obtain an approximation of a sampling distribution by using a computer simulation experiment. In the current chapter, we learn our first example of what I call a fancy math approximation. To that end, we need to summarize a probability histogram (equivalently, a sampling distribution) by determining its center (mean) and spread (standard deviation).

There is a simple formula for the mean of the sampling distribution of R_1 (Equation 7.2):

$$\mu = n_1(n+1)/2,$$

and two formulas for its variance. The first formula (Equation 7.4) is appropriate if the n observations in the combined data set are all distinct numbers (called the no-ties situation):

$$\sigma^2 = \frac{n_1 n_2 (n+1)}{12}.$$

The second formula (Equation 7.5) should be used whenever there are ties in the combined data set:

$$\sigma^2 = \frac{n_1 n_2 (n+1)}{12} - \frac{n_1 n_2 \sum(t_i^3 - t_i)}{12n(n-1)}.$$

If you don't recall the meaning of the t_i's in this formula, review the material in the text of this chapter immediately following Equation 7.5 on page 148.

Next, we learned about the family of Normal curves, pictured in Figure 7.4. A particular Normal curve is characterized by its center of gravity (mean) μ and its spread (standard deviation) σ; it is sometimes denoted as the N(μ,σ) curve. The Standard Normal curve corresponds to $\mu = 0$ and $\sigma = 1$.

There is a website that calculates areas under Normal curves and its use is demonstrated.

The main idea of this chapter is to use a Normal curve as an approximation of the probability histogram of R_1. We use the Normal curve that matches R_1 on its values of μ and σ (whichever version—ties or no ties—of σ is appropriate).

We always use the continuity correction when we obtain the Normal approximation to R_1. If we want to approximate $P(R_1 \geq r_1)$, we calculate the area under the Normal curve to the right of $(r_1 - 0.5)$. If we want to approximate $P(R_1 \leq r_1)$, we calculate the area under the Normal curve to the left of $(r_1 + 0.5)$.

Section 7.4 shows that the *vassarstats* website for the Mann-Whitney test will calculate the Normal approximation P-value for us. The site use the no-ties formula for the variance, which works reasonably well for a small number of ties. For a large number of ties—in particular, for an ordinal categorical response—the Mann-Whitney test site yields P-values that are a bit too large.

Also, in Section 7.4, we learned how to read Minitab output for the sum of ranks test.

7.6 Practice Problems

1. I have a balanced CRD with a total of six units. My six ranks are: 1, 2, 4, 4, 4 and 6.

   (a) Find six response values that would yield these ranks. Note that because ranks are obtained for the combined data set, you don't need to match observations to treatments to answer this question. Also note that there are an infinite number of correct answers.

   (b) Determine the exact sampling distribution of R_1 for these six ranks.

   (c) Draw the probability histogram of the sampling distribution of R_1 for these six ranks.

   (d) Show that the $\mu = 10.5$ with two different arguments.

   (e) Compute σ^2 and σ.

2. Refer to *Practice Problem* 3 on page 134 in Chapter 6; it presents a balanced CRD with $n = 50$ units and a four-category ordinal response.

   Using the same assignment of numbers to categories I used in Chapter 6, I entered my data into Minitab and directed Minitab to perform the sum of ranks test for the alternative $<$. I executed the Mann command and obtained the output below.

```
Mann-Whitney Test: C1, C2

C1          N =  25     Median =       2.000
C2          N =  25     Median =       3.000
W = 548.5
Test of  =  vs <  is significant at 0.0430
The test is significant at 0.0380 (adjusted for ties)
```

   In the Chapter 6 *Practice Problems* I reported that a 10,000 rep computer simulation experiment gave 0.0354 as an approximate P-value for the alternative $<$. Use this relative frequency to obtain the nearly certain interval (Formula 4.1 in Chapter 4) for the exact P-value. Comment on the Minitab approximate P-values.

3. Imagine that we have a CRD with $n_1 = 15$ units assigned to the first treatment and $n_2 = 20$ units assigned to the second treatment. Also, assume that there are no ties in the combined list of 35 response values.

   (a) Calculate the mean, variance and standard deviation of the sampling distribution of R_1.

   (b) Assume that the alternative is $>$. Use the Normal curve website to find the approximate P-values for the following actual values of r_1: 300, 330 and 370. Remember to use the continuity correction.

   (c) Assume that the alternative is $\neq$. Use the Normal curve website to find the approximate P-values for the following actual values of r_1: 300, 330 and 370. Remember to use the continuity correction.

(d) Assume that the alternative is $<$. Use the Normal curve website to find the approximate P-values for the following actual values of r_1: 250, 220 and 180. Remember to use the continuity correction.

7.7 Solutions to Practice Problems

1. (a) There is an infinite number of possible response values. I will proceed as follows. The three ranks equal to 4 mean that there are three tied values; call them 15, 15 and 15. There must be one observation larger than 15, which will have the rank 6, and two non-tied observations smaller than 15 to receive the ranks 1 and 2. I select observation values of 11, 14 and 17. Thus, my combined data set consists of 11, 14, 15, 15, 15 and 17.

   (b) There is one assignment that puts all three 4's on treatment 1, giving $r_1 = 12$.

   There are three assignments that put two 4's with one non-4 on treatment 1. The two 4's can be matched with 1, 2 or 6. Thus, this gives us nine more values of r_1: three each of 9, 10 and 14.

   There are three assignments that put one 4 with two non-4's on treatment 1. The one 4 can be matched with: 1 and 2; 1 and 6; or 2 and 6. Thus, this gives us nine more values of r_1: three each of 7, 11 and 12.

   Finally, there is one assignment that places all 4's on treatment 2. This assignment gives $r_1 = 1 + 2 + 6 = 9$.

   Combining the above, we see that three assignments each give the following values for r_1: 7, 10, 11 and 14. Also, four assignments each give the following values for r_1: 9 and 12. Thus,

   - $P(R_1 = r_1) = 0.15$ for $r_1 = 7, 10, 11, 14$.
   - $P(R_1 = r_1) = 0.20$ for $r_1 = 9, 12$.

   (c) The probability histogram is in Figure 7.6; I will explain its derivation.

   The possible values of r_1 are 7, 9, 10, 11, 12 and 14. They are marked and labeled in my picture. The minimum distance between consecutive values is 1. Because $\delta = 1$, the height of each rectangle equals the probability of its center's location; these probabilities are given in the answer above and presented in the probability histogram.

   (d) First, the probability histogram in Figure 7.6 is symmetric. Thus, its mean equals its point of symmetry, 10.5.

   Second, we can use Equation 7.2 on page 147:

$$\mu = n_1(n+1)/2 = 3(6+1)/2 = 21/2 = 10.5.$$

   (e) Because there are ties in the data set, we use Equation 7.5

$$\sigma^2 = \frac{n_1 n_2 (n+1)}{12} - \frac{n_1 n_2 \sum (t_i^3 - t_i)}{12n(n-1)}.$$

I will evaluate these terms individually and then subtract. The first term is:

$$\frac{n_1 n_2(n+1)}{12} = \frac{3(3)(7)}{12} = 63/12 = 5.25.$$

For the second term, first note that only one t_i is larger than 1; it is 3. The value of the second term is:

$$\frac{n_1 n_2 \sum(t_i^3 - t_i)}{12n(n-1)} = \frac{3(3)(3^3-3)}{12(6)(5)} = 216/360 = 0.60.$$

Thus,

$$\sigma^2 = 5.25 - 0.60 = 4.65 \text{ and } \sigma = \sqrt{4.65} = 2.1564.$$

2. For $\hat{r} = 0.0354$ and 10,000 reps, the nearly certain interval is:

$$0.0354 \pm 3\sqrt{\frac{0.0354(0.9646)}{10000}} = 0.0354 \pm 0.0055 = [0.0299, 0.0409].$$

Minitab gives us two approximate P-values. Whenever our data set has ties, we should use the adjusted (smaller) P-value. For some data sets, the two P-values are nearly identical, but not in this case. Thus, our Normal approximation is 0.0380. This value falls within our nearly certain interval; thus, the Normal approximation seems to be reasonable.

3. (a) From Equation 7.2,

$$\mu = n_1(n+1)/2 = 15(36)/2 = 270.$$

From Equation 7.4,

$$\sigma^2 = \frac{n_1 n_2(n+1)}{12} = \frac{15(20)(36)}{12} = 900 \text{ and } \sigma = \sqrt{900} = 30.$$

(b) We go to the website

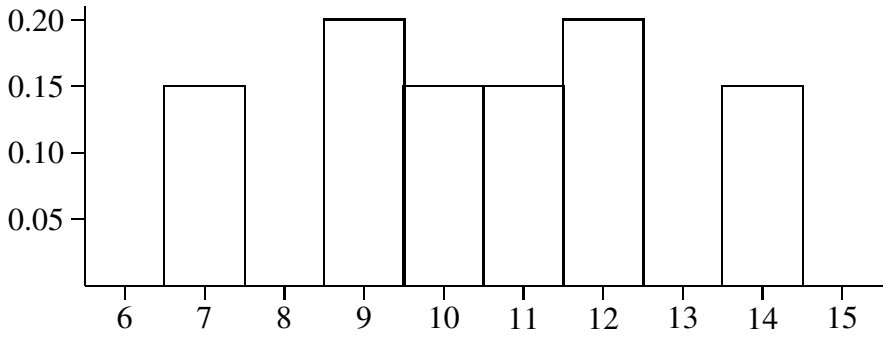
`http://davidmlane.com/hyperstat/z_table.html`

Following the instructions given in Section 7.4, we enter 270 in the *Mean* box and 30 in the *SD* box. We select the option *Above* and enter 299.5 in its box. We obtain the area 0.1627; this is the approximate P-value for $r_1 = 300$.

For the other values of r_1 we repeat the above, first typing 329.5 in the box and then 369.5. For 329.5, we get the area 0.02367, which is the approximate P-value for $r_1 = 330$. Finally, for 369.5, we get the area 0.0005, which is the approximate P-value for $r_1 = 370$.

(c) Here is a useful trick. The approximating curve (Normal) is symmetric. As a result, provided the one-sided approximate P-value does not exceed 0.5000, if you simply double it you obtain the approximate P-value for $\neq$. Hence, our approximate P-values are:

Figure 7.6: The probability histogram for the sampling distribution in Practice Problem 1.



- For $r_1 = 300$: $2(0.1627) = 0.3254$.
- For $r_1 = 330$: $2(0.0237) = 0.0474$.
- For $r_1 = 370$: $2(0.0005) = 0.0010$.

(d) We select the option *Below* and enter 250.5 in its box. We obtain the area 0.2578; this is the approximate P-value for $r_1 = 250$.

For the other values of r_1 we repeat the above, first typing 220.5 in the box and then 180.5. For 220.5, we get the area 0.0495, which is the approximate P-value $r_1 = 220$. Finally, for 180.5, we get the area 0.0014, which is the approximate P-value for $r_1 = 180$.

7.8 Homework Problems

1. An unbalanced CRD has $n_1 = 2$, $n_2 = 3$ and $n = 5$. Thus, there are 10 possible assignments. (You don't need to verify this fact.) The sorted numbers in the combined data set are:

$$3, 8, 8, 8 \text{ and } 20.$$

   (a) Determine the ranks for the combined data set.

   (b) Complete the following table, the sampling distribution for R_1.

r_1:	4	6	8
$P(R_1 = r_1)$:			

   (c) Draw the probability histogram for your table in (b).

   (d) Calculate the variance and standard deviation of the sampling distribution for R_1.

2. Refer to Homework Problem 3 on page 139, an unbalanced CRD with an ordinal response. Its data are reproduced below.

	Response			
Treatment	Disagree	Neutral	Agree	Total
1	8	7	5	20
2	3	5	7	15
Total	11	12	12	35

   I assign numbers 1 (Disagree), 2 (Neutral) and 3 (Agree) to these categories. I placed these data in Minitab and executed the Mann command. My output is below.

```
Mann-Whitney Test: C1, C2

C1          N =  20     Median =       2.000
C2          N =  15     Median =       2.000
W = 318.0
Test of =  vs  not =  is significant at 0.1666
The test is significant at 0.1424 (adjusted for ties)
```

   Use this output to answer the questions below.

   (a) What is the observed value of R_1?

   (b) Which alternative ($>$, $<$ or $\neq$) did Minitab use?

   (c) Which P-value is better to use? Why?

3. Recall Reggie's study of darts, first introduced in Homework Problems 5–7 in Chapter 1 (Section 1.8).

(a) Use the *vassarstats* website's Mann-Whitney command to obtain the Normal approximation to the P-value for $>$. (Using Minitab, I found that ignoring ties, as *vassarstats* website's Mann-Whitney does, has no impact on the approximation. Well, it changes the answer by 0.0001, which I am willing to ignore.)

(b) I performed a simulation with 10,000 reps and obtained 0.0075 for my approximate P-value for the alternative $>$. Compute the nearly certain interval (Formula 4.1) for the exact P-value for the alternative $>$.

(c) Compare your approximations from (a) and (b); comment.

4. Consider an unbalanced CRD with $n_1 = 36$ and $n_2 = 27$. Assume that the combined data set contains no ties.

(a) Calculate the mean and standard deviation of the sampling distribution of R_1.

(b) Use your answers from (a) and the website

`http://davidmlane.com/hyperstat/z_table.html`

to obtain the approximate P-values for the alternative $>$ and each of the following values of r_1: 1200, 1250 and 1300.

(c) Use your answers from (a) and the website

`http://davidmlane.com/hyperstat/z_table.html`

to obtain the approximate P-values for the alternative $<$ and each of the following values of r_1: 1120 and 1070.

(d) Use your answers from (a) and the website

`http://davidmlane.com/hyperstat/z_table.html`

to obtain the approximate P-values for the alternative $\neq$ and each of the following values of r_1: 1300 and 1070.

Chapter 8

Dichotomous Responses; Critical Regions

8.1 Introduction and Notation

In all previous studies in these notes, the response has been either a numerical variable or an ordered categorical variable with at least three categories. For a numerical response we compared the treatments by comparing the means or the ranks of their responses. For an ordered categorical response we compared the treatments by comparing the ranks of their responses.

In this chapter we consider studies that have a dichotomous response—a categorical response with two categories. We begin with four examples.

Example 8.1 (Therese's infidelity study.) Therese studied 20 of her adult female friends. The women were divided into two treatment groups, both of size 10, by randomization. Women assigned to the first treatment group read the following question:

- You are friends with a married couple and are equally fond of the man and the woman. You discover that the *husband* is having an affair. The *wife* suspects that something is going on and asks you if you know anything about *her husband* having an affair. Do you tell?

Women assigned to the second treatment group read the following question:

- You are friends with a married couple and are equally fond of the man and the woman. You discover that the *wife* is having an affair. The *husband* suspects that something is going on and asks you if you know anything about *his wife* having an affair. Do you tell?

Each subject was instructed to respond either yes or no.

Example 8.2 (Ruth's prisoner study.) Ruth's subjects were 50 male inmates at a minimum security federal prison camp in Wisconsin. All of the men were first-time nonviolent criminal offenders serving two or more years of prison time. The men were divided into two treatment groups of 25 each by randomization. Men assigned to the first treatment group were given the following question:

- The prison is beginning a program in which inmates have the opportunity to volunteer for community service with developmentally disabled adults. *Inmates who volunteer will receive a sentence reduction*. Would you participate?

Men assigned to the second treatment group were given the following question:

- The prison is beginning a program in which inmates have the opportunity to volunteer for community service with developmentally disabled adults. Would you participate?

Each subject responded either yes or no.

Example 8.3 (Thomas's golf putting study.) Thomas wanted to investigate the difference in difficulty between four and eight foot putts in golf. He performed a balanced study with randomization and a total of 50 putts. The first treatment was putting from four feet on a level surface and the second treatment was putting from eight feet on a level surface. Each putt was either made or missed.

Example 8.4 (The artificial Headache Study-2 (HS-2).) A researcher has 100 persons available for study. Each person routinely suffers mild tension headaches (not migraines). The researcher wants to compare two active drugs, call them A and B, for the treatment of mild headaches. The 100 subjects are divided into two groups of size 50 each by randomization. Each subject is given the following instructions:

> The next time you have a mild headache take the drug we have given you. Fifteen minutes later answer the following question with a response of either yes or no: Has your headache pain diminished?

When the response is a dichotomy, there are technical names for the two possible responses: one is called a **success** and the other is called a **failure**. The methods we learn will focus on *counting successes*. We use the following method for deciding which possible outcome gets the distinction of being called a success.

1. If one of the possible responses is very rare (admittedly vague), then it is labeled the success.
2. If neither possible response is very rare, then the more desirable response is labeled the success.
3. If neither of the previous two scenarios applies, then the researcher arbitrarily assigns the label success to one of the possible responses.

Here is the idea behind the first rule. Every time I drive a car I have the potential to be involved in a traffic accident. Fortunately, in my 47 years of driving I have been in only one accident and it was very minor. When something occurs only rarely, it is much easier to keep track of how many times it happens rather than how many times it fails to happen.

In our examples above, the researchers labeled as successes: telling, agreeing to volunteer, making a putt and reporting that the pain has diminished. Tables 8.1–8.4 present and summarize the data for each of our four studies.

Table 8.1: The 2×2 contingency table of observed counts for Therese's infidelity study.

	Tell?			Row Proportions	
Cheater was:	Yes	No	Total	Yes	No
The Husband	7	3	10	0.70	0.30
The Wife	4	6	10	0.40	0.60
Total	11	9	20		

Table 8.2: The 2×2 contingency table of observed counts for Ruth's prisoner study.

	Volunteer?			Row Prop.	
Version Read:	Yes	No	Total	Yes	No
Sentence Reduction	18	7	25	0.72	0.28
No Sentence Reduction	23	2	25	0.92	0.08
Total	41	9	50		

Table 8.3: 2×2 Contingency table of observed counts for Thomas's golf putting study.

	Putt was			Row Prop.	
Distance:	Made	Missed	Total	Made	Missed
Four feet	18	7	25	0.72	0.28
Eight feet	10	15	25	0.40	0.60
Total	28	22	50		

Table 8.4: 2×2 contingency table of observed counts for the artificial Headache Study-2 (HS-2).

	Pain relieved?			Row Prop.	
Drug :	Yes	No	Total	Yes	No
A	29	21	50	0.58	0.42
B	21	29	50	0.42	0.58
Total	50	50	100		

Table 8.5: General notation for a 2×2 contingency table of observed counts for a CRD with a dichotomous response.

	Response			Row Proportions	
Treatment :	S	F	Total	S	F
1	a	b	n_1	$\hat{p}_1 = a/n_1$	$\hat{q}_1 = b/n_1$
2	c	d	n_2	$\hat{p}_2 = c/n_2$	$\hat{q}_2 = d/n_2$
Total	m_1	m_2	n		

Table 8.5 presents our general notation for a CRD with a dichotomous response. When I develop ideas below it will be convenient to use the general notation.

First, a few comments:

1. The orientation for these tables in these notes will follow the four examples above; namely, the rows distinguish between treatments and the columns identify the possible responses. Many materials (texts, research papers, etc.) reverse this orientation. Thus, when reading other materials, be careful to identify which orientation is being used.

2. We summarize the table of counts by computing the **row proportions**: the $\hat{p}$'s and $\hat{q}$'s given above. There is a great deal of redundancy in these; namely, in each row the sum of its $\hat{p}$ and $\hat{q}$ is always one. Thus, after you get more familiar with these ideas I usually will suppress the $\hat{q}$'s.

3. In these tables, I do **not** calculate the row proportions for the *Total* row because in a CRD these numbers typically are not of interest.

There is a very simple, but useful, connection between the treatment (row) proportions of successes and the means ($\bar{x}$ and $\bar{y}$) of our earlier work for a CRD with a numerical response. I will illustrate the connection with Therese's data; the interested reader can show easily that the connection is also true for the general case.

Therese assigned 10 friends to her first treatment, the husband having an affair; seven responded *yes* (success) and three responded *no* (failure) giving $\hat{p}_1 = 7/10 = 0.70$. Alternatively, we can make Therese's response a number: 1 for *yes* and 0 for *no*. With this latter viewpoint, Therese's data consist of seven 1's and three 0's. Clearly, the sum of her 10 numbers is 7–which is the total number of successes. The mean of these 10 numbers, which we call $\bar{x}$, is $7/10 = 0.70$. In other words,

$$\hat{p}_1 = \bar{x} \text{ and, similarly, } \hat{p}_2 = \bar{y}.$$

This identification is important because:

> It shows that the Skeptic's Argument and all that follows from it—the Advocate, the hypotheses, the test of hypotheses, the rules for computing the P-value and so on—can be immediately adapted to a dichotomous response.

For convenience, we will call the observed value of our test statistic $x = \hat{p}_1 - \hat{p}_2$ which implies that we refer to our test statistic as X. We could, of course, call it U, but I fear that would create confusion. Let's keep U for comparing means and X for comparing proportions, even though they coincide. By the way, X is also mathematically equivalent to R_1, but you don't need to understand the details of the argument. You are welcome to ignore this latter connection or rejoice in it.

8.2 The Test of Hypotheses: Fisher's Test

Let me be precise about the test of hypotheses in this chapter. Define p_1 to be the proportion of successes that would be obtained if all units were assigned to treatment 1. Similarly, p_2 is the proportion of successes that would be obtained if all units were assigned to treatment 2. If the Skeptic is correct, then $p_1 = p_2$ because the treatment does not matter; i.e., some units will yield successes and others will yield failures, regardless of the treatment they receive. For example, if the Skeptic is correct for Therese's study of infidelity, then some of her friends are *tellers*, no matter the sex of the person having the affair, and the remaining friends *keep quiet*, again no matter the sex of the person having the affair.

If we could perform the clone-enhanced study we would know whether the Skeptic is correct. If the Skeptic is incorrect, the clone-enhanced study would reveal this fact as well as the values of p_1 and p_2.

For testing, the null hypothesis is that the Skeptic is correct. As with our test based on comparing means, there are three options for the alternative:

- $H_1 : p_1 > p_2$;
- $H_1 : p_1 < p_2$; and
- $H_1 : p_1 \neq p_2$.

As before, you may choose whichever alternative you like, provided you make your choice before collecting data. Also as before, I recommend using the Inconceivable Paradigm to make this choice.

In addition, when the response is a dichotomy there is a big bonus: On the assumption the Skeptic is correct, we don't need to perform a computer simulation experiment to approximate the P-value; there is a *simple mathematical formula* that allows us to calculate the exact P-value.

When I say a *simple mathematical formula* in the previous sentence, I am being a bit tricky. It is a simple formula to write down, but tedious to compute by hand. Fortunately, there is a website that, with minimal effort from us, will produce the exact P-value for each of the three possible alternative hypotheses. The website is

`http://www.langsrud.com/fisher.htm`

Note that I am deviating from my usual method of presenting material. In the previous chapters, the computational methods had their own section, named *Computing*. I believe that this chapter

will be easier for you to follow if I incorporate the website now. (As always, let me know if you think I am making a tactical or strategic error.)

The first thing I notice when I look at this website is the label **Fisher's Exact Test** in large, bold-faced type in its upper left corner. This test is named in honor of Sir Ronald A. Fisher (1890–1962), the most famous statistician of the first half of the twentieth century. Fisher also did well-respected work in genetics. Among Fisher's many important works, he was a strong advocate of the importance of randomization in scientific studies, which is part of the reason this test bears his name.

Immediately below the name of the test, you will find:

1. A box labeled 'COMPUTE.'

2. A rectangle containing four small windows. The windows contain the default data: 3, 1, 1 and 3.

3. A box labeled 'CLEAR TABLE.'

4. A box labeled 'CLEAR OUTPUT.'

I will now illustrate the use of this website for Ruth's data (see Table 8.2). Just follow the steps below.

1. Click on the box 'CLEAR TABLE.' This action will result in the default data disappearing from the four small windows.

2. Enter Ruth's counts, 18, 7, 23, and 2, in the four small windows in the same order that they appear in the contingency table.

3. Click on the box labeled 'COMPUTE.' Three P-values will appear in the output window on the right side of the screen. For Ruth's data, your display should read:

   - Left: p-value $= 0.0691666077613616$.
   - Right: p-value $= 0.9883922891274275$.
   - 2-Tail : p-value $= 0.1383332155227232$.

This website was **not** written by me; thus, unsurprisingly, it does not follow my notation. In particular, note the following:

- Left on the website gives the P-value for our alternative $<$.
- Right on the website gives the P-value for our alternative $>$.
- 2-Tail on the website gives the P-value for our alternative $\neq$.

Note that for Ruth's data, the P-value for $\neq$ is twice the smaller of the other two P-values. This is no surprise. We know from our earlier work that whenever a study is balanced—i.e., whenever $n_1 = n_2$—then the exact sampling distribution of U—and, hence, X—is symmetric about 0. If $X = 0$, however, this *doubling* does not work. The P-value for $\neq$ is 1 and the P-values for the other two alternatives are equal and both exceed 0.5000.

Before we discuss the meaning of these P-values, I want you to get some additional practice using this website. In particular,

1. Go to the website and enter the counts for Therese's infidelity study: 7, 3, 4 and 6. You should obtain the following P-values:

   - Left: p-value $= 0.9651107406525374$.
   - Right: p-value $= 0.18492498213860464$.
   - 2-Tail : p-value $= 0.3698499642772093$.

2. Go to the website and enter the counts for Thomas's golf putting study: 18, 7, 10 and 15. You should obtain the following P-values:

   - Left: p-value $= 0.9952024757062247$.
   - Right: p-value $= 0.02250227023961592$.
   - 2-Tail : p-value $= 0.04500454047923211$.

3. Go to the website and enter the counts for HS-2: 29, 21, 21 and 29. You should obtain the following P-values:

   - Left: p-value $= 0.964328798255893$.
   - Right: p-value $= 0.08060057614938207$.
   - 2-Tail : p-value $= 0.16120115229876414$.

Let's look at one of these P-values, say, the 'Right' P-value for Thomas's golf putting study: 0.02250227023961592. This is a pretty absurd answer. Correct and absurd. It is absurd because it is so precise. Do I really care that the 17th digit after the decimal point is a 2? Would my interpretation change if it were a 7? Of course not! As a result, in these notes I will report a Fisher's Exact Test P-value to four digits after the decimal, with the exception noted below. The exception can be illustrated by what I will call the **really big and phony golf putting study (RBPGP).**

I don't want to spend much time on the RBPGP study, because, as its name suggests, it is make-believe. Consider the actual putting study performed by Thomas, but now suppose that his counts are all multiplied by ten, becoming: 180, 70, 100 and 150. I entered these counts into the Fisher website and obtained the following P-values:

- Left: p-value $= 0.9999999999999133$.
- Right: p-value $= 3.4235675124234885$e-13.

- 2-Tail : p-value $= 6.847135024846977$e-13.

Look at the P-value for Right ($>$). Reading left-to-right we begin with a curious result: the P-value equals 3.42 ... which cannot be correct because it is larger than 1! This alerts us to keep reading. We are rewarded for our persistence when we find the e-13 at the end of the P-value. This is the website's way of writing scientific notation. The correct way to interpret this P-value is that it equals

$$3.424 \times 10^{-13}, \text{ or, if you prefer, } 0.0000000000003424,$$

a really small P-value!

To summarize the above, I *usually* will report my P-values to four digits after the decimal point *unless* the result is close to 0.0000, in which case I *usually* will use scientific notation with four significant digits. (As Oscar Wilde said, "Consistency is the last refuge of the unimaginative.")

Let us now look at Ruth's prisoner study again. Ruth's data, as you may have noted, display a remarkable pattern, which we will get to shortly. Let's go through Ruth's reasoning that led to her choice of alternative. The Skeptic's Argument for Ruth's study is that some men would volunteer to work with developmentally disabled adults and some would not, but that the treatment was irrelevant. Ruth decided that the only conceivable alternative is that the offer of a sentence reduction would **increase** the number of men who would volunteer. Thus, Ruth felt that the $<$ alternative was inconceivable and she opted for the alternative $>$. Now look again at Ruth's data: her $\hat{p}_2$ is larger than her $\hat{p}_1$ by 0.20, twenty percentage points. In words, her data support the alternative she labeled as inconceivable!

You might be wondering *why* I have included Ruth's study in these notes. Let me assure you, my motive is **not** to ridicule Ruth. In fact, I admire Ruth quite a lot. Not many of my students traveled to a prison to collect data! (Ruth's major was Social Work.) If I had been asked to select the alternative—remember, we do this before we collect data—I definitely would have made the same choice that Ruth did.

Thus, why have I included Ruth's somewhat embarrassing study? First, to illustrate that *inconceivable* is **not** the same as *impossible*. Yes, it might be embarrassing, but I conjecture that every scientist will occasionally obtain data that support the inconceivable.

When data support the inconceivable, it is natural to wonder: What went wrong? So, take a minute and think about Ruth's study. Can you think of any reason(s) why the mention of a sentence reduction would lead to **less** participation in the volunteer program? Below, I have listed some ideas of mine. Do any of these match your reason(s)? Do any of these seem particular clever or stupid?

1. Perhaps the experience of being a prisoner makes the men *not trust* prison officials—or a researcher—to the extent that the promise of a reward has a negative impact.

2. Prisons are very expensive to operate. Perhaps it is routine for *first-time nonviolent criminal offenders serving two or more years* to receive a sentence reduction for *good behavior*. In this case, the prisoner might feel like he is being tricked; being offered a reward that he would likely obtain anyways. Thus, perhaps the question should have read,

> Inmates who volunteer will receive a sentence reduction *in addition to any other sentence reductions*.

I am not going to *obsess* about possible explanations; my reason will become clear after we examine Ruth's three P-values.

As we have seen earlier, Ruth's exact P-value for her selected alternative $>$ (Right) is 0.9884. This incredibly large P-value—remember the maximum possible value is 1—reflects the fact that, for her chosen alternative, Ruth obtained almost the *weakest possible evidence*. But look at her other two P-values: 0.0692 for $<$ and 0.1383 for $\neq$. Under the usual classical approach to tests of hypotheses, *neither* of these P-values is small enough to reject the null hypothesis. Thus, the pattern in Ruth's data, while surprising, is within the usual bounds of chance behavior, computed, of course, under the assumption the null hypothesis is correct. You might argue that 0.0692 is very close to the magic threshold of 0.05; but I would respond that, in my opinion and experience, it would be very difficult to convince many people that it is reasonable—before collecting data—to argue that $>$ is *inconceivable* for Ruth's study.

For the golf putting study, Thomas chose the alternative $>$. (Why does this make sense, before collecting the data?) Thus, his P-value is 0.0225 and the classical approach says to reject the null in favor of the alternative. In words, his data convinced Thomas that he was better at making a four foot putt rather than an eight foot putt.

Now, you might be thinking, "Of course, a shorter putt is easier. What a waste of time!" I have two comments to make on this attitude:

1. Quite often, especially in hindsight, research confirms the obvious. For example, it might seem *obvious* that smoking tobacco is harmful to a person's health, but this conclusion was not reached easily, in part because it would have been unethical to use randomization.

2. Don't fall into the trap of thinking you *know* how the world works. Subtle issues can operate in a study. For example, consider a highly skilled basketball player shooting baskets in a practice setting. Let treatment 1 be shooting a free throw and let treatment 2 be shooting from 12 inches in front of the free throw line. Now, you could argue that treatment 2 *must* be easier than treatment 1 because it is a shorter shot. I am not sure about this. Treatment 2 is a shorter shot, but my guess is that the player has previously spent a huge amount of time practicing treatment 1 and virtually no time practicing treatment 2. Thus, the player *might* actually be better from the longer distance. I do admit that this basketball argument seems, to me, not to apply to golf putts.

I will end this chapter with a few comments about the HS-2. I see three possible scenarios for this study:

1. If drug A is an active drug and drug B is, in fact, a placebo, then the alternative $>$ seems obvious and the P-value is 0.0806.

2. If drug B is an active drug and drug A is, in fact, a placebo, then the alternative $<$ seems obvious and the P-value is 0.9643. This would be a strange and alarming result: drug B seems worthless for headache relief and, indeed, is borderline contraindicated.

3. If both drugs A and B are reasonable treatments for headaches (see the variety of products available at a pharmacy) then $\neq$ is likely the alternative of choice, giving a P-value of 0.1612, Unless, for example, A was an *extra strength* version of B, in which case I would opt for the alternative $>$.

8.3 The Critical Region of a Test

8.3.1 Motivation: Comparing P-values

Our focus on tests of hypotheses has been to find the P-value of a test. We can compare P-values *within* tests and *between* tests. Let me explain what I mean by these two forms of comparison.

Consider Dawn's study of her cat Bob with the alternative $>$. Recall that for the test of means, the observed value of the test statistic is $u = 2.2$. Table 4.1 on page 76 presents an approximation to the sampling distribution of U based on a simulation experiment with $m =$ 10,000 reps. This approximation yields 0.0198 as the approximate P-value for Dawn's data and the alternative $>$. Thus, the actual value of u, 2.2, provides fairly strong evidence in support of the alternative $>$.

I argued that a value of u that is *larger than* the actual 2.2 would yield even stronger evidence for the alternative $>$. In particular, you can verify from Table 4.1 that if the observed value of the test statistic had been $u = 2.6$, then the approximate P-value for the alternative $>$ would have been:

$$\text{Rel. Freq.}(U \geq 2.6) = 0.0023 + 0.0017 + 0.0008 + 0.0003 = 0.0051.$$

The above is what I mean by a *within test* comparison of P-values. The value of u which provides stronger evidence for $>$ (2.6 versus 2.2) yields the smaller P-value. This comparison is pretty noncontroversial: Within a study, stronger evidence yields a smaller P-value. By contrast, a *between test* comparison of P-values raises important issues.

Recall Sara's study of golf, with data presented in Table 2.2 on page 29. For the alternative $>$ and the test based on means, I stated in Table 4.3 on page 77 that the approximate P-value—based on a simulation experiment with 10,000 reps—is 0.0903. By contrast, in Example 6.4 on page 127 I stated that the approximate P-value for the alternative $>$ and the sum of ranks test is 0.0293. Thus, for Sara's data and the alternative $>$ the approximate P-value for the sum of ranks test is much smaller than the approximate P-value for the test that compares means; this is what I mean by a *between test* comparison of P-values.

I am tempted to say (and I have heard many scientists and statisticians in similar situations say) that for Sara's data the sum of ranks test is better than the test that compares means because its P-value is much smaller. There are two reasons I resist saying this:

1. It is not literally true, even though arguably it is suggestive.

2. It sounds like I am cheating. (Perhaps) I prefer to reject the null hypothesis in favor of the alternative; thus, I state that the test that does what I want is better than the test that doesn't!

The difficulty is that we have not, as yet, developed the tools needed to compare tests. We will do so in Chapter 9 when I introduce the notion of the power of a test. *Power* is a difficult topic and I have chosen to begin, in this chapter, the journey to your understanding power.

Table 8.6: A partial 2×2 contingency table of observed counts for a balanced CRD.

	Response			
Treatment	S	F	Total	$\hat{p}$
1	a	b	27	$a/27$
2	c	d	27	$c/27$
Total	31	23	54	

8.3.2 From P-values to Critical Regions

If a researcher has a CRD with a dichotomous response, then

`http://www.langsrud.com/fisher.htm`

is an incredibly useful site because it provides exact P-values with almost no effort from the researcher. As a teacher, however, the site is disappointing because it does not provide the entire sampling distribution of X. With some work, however, the site can be used to obtain the entire sampling distribution of X.

Example 8.5 (A balanced CRD on a total of 54 units that yields a total of 31 successes.) The table of data described in the name of this example is given in Table 8.6. Note that I am **not** telling you the actual cell counts, a, b, c and d. My goal is to examine the sampling distribution of X for this table; for this goal, as you recall, we don't need the actual data, just the marginal totals.

A natural question for you to have is:

> Bob, why did you choose a balanced study with 27 units on each treatment? Twenty-seven makes the arithmetic messy, to say the least.

My answer is twofold:

- We are going to avoid any arithmetic that involves dividing by 27; thus its messiness is irrelevant; and
- These data give me a number for a particular P-value that I like; I will reveal why soon.

From Table 8.6 we see that, given the cell counts:

$$\hat{p}_1 = a/27 \text{ and } \hat{p}_2 = c/27; \text{ thus, } x = \hat{p}_1 - \hat{p}_2 = (a-c)/27$$

is the observed value of the test statistic X. Also, $a + c = 31$ or $a - 31 = -c$. Thus,

$$x = (a-c)/27 = (a + a - 31)/27 = (2a - 31)/27.$$

My point is that the value of a in Table 8.6 determines the value of x; as it must, because the value of a, given the marginal totals, determines the values of b, c and d.

Let's go to the site

Table 8.7: The sampling distribution of X for the 2×2 contingency table presented in Table 8.6.

		P-values for:		
		$<$	$>$	$\neq$
a	x	$P(X \leq x)$	$P(X \geq x)$	$P(\|X\| \geq x)$
8	$-15/27$	0.0000	1.0000	0.0001
9	$-13/27$	0.0004	1.0000	0.0008
10	$-11/27$	0.0027	0.9996	0.0054
11	$-9/27$	0.0133	0.9973	0.0267
12	$-7/27$	0.0489	0.9867	0.0978
13	$-5/27$	0.1355	0.9511	0.2709
14	$-3/27$	0.2913	0.8645	0.5826
15	$-1/27$	0.5000	0.7087	1.0000
16	$1/27$	0.7087	0.5000	1.0000
17	$3/27$	0.8645	0.2913	0.5826
18	$5/27$	0.9511	0.1355	0.2709
19	$7/27$	0.9867	0.0489	0.0978
20	$9/27$	0.9973	0.0133	0.0267
21	$11/27$	0.9996	0.0027	0.0054
22	$13/27$	1.0000	0.0004	0.0008
23	$15/27$	1.0000	0.0000	0.0001

`http://www.langsrud.com/fisher.htm`

and investigate what happens when I plug-in the value $a = 21$, which implies that $b = 6$, $c = 10$ and $d = 17$—otherwise, the margins in Table 8.6 would be wrong. The observed value of the test statistic is

$$x = (a - c)/27 = (21 - 10)/27 = 11/27.$$

I obtain the following three P-values, after rounding to four digits after the decimal point:

- Left ($<$): 0.9996; i.e., $P(X \leq 11/27) = 0.9996$.
- Right ($>$): 0.0027; i.e., $P(X \geq 11/27) = 0.0027$.
- 2-Tail($\neq$): 0.0054; i.e., by symmetry:

$$P(|X| \geq 11/27) = 2P(X \geq 11/27) = 2(0.0027) = 0.0054.$$

I repeated the above computation for all possible values of a in Table 8.6; my results are presented in Table 8.7. There is a tremendous amount of information in this table; let me begin by making a few comments about it.

1. Table 8.7 does not actually present *all possible values of* a as I claimed above. Why not? Well, for $a < 8$, equivalently $x < -15/27$,

   $$P(X \leq x) = 0.0000, P(X \geq x) = 1.0000 \text{ and } P(|X| \geq x) = 0.0000.$$

   In words, for $x < -15/27$, the P-value is really small for both $<$ and $\neq$. **For my current purposes, we won't be concerned with *how small.***

   Similarly, for $a > 23$, equivalently $x > 15/27$,

   $$P(X \leq x) = 1.0000, P(X \geq x) = 0.0000 \text{ and } P(|X| \geq x) = 0.0000.$$

   In words, for $x > 15/27$, the P-value is really small for both $>$ and $\neq$.

2. For $x < 0$, the P-value for $\neq$ is twice the P-value for $<$, except for possible round-off error. Similarly, For $x > 0$, the P-value for $\neq$ is twice the P-value for $>$, except for possible round-off error. These relationships are no surprise; they follow from the fact that, because the study is balanced, the sampling distribution of X is symmetric around zero.

Recall the **classical approach** to interpreting a P-value, given in Chapter 5 on page 99 and reproduced below:

> Reject the null hypothesis in favor of the alternative if, and only if, the P-value is less than or equal to 0.05.

Let's consider the alternative $>$. Look at Table 8.7 again and note that the P-value is 0.05 or smaller, if, and only if, $x \geq 7/27$ ($a \geq 19$). As a result, if my primary interest is on whether or not I reject the null hypothesis, then I have the rule

> Reject the null hypothesis if, and only if, $X \geq 7/27$.

Being lazy, statisticians summarize the above by saying that the **critical region** for the test is

$$(X \geq 7/27).$$

(I will add parentheses, when deemed necessary, to set off the formula for a critical region from its neighboring text.) Thus, the null hypothesis is rejected if, and only if, the observed value of the test statistic falls in the critical region. The point being that it is easier to say *the critical region is* $(X \geq 7/27)$ *instead of the whole if, and only if, stuff.*

By similar reasoning, the critical region for the alternative $<$ is $(X \leq -7/27)$. Finally, the critical region for the alternative $\neq$ is $(|X| \geq 9/27)$.

Let's pause for a moment. You might be wondering, "Why are we learning about critical regions?" This is a fair question. Sadly, you won't see the answer until we have finished Chapter 9. Thus, please be patient. As Andre Gide said:

> One does not discover new lands without consenting to lose sight of the shore for a very long time.

Table 8.8: Types 1 and 2 errors in a test of hypotheses.

Action (by researcher)	Truth (Only Nature knows)	
	H_0 is correct	H_1 is correct
Fails to reject H_0	Correct action	Type 2 Error
Rejects H_0	Type 1 Error	Correct action

8.3.3 Two Types of Errors

Please look at Table 8.8. Statisticians find Table 8.8 to be a very useful way to think about a test of hypotheses. I will attempt to explain it to you.

The table consists of two rows and two columns, but unlike such tables earlier in this chapter, this table is not for data presentation.

1. The two rows of the table correspond to the two possible actions, or conclusions, available to the researcher:
    - The researcher can fail to reject the null hypothesis; or
    - The researcher can reject the null hypothesis.
2. The two columns correspond to the two possibilities for reality:
    - The null hypothesis can be correct; or
    - The alternative can be correct.

We see that the columns are an idealization; as we saw in Chapter 5, it is possible for the Skeptic (null hypothesis) to be incorrect and also for every alternative—even $\neq$—to be incorrect. For the current purposes, we are not concerned with this possibility. Thus, for example, for ease of exposition, I will sometimes refer to the second column as corresponding to the null hypothesis being false.

The four cells in the table represent all possible combinations of the two rows with the two columns.

1. The cells on the **main diagonal** correspond to correct actions, or conclusions:
    - The researcher fails to reject a correct null hypothesis; or
    - The researcher rejects a false null hypothesis.
2. The cells on the **off diagonal** correspond to incorrect actions, or conclusions:
    - The researcher rejects a true null hypothesis, called a **Type 1 error**; or
    - The researcher fails to reject a false null hypothesis, called a **Type 2 error**.

Rather obviously, an honest researcher prefers to take a correct action and to avoid our two types of errors.

Before data are collected, the researcher is uncertain about which action will be taken (which row will occur). In addition the researcher is always uncertain about the truth. The uncertainty in the rows (actions) will be dealt with using the ideas of probability, as you will see. The uncertainty in the columns, however, is a different matter. I will explain the popular method and then make a few comments about a different possible approach.

The popular approach is to *condition on* one of the columns being true. This, indeed, is what I did in Chapters 3–5 and, in fact, what I *will do* for all of the tests in these *Course Notes*. In particular, all of our analyses to date have been conditional on the assumption that the Skeptic is correct; i.e., that the null hypothesis is true. (In Chapter 9 we will consider what happens if we condition on the *second column being correct.)*

Look at the first column in Table 8.8 again. Conditional on the null hypothesis being correct, there are only two possible actions: the correct action—failing to reject the null hypothesis—and the incorrect action—rejecting the null hypothesis and making a Type 1 error. Thus, if you read that a test has, say, an 8% chance of making a Type 1 error, literally this means that *on the assumption that the null hypothesis is true* there is an 8% probability that the test will mistakenly reject the null hypothesis. The usually unstated implication is, of course, that there is a 92% probability that the test will correctly fail to reject the null hypothesis. Note that we know **nothing** about how the test performs if the null hypothesis is false—this will be the topic of Chapter 9.

There is another approach which I will briefly describe; view the material in the remainder of this subsection as optional enrichment. If you have heard of **Bayesian analysis** or the **Bayesian approach** and are curious about it, then you might want to continue reading.

The other approach is called the **Bayesian approach**. Before collecting data, the researcher specifies his/her personal probability that the null hypothesis is true. For example, Bert the researcher might state, "The probability that the null hypothesis is true is equal to 10%." After the data are collected, Bert can update his personal probability. Updating is achieved by applying something called **Bayes' formula**, which you will learn about later in these *Course Notes*. As a result of the updating, after analyzing the data, Bert might say, "My new probability that the null hypothesis is true is equal to 20%."

Bayes' formula for updating probabilities is great, but I don't like applying it to test of hypotheses. There are two reasons I feel this way.

1. It can be difficult to obtain widespread acceptance of *personal probabilities*. For example, suppose that another researcher—call her Sally—is investigating the same phenomenon as Bert. She states, "The probability that the null hypothesis is true is equal to 90%." Without going into details, you will likely agree that it seems reasonable that Bert and Sally could draw wildly different conclusions from the same data.

2. This is my stronger reason. I feel that, to a certain extent, the probability that the null hypothesis is correct is not that interesting. Why? Because I believe in Occam's razor. A main idea of tests of hypotheses is to discard a simpler explanation only if it appears to be inadequate; whether the simpler explanation is *likely to be true*, to me, is largely irrelevant.

8.3.4 The Significance Level of a Test

Let's return to my earlier example of Fisher's test with margins $n_1 = n_2 = 27$, $m_1 = 31$ and $m_2 = 23$. I proposed the following critical regions for the three possible alternatives:

- For the alternative $>$, the critical region is $(X \geq 7/27)$ which gives 0.0489 as the probability that the test will make a Type 1 error.
- For the alternative $<$, the critical region is $(X \leq -7/27)$ which gives 0.0489 as the probability that the test will make a Type 1 error.
- For the alternative $\neq$, the critical region is $(|X| \geq 9/27)$ which gives 0.0267 as the probability that the test will make a Type 1 error.

The probability of a Type 1 also is called the **significance level** of a test and is denoted by α (the lower case Greek letter 'alpha'). Thus, for the critical regions given above, $\alpha = 0.0489$ for the alternatives $>$ or $<$ and $\alpha = 0.0267$ for the alternative $\neq$.

Many textbooks state, "Researchers usually take $\alpha = 0.05$," which begs the question, "Why didn't I use $\alpha = 0.05$?" This is an easy question to answer: It is impossible to have $\alpha = 0.05$ for the margins that I have given you. Indeed, there are only a finite number of possible choices for α for this study; in particular, the possible values of α for the alternative $>$ coincide with the entries in the '$>$' column in Table 8.7; e.g., α could be 0.0133 or 0.0489 or 0.1355 to name three possibilities, but it cannot be 0.05. (In fact, my strange choice of margins—27, 27, 31 and 23—reflected my desire to have a significance level that is close to 0.05; I discovered these happy margins by trial-and-error.)

8.4 Two Final Remarks

In this section I will examine two issues that frequently arise while using Statistics in scientific problems. In my experience, there is a great deal of confusion about these issues, but both may be resolved quite easily, thanks to the concept of critical regions.

8.4.1 Choosing the Alternative after Looking at the Data: Is it Really Cheating?

Yes, it is cheating, as I will now demonstrate. To keep this presentation brief, I will illustrate my conclusion with only one example. I hope that you will see that this example can be generalized; if not, I hope that you will trust me on this.

Refer to Example 8.5 on page 177 and its marginal counts, presented in Table 8.7 on page 178. Recall that we found critical regions and significance levels (α's) for the three possible alternatives:

- For the alternative $>$, the critical region is $(X \geq 7/27)$ and the significance level is $\alpha = 0.0489$.

- For the alternative $<$, the critical region is $(X \leq -7/27)$ and the significance level is $\alpha = 0.0489$.
- For the alternative $\neq$, the critical region is $(|X| \geq 9/27)$ and the significance level is $\alpha = 0.0267$.

Next, consider a hypothetical researcher who I will call **Hindsight Hank** (Hank, for short). Hank wants to have α smaller than 0.05, but as close to 0.05 as possible. Thus, once Hank chooses his alternative, he should use one of the critical regions listed above.

Hank, however, refuses to select his alternative before collecting data and, instead, proceeds as follows.

- If the observed value of the test statistic, x, is greater than 0, then Hank says, "Well, obviously, $<$ is inconceivable. Thus, my alternative is $>$, my critical region is $(X \geq 7/27)$ and $\alpha = 0.0489$."
- If the observed value of the test statistic, x, is smaller than 0, then Hank says, "Well, obviously, $>$ is inconceivable. Thus, my alternative is $<$, my critical region is $(X \leq -7/27)$ and $\alpha = 0.0489$."

  (Note that for the sampling distribution given in Table 8.7, the observed value of the test statistic **cannot equal zero**. For situations in which zero is a possible value of the test statistic, the argument below needs to be modified, but the basic important idea remains the same.)

We see that Hank's actual critical region is $(|X| \geq 7/27)$; i.e., if the observed value x satisfies:

$$x \geq 7/27 \text{ or } x \leq -7/27,$$

then Hank will reject the null hypothesis. From Table 8.7 we can see that Hank's actual α is equal to 0.0978.

> In summary, I label Hank's behavior to be cheating because he claims to have $\alpha = 0.0489$, but his actual α is twice as large!

8.4.2 The Two-Sided Alternative Revisited

Suppose that you choose the alternative $\neq$ and have the sampling distribution given in Table 8.7. You choose the critical region $(|X| \geq 9/27)$, which gives $\alpha = 0.0267$. After data are collected, your observed value of the test statistic, x, equals or exceeds $9/27$. Thus, the action is to reject the null hypothesis. Here is the question I address in this subsection:

> Which of the following is correct?
>
> 1. The scientific conclusion is that $p_1 > p_2$.
> 2. The scientific conclusion is that $p_1 \neq p_2$.

(If the answer seems obvious, please bear with me.) This can be an important issue for a scientist. For example, if the treatments are different medical protocols and a success is preferred to a failure, then the first conclusion indicates that that treatment 1 is preferred to treatment 2, whereas the second conclusion simply indicates that the treatments differ.

I have witnessed rather heated discussions over which conclusion is correct. Thus, I will spend a few minutes explaining why the first conclusion is correct; i.e., it is proper to conclude that treatment 1 is better than treatment 2. (Remember that successes are preferred. Also, remember that our conclusions in Part I of these notes are quite limited; I am saying that the proper conclusion is that **for the units being studied** treatment 1 is superior to treatment 2.)

My argument is actually quite simple, once I reveal a fact to you.

You have noted, no doubt, that I allow only two hypotheses in our tests of hypotheses. There are two reasons that all introductory statistics texts make this restriction.

1. Having exactly two hypotheses is the norm in scientific work.

2. Allowing for three or more hypotheses will result, for the most part, in more work for you with little benefit.

The current situation, however, is an exception to the *for the most part* disclaimer.

To a professional statistician, the pair of hypothesis:

$$H_0\text{: The Skeptic is correct; and } H_1\text{: } p_1 \neq p_2,$$

is *shorthand* for the three hypothesis problem:

$$H_0\text{: The Skeptic is correct; } H_1\text{: } p_1 > p_2\text{; and } H_2\text{: } p_1 < p_2.$$

The critical region is actually:

$$\text{If } X \geq 9/27, \text{ then reject } H_0 \text{ in favor of } H_1\text{; and if } X \leq -9/27, \text{ then reject } H_0 \text{ in favor of } H_2.$$

We can see that the probability of a Type 1 error (rejecting a true null hypothesis for either conclusion) is:

$$P(X \geq 9/27) + P(X \leq -9/27) = P(|X| \geq 9/27) = 0.0267.$$

In summary, if one selects the two-sided alternative and rejects the null hypothesis, then the proper scientific conclusion is:

- $p_1 > p_2$ if $x > 0$; or

- $p_1 < p_2$ if $x < 0$.

Although the example of this subsection is for a dichotomous response and Fisher's test, the conclusion remains the same for the other two tests of Part I of these notes and, indeed, for analogous situations in population-based inference in Part II of these *Course Notes*.

8.5 Summary

In this chapter we consider CRDs with a dichotomous response. When the response is dichotomous, one possible response is labeled a **success** and the other a **failure**. Our suggested three *rules* for assigning these labels are given on page 168. We summarize our data by computing $\hat{p}_1$ [$\hat{p}_2$], which is the proportion of successes on treatment 1 [2] in the data. Often, we also compute the (redundant) $\hat{q}_1$ [$\hat{q}_2$], which is the proportion of failures on treatment 1 [treatment 2] in the data. Note that I say *redundant* because:

$$\hat{p}_1 + \hat{q}_1 = 1 \text{ and } \hat{p}_2 + \hat{q}_2 = 1.$$

If we identify the number '1' with a success and the number '0' with a failure, we see that we can match our current notation with our earlier notation for a numerical response, namely:

$$\hat{p}_1 = \bar{x} \text{ and } \hat{p}_2 = \bar{y}.$$

This identification is very helpful because it implies that **all** of our earlier work on:

> the Skeptic's argument; specify the hypotheses; the test statistic U; sampling distributions; and the rules for computing P-values

apply immediately to the studies of this chapter. In particular, we define p_1 [p_2] (note, no hat) to be the proportion of successes the researcher would have obtained if the **All Treatment-1 [2] study had been performed**. If the **clone-enhanced study** could be performed, then we would know the values of both p_1 and p_2.

As before, the null hypothesis is that the Skeptic is correct. The three options for the alternative are given on page 171. With a dichotomous response, our test is called Fisher's test.

Finally, there is a wonderful website,

`http://www.langsrud.com/fisher.htm`,

that is easy to use and gives us the exact P-value for Fisher's test for every choice of alternative. When using this website, recall that:

- *Left* represents the alternative $<$;
- *Right* represents the alternative $>$; and
- *2-Tail* represents the alternative $\neq$.

It is convenient to use the symbol X to represent the test statistic for Fisher's test, with observed value

$$x = \hat{p}_1 - \hat{p}_2. \tag{8.1}$$

We don't actually compute x to obtain our P-value; the website above takes the counts a, b, c and d as input, saving us from the tedium of the dividing and subtracting needed to obtain x.

The **critical region** of a test consists of the collection of all values of the test statistic that would result in the rejection of the null hypothesis.

Table 8.8 displays the four possible combinations of action (by the researcher) and truth (known only to Nature) that could occur in a test of hypotheses. It is reproduced below:

Action	Truth (Only Nature knows)	
(by researcher)	H_0 is correct	H_1 is correct
Fails to reject H_0	Correct action	Type 2 Error
Rejects H_0	Type 1 Error	Correct action

After data are collected, the researcher rejects the null hypothesis if, and only if, the observed value of the test statistic lies in the critical region. Thus, the data determine which action and, hence, row of the table is relevant. The columns of the table are more problematic. The standard approach—which we will follow in these *Course Notes*—is to study the columns one-at-a-time. In particular, first we condition on the null hypothesis being true. You are familiar with this idea; we used it to obtain the sampling distributions for our three tests to date, comparisons of: means, mean ranks and proportions.

When statisticians talk about the probability of a Type 1 error, they really mean the probability of making a Type 1 error *conditional* on the assumption that the null hypothesis is true. (Conditional probabilities will be discussed more carefully in Chapter 16.) The probability of a Type 1 error is called the **significance level** of the test and is denoted by α.

Typically, before collecting data a researcher selects a target value for α; usually—but not exclusively— the target is 0.05. After the exact or approximate sampling distribution is obtained, using trial-and-error the researcher determines a critical region which gives a value of α which is close to the target value. I have shown you one example of this *search* for a critical region and will show you others in the Practice Problems below.

Sometimes a researcher specifies that the actual α should be as close as possible to the target, **without exceeding the target**. Thus, for example, if the target is 0.05, such a researcher would prefer $\alpha = 0.0450$ over, say, $\alpha = 0.0520$, even though the latter is closer to the target. I call this approach **The Price is Right paradigm**, but this name is unlikely to become widely used. (See also Practice Problem 4.)

Section 8.4 shows why it is considered to be cheating to look at one's data before selecting the alternative hypothesis. Also, Section 8.4 shows that if one rejects the null hypothesis after selecting a two-sided alternative, the proper scientific conclusion is the one-sided alternative supported by the data.

Finally, Chapter 9 will address the issue of determining the probability of a Type 2 error of a test.

Table 8.9: The 2×2 contingency table of observed counts for Sara's golf study. A response is a success if, and only if, it equals or exceeds 100 yards.

	Response			Row Proportions	
Club:	S	F	Total	S	F
3-Wood	31	9	40	0.775	0.225
3-Iron	20	20	40	0.500	0.500
Total	51	29	80		

Table 8.10: The 2×2 contingency table of observed counts for Reggie's dart study. A response is a success if, and only if, it equals or exceeds 200 points.

	Response			Row Proportions	
Distance:	S	F	Total	S	F
10 Feet	9	6	15	0.60	0.40
12 Feet	5	10	15	0.33	0.33
Total	14	16	30		

8.6 Practice Problems

1. Refer to Sara's golf data that are presented in Table 2.2 on page 29. Suppose that, being a novice golfer, Sara is mostly interested in *not embarrassing* herself. Thus, she decides that a response of 100 yards or more is a success and a response of less than 100 yards is a failure. With this definition, Sara's data are presented in Table 8.9. (I recommend that you verify the counts in this table.)

   (a) Suppose that Sara chooses the alternative $>$. Explain what this means in terms of the *Inconceivable Paradigm*.

   (b) Find the exact P-value for Sara's data for Fisher's test and the alternative $>$.

   (c) Recall that, based on simulation experiments, approximate P-values for Sara's data and the alternative $>$ are:

   - 0.0293 for the sum of ranks test; and
   - 0.0903 for the test that compares means.

   Compare these P-values to the answer you obtained in (b) and comment.

2. Refer to Reggie's dart data that are presented in the Chapter 1 Homework Problems on page 25. Suppose that Reggie decided that a response of 200 or more points is a success

and a response of fewer than 200 points is a failure. With this definition, Reggie's data are presented in Table 8.10. (I recommend that you verify the counts in this table.)

(a) Suppose that Reggie chooses the alternative $>$. Explain what this means in terms of the *Inconceivable Paradigm*.

(b) Find the exact P-value for Reggie's data for Fisher's test and the alternative $>$.

(c) Based on simulation experiments, approximate P-value for Reggie's data and the alternative $>$ are:

- 0.0074 for the sum of ranks test; and
- 0.0050 for the test that compares means.

Compare these P-values to the answer you obtained in (b) and comment.

3. In this problem I want to reinforce the ideas of finding a critical region for Fisher's test. Suppose that you want to perform a Fisher's test for a table with the following marginal totals:

$$n_1 = n_2 = m_1 = m_2 = 20.$$

(a) Use trial-and-error to find the critical region for the alternative $>$ and α as close to 0.05 as possible.

(b) Use trial-and-error to find the critical region for the alternative $>$ and α as close to 0.10 as possible.

(c) Use trial-and-error to find the critical region for the alternative $<$ and α as close to 0.05 as possible.

(d) Use trial-and-error to find the critical region for the alternative $\neq$ and α as close to 0.05 as possible.

4. This problem introduces the idea of finding the critical region for the test that compares means. In this problem I will use an exact sampling distribution. In Chapter 9 we will consider using an approximate sampling distribution.

Table 8.11 presents the frequency distribution of u for the 252 possible assignments for Kymn's study of rowing. (If this table seems familiar to you it's because you saw it in Table 3.7 on page 65.)

(a) Find the critical region for the alternative $>$ for α as close to the target value 0.05 as possible.

(b) Many statisticians are fans of the television show *The Price is Right*. (I don't actually know this; it's just my segue to the following idea.) In particular, they want α to be *as close to the target as possible without exceeding it*. Repeat part (a) with this **The Price is Right paradigm**.

(c) Find the critical region for the alternative $<$ for α as close to the target value 0.05 as possible.

Table 8.11: Frequency table for u for the 252 possible assignments for Kymn's study.

u	Freq.	u	Freq.	u	Freq.	u	Freq.	u	Freq.	u	Freq.
-7.2	1	-4.8	3	-2.4	10	0.4	12	2.8	10	5.2	4
-6.8	1	-4.4	5	-2.0	8	0.8	10	3.2	8	5.6	3
-6.4	1	-4.0	8	-1.6	14	1.2	13	3.6	6	6.0	1
-6.0	1	-3.6	6	-1.2	13	1.6	14	4.0	8	6.4	1
-5.6	3	-3.2	8	-0.8	10	2.0	8	4.4	5	6.8	1
-5.2	4	-2.8	10	-0.4	12	2.4	10	4.8	3	7.2	1
				0.0	16					Total	252

(d) Find the critical region for the alternative $\neq$ for α as close to the target value 0.05 as possible.

5. Refer to the sampling distribution given in the previous problem, see Table 8.11. Recall **Hindsight Hank**, who was introduced in Section 8.4. Hank decides to proceed as follows:

   - If $x \geq 0$, he declares $>$ to be his alternative and $(U \geq 4.8)$ to be his critical region.
   - If $x < 0$, he declares $<$ to be his alternative and $(U \leq -4.8)$ to be his critical region.

   (a) According to Hank, what is his value of α?

   (b) What is Hank's actual value of α?

6. Recall that in Kymn's actual study, $u = 7.2$. Suppose that she had chosen the alternative $\neq$; what would be her appropriate scientific conclusion?

8.7 Solutions to Practice Problem

1. (a) Sara decided that the alternative $<$ was inconceivable. This inconceivable alternative states that if Sara could have performed the clone-enhanced study, then there would have been more successes with the 3-Iron than with the 3-Wood.

   (b) I entered the values 31, 9, 20 and 20 in the Fisher's test website and obtained 0.0096 as the P-value for the alternative $>$.

   (c) The sum of ranks test gives a much smaller P-value than the test that compares means. In addition, Fisher's test gives a much smaller P-value than the sum of ranks test.

2. (a) Reggie decided that the alternative $<$ was inconceivable. This inconceivable alternative states that if Reggie could have performed the clone-enhanced study, then there would have been more successes from 12 feet than from 10 feet. In short, Reggie felt that it was inconceivable for the greater distance to result in better accuracy.

(b) I entered the values 9, 6, 5 and 10 in the Fisher's test website and obtained 0.1362 as the P-value for the alternative $>$.

(c) The sum of ranks test and the test that compares means give similar and very small approximate P-values. The P-value from Fisher's test is much larger than the other two P-values.

3. (a) My first guess is $a = 13$, which implies $b = c = 7$ and $d = 13$. The site gives me 0.0564 for the exact P-value for $>$. This is a good start!

I next try $a = d = 14$ and $b = c = 6$, which yields 0.0128 for the exact P-value for $>$.

Next, $a = 13$ gives $x = 13/20 - 7/20 = 6/20 = 0.30$. Thus, the critical region is $(X \geq 0.30)$ which has significance level $\alpha = 0.0564$.

(b) Building on my answer to (a), I next try $a = d = 12$ and $b = c = 8$, which yields 0.1715 for the exact P-value for $>$. Thus, the critical region in part (a) yields the value of α closest to 0.10.

(c) By symmetry, the critical region is $(X \leq -0.30)$ which gives $\alpha = 0.0564$ for the alternative $<$.

(d) By symmetry and the work above, the critical region $(|X| \geq 0.30)$ gives $\alpha = 2(0.0564) = 0.1128$ for the alternative $\neq$; and the critical region $(|X| \geq 0.40)$ gives $\alpha = 2(0.0128) = 0.0256$ for the alternative $\neq$. Thus, the latter of these is the answer I seek.

4. (a) Because there are 252 equally likely possible assignments, all probabilities for U will be of the form $k/252$, for some value of k. My target is to obtain the probability 0.05. Thus, I begin by setting

$$k/252 = 0.05 \text{ and solving for } k : k = 0.05(252) = 12.6.$$

By trial-and-error in Table 8.11, I find:

$$P(U \geq 4.8) = 14/252 = 0.0556 \text{ and } P(U \geq 5.2) = 11/252 = 0.0437.$$

Thus, the critical region is $(U \geq 4.8)$ and $\alpha = 0.0556$.

(b) The critical region chosen in part (a) violates *The Price is Right* paradigm because its α exceeds the target, 0.05. Instead, we use the critical region $(U \geq 5.2)$ which gives $\alpha = 0.0437$.

(c) By symmetry, the critical region is $(U \leq -4.8)$ and $\alpha = 0.0556$.

(d) By symmetry, I want to find the number c such that $P(U \geq c)$ is as close as possible to 0.025, one-half of the target value. I begin by setting

$$k/252 = 0.025 \text{ and solving for } k : k = 0.025(252) = 6.3.$$

By trial-and-error in Table 8.11, I find:

$$P(U \geq 5.6) = 7/252 = 0.0278$$

is the closest value to 0.025. Thus, the critical region is $(|U| \geq 5.6)$ and $\alpha = 2(0.0278) = 0.0556$.

5. (a) According to Hank, if his $x \geq 0$, he would say:

$$\alpha = P(U \geq 4.8) = 14/252 = 0.0556.$$

Alternatively, if his $x < 0$, he would say:

$$\alpha = P(U \leq -4.8) = 14/252 = 0.0556.$$

(b) Actually,

$$\alpha = P(U \geq 4.8) + P(U \leq -4.8) = 28/252 = 0.1111.$$

6. Kymn would reject the null hypothesis and, because $x = 7.2 > 0$, her appropriate conclusion is that $\mu_1 > \mu_2$.

8.8 Homework Problems

1. Mary, a student in my class, conducted a study that she called *Is Pearl Ambidextrous?* The balanced study consisted of a total of 50 trials. Treatment 1 [2] was a canter depart on a left [right] lead. According to Mary—I don't know about these things—a canter can be executed successfully or not. Pearl, a seven year-old mare, obtained a total of 42 successes, with 22 successes coming on treatment 1.

   (a) Present Mary's data in a 2×2 table; calculate the proportions of successes; and comment.

   (b) Find Mary's exact P-value for each of the three possible alternatives. Comment.

2. Robert, a student in my class, enjoyed target shooting with his rifle. He performed a balanced study on a total of 100 trials. Treatment 1 [2] was shooting from the prone [kneeling] position. A shot was labeled a success if it hit a specified region of the target. Robert obtained a total of 67 successes, with 25 coming from the kneeling position.

   (a) Present Robert's data in a 2×2 table; calculate the proportions of successes; and comment.

   (b) In his report, Robert stated:

   > Everyone believes that shooting from the prone position is more accurate than shooting from the kneeling position.

   Given this belief, what should Robert use for his alternative hypothesis?

   (c) Find Robert's exact P-value for each of the three possible alternatives. Comment on the P-value for your answer in part (b).

3. Suppose that you want to perform a Fisher's test for a table with the following marginal totals:

$$n_1 = n_2 = m_1 = m_2 = 25.$$

(If you have difficulty answering the questions below, refer to Practice Problem 3.)

Table 8.12: Frequency table for the values r_1 of R_1 for the 252 possible assignments for a balanced CRD with $n = 10$ units and 10 distinct response values.

r_1	Freq.	r_1	Freq.	r_1	Freq.	r_1	Freq.	r_1	Freq.	r_1	Freq.
15	1	19	5	23	14	28	20	33	11	37	3
16	1	20	7	24	16	29	19	34	9	38	2
17	2	21	9	25	18	30	18	35	7	39	1
18	3	22	11	26	19	31	16	36	5	40	1
				27	20	32	14				
										Total	252

(a) Use trial-and-error to find the critical region for the alternative $>$ and α as close to 0.05 as possible.

(b) Use trial-and-error to find the critical region for the alternative $<$ and α as close to 0.05 as possible.

(c) Use trial-and-error to find the critical region for the alternative $\neq$ and α as close to 0.05 as possible.

4. Table 8.12 presents the frequency distribution of the values of r_1 for the sum of ranks test for a balanced CRD with a total of $n = 10$ units and no tied values. (You saw this table previously in Table 6.11.) If you have difficulty with the questions below, refer to Practice Problem 4.

   Find the critical region for the test statistic R_1 and the alternative $>$ that gives α as close as possible to my target value 0.05, **without exceeding the target**.

Chapter 9

Statistical Power

9.1 Type 2 Errors and Power

Table 9.1 is a reproduction of Table 8.8 in Section 8.3 that presented the ideas of Type 1 and Type 2 errors. In Chapter 8 we focused on the first column of this table, the column which states that the null hypothesis is correct. In particular, by assuming that the null hypothesis is correct, we are able to obtain the exact or an approximate sampling distribution for a test statistic. This work led to the new notion of a critical region and its associated significance level for a test. For most of the current chapter we will examine what happens when we assume that the alternative hypothesis is true.

There are two things to remember about assuming that the alternative hypothesis is true:

1. As demonstrated in Chapter 5, whereas there is only one way for the null hypothesis to be true—namely, that the Skeptic is correct—there are figuratively, and sometimes literally, an infinite number of ways for the alternative hypothesis to be true. As a result, we will be able to obtain probabilistic results only by assuming that the alternative is true **in a particular way**.

2. The first step to focusing on the second column—i.e., seeing what happens if one assumes the alternative hypothesis is correct—is to obtain the critical region of a test. In other words, we must study what happens when the null hypothesis is correct before we can hope to study what happens when the alternative is correct.

Table 9.1: Types 1 and 2 errors in a test of hypotheses.

Action	Truth (Only Nature knows)	
(by researcher)	H_0 is correct	H_1 is correct
Fails to reject H_0	Correct action	Type 2 Error
Rejects H_0	Type 1 Error	Correct action

As you might suspect, the ideas behind this chapter are pretty complicated. As a result, in the next section I will present an extended complete analysis of a very simple and familiar study. First, however, let me present a few basic ideas.

We are in a situation in which we have a sampling distribution for a test statistic (exact or approximate) and a target value for α, the significance level of the test. We have found a critical region which has significance level which is close to the target. (You have had practice doing this, in the Practice Problems and Homework of Chapter 8.)

The critical region serves the following purpose. After the data are collected, the observed value of the test statistic is found either to lie in the critical region or not to lie in the critical region; in the former case, the null hypothesis is rejected and in the latter case the researcher fails to reject the null hypothesis. In other words, by comparing the critical region to the data—as summarized by the observed value of the test statistic—we find out which **row** of Table 9.1 is occurring.

When we focused on the first column of Table 9.1, the possible actions were: fail to reject the null hypothesis (the correct action for column 1); and reject the null hypothesis (a Type 1 error). Each of these actions has a probability of occurring and because there are only two possible actions, the two probabilities must sum to one. I mentioned that we denote the probability of a Type 1 error by the symbol α. Obviously, the probability of **correctly failing to reject a true null hypothesis** is equal to $(1 - \alpha)$. I have never witnessed **anybody** referring to this latter probability—let alone giving it a name—either verbally or in writing. So, why do I mention it? You will see very soon.

As mentioned above, when we attempt to calculate probabilities on the assumption that the alternative hypothesis is correct, we must specify **exactly how it is correct**. Once that specification is made, it is sensible to seek the probability of a Type 2 error. The probability of a Type 2 error typically is denoted by β (get it? Type 1: α; Type 2: β; the first two letters of the Greek alphabet). In column 2, however, unlike in column 1, we give a name to $(1 - \beta)$; it is called the **power** of the test for the particular alternative being considered.

The idea is that we would like to have a test that has a low probability of making a Type 2 error; in other words, we would like to have a test that has a large **power**.

9.2 An Extended Example: Cathy's Study of Running

Cathy's study of her running was introduced near the end of Chapter 2. I am guessing that you remember the general motivation of the study, but have not memorized Cathy's data. If I am wrong about the former, please read about it again (Section 2.4 on page 40); if I am correct about the latter, you will appreciate my reproduction of Cathy's data in Table 9.2.

For the purpose of this section, we will suppose that Cathy chose the alternative $>$; i.e., that her times on treatment 1 (the high school) would be larger (i.e., she would run slower) than her times on treatment 2 (the park). This is **not** a ridiculous choice; Cathy might have believed that the natural beauty of the park energized her, resulting in faster times.

In any event, the sampling distribution of U for Cathy's study is given in Table 9.3; it is easy to see that the critical region

$$(U \geq 9.67) \text{ gives } \alpha = 0.05, \text{ exactly.}$$

Table 9.2: Cathy's times, in seconds, to run one mile. HS means she ran at the high school and P means she ran through the park.

Trial:	1	2	3	4	5	6
Location:	HS	HS	P	P	HS	P
Time:	530	521	528	520	539	527

Table 9.3: The sampling distribution of U for Cathy's CRD.

u	$P(U = u)$	u	$P(U = u)$	u	$P(U = u)$
-9.67	0.05	-3.00	0.10	3.67	0.05
-9.00	0.05	-2.33	0.05	4.33	0.05
-7.67	0.05	-1.67	0.05	5.00	0.05
-5.00	0.05	1.67	0.05	7.67	0.05
-4.33	0.05	2.33	0.05	9.00	0.05
-3.67	0.05	3.00	0.10	9.67	0.05

Cathy's actual value of u is 5.00. This value is **not** in the critical region; thus, the action is that Cathy fails to reject the null hypothesis. (Earlier, when we were focusing on P-values, we found that the P-value for Cathy's data and the alternative $>$ is equal to 0.20. Then, as now, the conclusion/action was to fail to reject.)

At this point a student might be tempted to ask: Tell me, Bob, did Cathy take the correct action? My answer, of course, is: Who do you think I am, Nature?

We are now approaching the tricky part of this story.

I am now going to focus on the idea that the alternative hypothesis is correct. I need to specify exactly how it is correct. I make this *executive decision* in two steps:

1. I assume that there is a constant treatment effect, as given in Definition 5.1 on page 91. Because the alternative is $>$, this constant treatment effect must be a positive number.

2. Having decided on the *form of the alternative*—i.e., a constant treatment effect—I must decide on its size. To get this started, I will explore the idea that there is a constant treatment effect of seven seconds.

These two points might seem to be a bit abstract. Let's get more specific.

Cathy's sorted times are:

- 521, 530 and 539 on treatment 1; and

- 520, 527 and 528 on treatment 2.

Table 9.4: The sampling distribution of U for Cathy's CRD on the assumption that there is a constant treatment effect of seven seconds.

u	$P(U=u)$	u	$P(U=u)$	u	$P(U=u)$
-3.00	0.05	5.00	0.10	11.00	0.05
-0.33	0.05	5.67	0.05	11.67	0.05
0.33	0.05	6.33	0.05	12.33	0.05
1.67	0.05	7.67	0.05	13.67	0.05
2.33	0.05	8.33	0.05	14.33	0.05
3.00	0.05	9.00	0.10	17.00	0.05

The assumption of a constant treatment effect of seven seconds means that the trials that gave times 521, 530 and 539 on treatment 1 would have given times $521 - 7 = 514$, $530 - 7 = 523$ and $539 - 7 = 532$, respectively, if they had been assigned to treatment 2. Similarly, the times 520, 527 and 528 that were actually obtained on treatment 2, would have given times $520 + 7 = 527$, $527 + 7 = 534$ and $528 + 7 = 535$, respectively, if they had been assigned to treatment 1.

Now, the details get messy and **I won't ever ask you to produce them**, but I hope you can see that the current situation is much like assuming the Skeptic is correct; namely, we have a theory that tells us exactly what responses would have been obtained for any particular assignment. Because there are only 20 assignments to consider, it is easy for me to determine the exact distribution of U **on the assumption that there is a constant treatment effect of seven seconds**.

The distribution of U for Cathy's study on the assumption that the alternative hypothesis is correct in that there is a constant treatment of seven seconds is given in Table 9.4. I want to make a few comments about the distribution in Table 9.4:

1. If you examine the entries carefully, you will see that the distribution is symmetric about the number 7. This means that both the values of u and the probabilities are symmetric around 7. Let's first look at the values of u. The mean of the smallest possible value of u (-3.00) and the largest possible value of u (17.00) is 7.00. Thus, they are equal distance from 7.00. This pattern continues for all such pairs: the mean of -0.33 and 14.33; the mean of 0.33 and 13.67; and so on, are all equal to 7. Next, note that all probabilities, except for two, are equal to 0.05. The other two, both equal to 0.10, belong to the values $u = 5.00$ and $u = 9.00$ which are symmetric around 7.

2. Based on the previous remark, the mean of the sampling distribution in Table 9.4—being equal to its center of gravity—is 7. This is hardly surprising because the assumption is that there is a constant treatment effect of seven seconds for each trial.

3. Sadly, the distribution in Table 9.4 is **not** obtained by simply shifting the distribution in Table 9.3 *to the right* by seven seconds. (For one of many possible examples of this fact: if you add 7 to the smallest value in the latter table, $u = -9.67$, you **do not obtain** the smallest value in the former table, $u = -3.00$.)

Table 9.5: The sampling distribution of U for Cathy's CRD on the assumption that there is a constant treatment effect of 15 seconds.

u	$P(U=u)$	u	$P(U=u)$	u	$P(U=u)$
2.33	0.05	13.00	0.05	19.67	0.05
5.00	0.05	13.67	0.05	21.67	0.05
7.00	0.05	14.33	0.05	22.33	0.05
7.67	0.05	15.67	0.05	23.00	0.05
8.33	0.05	16.33	0.05	25.00	0.05
10.33	0.05	17.00	0.05	27.67	0.05
11.00	0.05	19.00	0.05		

In fact, there is no simple relationship between the null sampling distribution of U (computed by assuming that the Skeptic is correct) and the numerous sampling distributions that can be obtained by varying the size of the constant treatment effect. (See the next example with a constant treatment effect of 15 seconds.) As a result, studying power for randomization-based inference is a frustrating and tedious process. By contrast, population-based inference (Part II of these notes) is more amenable to studies of power.

Finally, I am ready to say something useful about power and Cathy's study. Remember the critical region for $\alpha = 0.05$ is $(U \geq 9.67)$. In words, the null hypothesis is rejected if, and only if, the observed value of U equals or exceeds 9.67. We see from Table 9.4, that, on the assumption there is a constant treatment effect of seven seconds,

$$P(U \geq 9.67) = 6(0.05) = 0.30.$$

Thus, the power for our chosen alternative is (only) 30%. In words, if it is true that running at the high school adds seven seconds (compared to running through the park) to Cathy's time, there was only a 30% chance that Cathy's study would detect it and correctly reject the null hypothesis.

I repeated the above analysis—details, thankfully, suppressed—for a constant treatment effect of 15 seconds. My results are presented in Table 9.5. From this table, we can calculate the power, based on the assumption of a constant treatment effect of 15 seconds:

$$P(U \geq 9.67) = 1 - P(U < 9.67) = 1 - 5(0.05) = 0.75.$$

As above, if it is true that running at the high school adds 15 seconds (compared to running through the park) to Cathy's time, there was a 75% chance that Cathy's study would detect it and correctly reject the null hypothesis. I consider 75% to be a pretty large value for power. On the other hand, to me—never much of a distance runner—15 seconds seems to be a huge treatment effect. Thus, I am not convinced that my computations are very useful scientifically. Of course, Cathy's study was very small and it is impressive that it has any power!

The above two sampling distributions reflect a general fact that you can rely upon: For the alternative $>$ and the test statistic U, **if we increase the size of the treatment effect, then the power will increase or stay the same.**

Table 9.6: The effect of the choice of α on the power of the test statistic U for Cathy's data.

α	Critical Region	Power for a constant treatment effect of: 7 seconds	15 seconds
0.05	$U \geq 9.67$	0.30	0.75
0.10	$U \geq 9.00$	0.40	0.75
0.15	$U \geq 7.67$	0.50	0.85
0.20	$U \geq 5.00$	0.70	0.95

I end this section by looking at what happens if we change the value of α. My results are summarized in Table 9.6. Let me make a few comments about this table.

1. The table has four rows, corresponding to four choices for α: 0.05, 0.10, 0.15 and 0.20. In my experience, it is rare for a researcher to choose an α that is larger than 0.10, but with such a small study (only 20 possible assignments) please grant me some latitude.

2. The first row of the table is a restatement of the earlier work in this subsection. Namely, for $\alpha = 0.05$, from Table 9.3 the critical region is $(U \geq 9.67)$ and the powers, 0.30 and 0.75, were found by examining Tables 9.4 and 9.5, respectively.

3. In a similar manner, for $\alpha = 0.10$, we can see from Table 9.3 that

$$P(U \geq 9.00) = 0.10;$$

   thus, we have our critical region for $\alpha = 0.10$. For power, from Table 9.4 we find

$$P(U \geq 9.00) = 0.40,$$

   and from Table 9.5 we find

$$P(U \geq 9.00) = 0.75.$$

4. In similar manners, you can verify the remaining entries in Table 9.6.

The obvious conclusion from Table 9.6 is: if we increase the value of α, then the power will increase or stay the same. (For population-based inference, it will increase.)

9.3 Simulation Results: Sara's Golfing Study

In this section I will use Sara's study, first introduced in Chapter 2, to examine several problems, listed below. **Note that throughout this section, I will use the alternative $>$.**

1. How to find the critical region for the test that compares means.

2. How to calculate power for the test that compares means.

3. How to find the critical region for the sum of ranks test.

4. How to calculate power for the sum of ranks test.

5. How to decide which test is better for Sara's data: the test that compares means or the sum of ranks test.

In the previous section, I was able to obtain and present exact results on the significance level and power for Cathy's running study because there were only 20 possible assignments, a very manageable number. By contrast, for Sara's study, the number of possible assignments is

$$1.075 \times 10^{23} \text{ (see page 57).}$$

Because this number is so **huge**, we will opt for an approximation based on a computer simulation experiment with $m = 10{,}000$ reps.

You might recall that in Chapter 4 I performed a computer simulation experiment on Sara's data and the test statistic U. With 10,000 reps, I obtained 723 distinct observed values of U, far too many for an analysis in which I show you all of the details! As a result, I will provide you with only the necessary summaries of my simulation results. Thus, I am leaving you *in the dark* in two ways:

1. I am not going to explain the computer code needed for my simulation experiments; and

2. I am not going to show you the raw results of my simulation experiments; they are just too cumbersome and tedious!

Recall that in earlier simulation studies, the P-value I obtained for the test that compares means was much larger than the P-value I obtained for the sum of ranks test (0.0903 for the former and 0.0293 for the latter). As a result, I am particularly interested in determining which test is *better*—i.e., more powerful—for Sara's data.

It would be possible to perform my analysis with two different simulation experiments: one for the test statistic U and another for the test statistic R_1. It turns out to be a superior strategy, however, to investigate the two test statistics simultaneously. (The sense in which this new strategy is better will be discussed much later in these *Course Notes*; sorry.)

The obvious question is: What do I mean by *investigating the two test statistics simultaneously?* It is easiest to see what I mean if we look back at Cathy's study. Please refer to her actual study results, presented in Table 9.2 on page 195. I would never, of course, use a computer simulation experiment to approximate the sampling distribution of either U or R_1 for Cathy's study; with only 20 possible assignments the exact sampling distributions are easily obtained. (I am reminded of former President Richard Nixon who reportedly said, "I would **never** bribe everyone who knows the truth about Watergate, but how much would it cost?") Suppose, for example, that on the first rep, my simulation experiment selects the assignment 3,4,5. With this choice, the sorted data become:

- Treatment 1: 520, 528, 539; with ranks 1, 4, 6; and
- Treatment 2: 521, 527, 530; with ranks 2, 3, 5.

Thus,

$$u = \bar{x} - \bar{y} = \frac{520 + 528 + 539}{3} - \frac{521 + 527 + 530}{3} = 529 - 526 = 3; \text{ and } r_1 = 1 + 4 + 6 = 11.$$

In words, every rep yields observed values for both U and R_1. By performing 10,000 reps, we can build up a **joint distribution** of values of U and R_1. This is what I did for Sara's data; I will describe my findings below.

First, I need to determine my approximate critical region for each test. By the way, I hate to keep typing *approximate*; can we agree to remember that all answers from computer simulation experiments are necessarily approximations? Assuming your answer is yes, I will dispense with the typing of the word *approximate*.

I choose to have a target of 0.05 for α for both tests. Why? Two reasons:

1. Why 0.05? Because in my experience, it is the most popular choice among statisticians and researchers.
2. Why the same target for both tests? So that the comparison will be *fair*. As shown in Table 9.6, as the value of α increases, a test becomes more powerful. If I were to show, for example, that the test statistic U with $\alpha = 0.10$ is more powerful than the test statistic R_1 with $\alpha = 0.01$, what would I have accomplished? (My answer: Nothing of value.)

My simulation experiment was quite successful in achieving an α close to my target value of 0.05:

- For the test that compares means,

$$\text{Rel. Freq. } (U \geq 10.65) = 0.0499; \text{ and}$$

- For the sum of ranks test,

$$\text{Rel. Freq. } (R_1 \geq 1788.0) = 0.0499.$$

Thus, my two critical regions are:

$$(U \geq 10.65) \text{ and } (R_1 \geq 1788.0),$$

both yielding $\alpha = 0.0499$.

In the previous paragraph, I consider the tests separately; i.e., I don't describe how U and R_1 vary together. For the latter, please refer to the results in Table 9.7. I will take a few minutes to describe the information displayed in this table.

Table 9.7: Decisions made by U and R_1 for the 10,000 data sets that were obtained in the computer simulation experiment on Sara's data under the assumption that the Skeptic is correct. The critical regions are $(U \geq 10.65)$ and $(R_1 \geq 1788.0)$.

	R_1		
U	Fail to Reject H_0	Reject H_0	Total
Fail to reject H_0	$a = 9{,}374$	$b = 127$	$a + b = 9{,}501$
Reject H_0	$c = 127$	$d = 372$	$c + d = 499$
Total	$a + c = 9{,}501$	$b + d = 499$	$m = 10{,}000$

1. Literally, Table 9.7 has two rows and two columns of numbers—along with various totals—but it is importantly different from our data tables for a CRD in Chapter 8. (We will see tables like this again in Chapter 16.)

   - The two rows correspond to the two possible actions—fail to reject and reject—possible with test statistic U; and
   - The two columns correspond to the two possible actions—again, fail to reject and reject—possible with test statistic R_1.

2. The table's cells correspond to the four possible combinations of the tests' joint actions. For example, the upper left cell corresponds to both tests failing to reject and the lower right cell corresponds to both tests rejecting.

3. I use the *Chapter 8 notation* for the cell counts: a, b, c and d. **I do not use the *Chapter 8 notation* for the marginal totals.** For example, following Chapter 8, if I denoted the first row total, 9,501, by n_1, this would be *wrong* because, in Sara's study, the symbols n_1 and n_2, both 40, have already been claimed.

4. It is extremely important for you to remember that Table 9.7 was generated under the assumption that the null hypothesis is true. Dating back to Chapters 3–5, we obtain P-values and critical regions by assuming the Skeptic is correct. Soon we will see what happens if we assume that the Skeptic is wrong.

5. Let's look at the counts in the four cells in Table 9.7. The largest count, by far, is $a = 9{,}374$. This tells us that for 9,374 of the assignments, the tests *correctly agreed* that the true null hypothesis should not be rejected.

   The next largest count, $d = 372$ tells us that for 372 of the assignments, the tests *incorrectly agreed* to reject the true null hypothesis.

   The remaining counts are equal, $b = c = 127$. The equality is no surprise; they must be equal because the two tests have the same α. What *is interesting* is the relative sizes of 372 and 127: these give us a rough idea on how much the tests agree in what they see in the data. For example, if d had been 499 and both b and c had been 0, then the two tests would be in

total agreement on when to reject; in short, there would seem to be no reason to compare the two tests. At the other extreme, if $d = 0$ and $b = c = 499$, then the tests would never agree on when to reject the null hypothesis. Because the two tests are looking at the same data (remember, we analyze each assignment—each data set—with both tests), if I obtained $d = 0$, I would be quite sure that I had made a serious error in my computer program!

It the explanation of this item seems incomplete, note that I will return to this topic later.

9.3.1 A Simulation Study of Power

In this subsection we will focus on what happens in Sara's study if we assume that the alternative hypothesis is true. As discussed previously in these notes, there is figuratively, perhaps even literally, an infinite number of ways for the alternative hypothesis to be correct. As I argued earlier with Cathy's study of running, I will focus on alternatives that correspond to the notion of a constant treatment effect. The alternative hypothesis $>$ implies that that constant treatment effect must be a positive number.

I performed six simulation experiments on Sara's data. Each simulation consisted of 10,000 reps in which each rep evaluated the decisions made by both test statistics, U and R_1. I used the critical regions presented above, namely

$$(U \geq 10.65) \text{ and } (R_1 \geq 1788.0),$$

because with these choices, both tests have the same value for α, 0.0499. The six simulations examined six possible values of the constant treatment effect: 3, 6, 9, 12, 15 and 18 yards. The results of these six simulation experiments are presented in Table 9.8. There is a **huge** amount of information in this table; thus, I will spend a large amount of time explaining it.

1. Let's begin by looking at the first 2×2 table within the larger Table 9.8. We will begin by focusing on the marginal totals.

   The rows of the table correspond to the possible actions taken by test statistic U. We see that for 8,711 assignments, the test fails to reject the null hypothesis (because the assignment yielded $u < 10.65$). For the remaining 1,289 assignments the test rejects the null hypothesis (because the assignment yielded $u \geq 10.65$).

   The simulation was performed on the assumption that there is a constant treatment effect of 3 yards; thus, the Skeptic is wrong, the alternative $>$ is correct, and the correct action by the researcher would be to reject the null hypothesis. The power of the test U for this constant treatment effect of 3 yards is $1{,}289/10{,}000 = 0.1289$, just under 13 percent.

   The columns of the table correspond to the possible actions taken by test statistic R_1. We see that for 8,466 assignments, the test fails to reject the null hypothesis (because the assignment yielded $r_1 < 1788.0$). For the remaining 1,534 assignments the test rejects the null hypothesis (because the assignment yielded $r_1 \geq 1788.0$). The power of the test R_1 for this constant treatment effect of 3 yards is $1{,}534/10{,}000 = 0.1534$, just over 15 percent.

Table 9.8: (Approximate) Power for U and R_1 from repeated simulation experiments for Sara's golf study. 'CTE' stands for the size, in yards, of the constant treatment effect in each simulation. The critical regions for the tests are ($U \geq 10.65$) and ($R_1 \geq 1788.0$).

CTE$= 3$	R_1		
U	Fail to Reject H_0	Reject H_0	Total
Fail to reject H_0	8,313	398	8,711
Reject H_0	153	1,136	**1,289**
Total	8,466	**1,534**	10,000

CTE$= 6$	R_1		
U	Fail to Reject H_0	Reject H_0	Total
Fail to reject H_0	6,934	682	7,616
Reject H_0	209	2,175	**2,384**
Total	7,143	**2,857**	10,000

CTE$= 9$	R_1		
U	Fail to Reject H_0	Reject H_0	Total
Fail to reject H_0	5,006	987	5,993
Reject H_0	174	3,833	**4,007**
Total	5,180	**4,820**	10,000

CTE$= 12$	R_1		
U	Fail to Reject H_0	Reject H_0	Total
Fail to reject H_0	2,973	1,216	4,189
Reject H_0	97	5,714	**5,811**
Total	3,070	**6,930**	10,000

CTE$= 15$	R_1		
U	Fail to Reject H_0	Reject H_0	Total
Fail to reject H_0	1,614	856	2,470
Reject H_0	78	7,452	**7,530**
Total	1,692	**8,308**	10,000

CTE$= 18$	R_1		
U	Fail to Reject H_0	Reject H_0	Total
Fail to reject H_0	737	582	1319
Reject H_0	35	8646	**8681**
Total	772	**9228**	10,000

We see that for a constant treatment effect of 3 yards, the test based on R_1 is more powerful than the test based on U; in words, of the two tests, R_1 is more likely to correctly reject the null hypothesis when there is a constant treatment effect of 3 yards.

This would be a good time to remember that these powers are *approximations* based on a simulation experiment. As you will see later in this section (see the *nearly certain interval*), this difference seems to be real; i.e., too large to be attributed to chance.

At this time, we will ignore the cell counts (a–d) in this table.

2. We did all the *heavy lifting* in the above consideration of a constant treatment effect of 3 yards. The remaining small tables in the large Table 9.8 are now easy to understand. This stated ease, however, will not dissuade me from discussing these small tables.

3. For a constant treatment effect of 6 yards, the power of R_1, 0.2857, exceeds the power of U, 0.2384.

4. For a constant treatment effect of 9 yards, the power of R_1, 0.4820, exceeds the power of U, 0.4007.

5. For a constant treatment effect of 12 yards, the power of R_1, 0.6930, exceeds the power of U, 0.5811.

6. For a constant treatment effect of 15 yards, the power of R_1, 0.8308, exceeds the power of U, 0.7530.

7. Finally, for a constant treatment effect of 18 yards, the power of R_1, 0.9228, exceeds the power of U, 0.8681.

Summarizing the above comments:

- For both tests, an increase in the size of the constant treatment effect leads to an increase in the power.

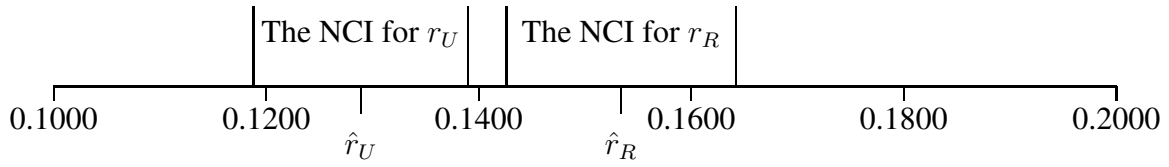
- For every constant treatment effect considered, R_1 is more powerful than U. I will summarize this fact by saying—with a level of generalization that seems excessive—that R_1 is more powerful than U. More colorfully, I will say that for Sara's data, R_1 is a better test than U.

Let me return to something I typed in Chapter 8. I noted that for Sara's data, the P-value for the sum of ranks test (using R_1) was 0.0293, but the P-value for the test that compares means (using U) was much larger, 0.0903. I stated that the smaller P-value for R_1 was *suggestive* that R_1 was the better test, but *not conclusive*.

Based on my above remarks, we have decided that, indeed, R_1 is more powerful than U; well, it is for the six alternatives I examined. Thus, it is reasonable to wonder why we can't somehow reach the same conclusion by looking at P-values. I will now explain why we cannot do so.

Look at any of the smaller tables within Table 9.8 and focus on the number in the lower left cell—the cell whose count is denoted by c. For example, for the table for a constant treatment

Figure 9.1: Two nearly certain intervals (NCI's) for the true power of two tests when there is a constant treatment effect of 3 yards.



effect of 3 yards, $c = 153$. This means that for 153 of the assignments the test based on R_1 failed to reject the null hypothesis while the test based on U correctly rejected the null hypothesis. In particular, this implies that for these 153 assignments, the P-value for U is less than or equal to 0.0499 and the P-value for R_1 is greater than 0.0499. In other words, for these 153 assignments—admittedly only 1.5% of the assignments examined, but not 0%—the test with the smaller P-value was the test with less power. The moral is: If you want to decide which test is better, you need to perform an analysis of power, as we have done in this section.

9.3.2 A New Nearly Certain Interval

Let's return to our earlier consideration of a comparison of power for our two test statistics, U and R_1. In particular, let's look at the approximate power obtained when the constant treatment effect is equal to 3 yards.

The approximate power using R_1 is 0.1534. Using the ideas of Chapter 4, we can obtain the nearly certain interval for the exact power. In particular, let r_R denote the exact (unknown) power and let $\hat{r}_R = 0.1534$ denote its approximation. The nearly certain interval for r_R is

$$0.1534 \pm 3\sqrt{\frac{0.1534(0.8466)}{10{,}000}} = 0.1534 \pm 0.0108.$$

Similarly, the approximate power using U is 0.1289. Let r_U denote the exact (unknown) power and let $\hat{r}_U = 0.1289$ denote its approximation. The nearly certain interval for r_U is

$$0.1289 \pm 3\sqrt{\frac{0.1289(0.8711)}{10{,}000}} = 0.1289 \pm 0.0101.$$

These two nearly certain intervals are presented in Figure 9.1. The two intervals **do not overlap**; thus, we can safely say that R_1 is more powerful than U. Their boundaries, however, are close and this figure suggests a great deal of uncertainty in the difference in the values of the power. If we want to have an idea of *how much more powerful* R_1 is, we need to introduce a new nearly certain interval.

In particular, my goal is to approximate the value of $r_R - r_U$. In the formula below, I will use our standard notation of (a)–(d) for the counts in the appropriate 2×2 table. Thus, in particular,

$$(\hat{r}_R - \hat{r}_U) = (b+d)/m - (c+d)/m = (b-c)/m,$$

where m is the number of reps in the simulation experiment. (Note that m = 10,000 for all examples in this chapter.)

Result 9.1 (The Nearly Certain Interval for the Difference in Power.) Following the notation given above, the nearly certain interval for $r_R - r_U$ is given by:

$$(\frac{b-c}{m}) \pm (3/m)\sqrt{\frac{m(b+c)-(b-c)^2}{m-1}} \tag{9.1}$$

Also, the nearly certain interval for $r_U - r_R$ is given by:

$$(\frac{c-b}{m}) \pm (3/m)\sqrt{\frac{m(b+c)-(c-b)^2}{m-1}} \tag{9.2}$$

I will illustrate the use of Equation 9.1 for our discussion of power for the constant treatment effect of 3 yards. From Table 9.8, we have

$$b = 398 \text{ and } c = 153, \text{ which give } (b+c) = 551 \text{ and } (b-c) = 245.$$

Thus, the nearly certain interval is

$$0.0245 \pm 0.0003\sqrt{\frac{5510000-(245)^2}{9999}} = 0.0245 \pm 0.0070.$$

9.4 Simulation Results: Doug's Study of 301

Practice Problem 1 in Section 4.6 introduced Doug's study of the dart game 301. To summarize: Doug's response was the number of rounds he required to complete a game of 301; his study factor was type of dart, with treatment 1 being his personal darts and treatment 2 being bar darts. For Doug's study, the smaller the value of the response, the better. Doug labeled the alternative $>$ inconceivable because he believed that, if the Skeptic was wrong, then he played better with his personal darts. Thus, his choice of alternative was $<$.

In this section I will use Doug's data to reinforce the ideas presented above for Sara's study. I could, of course, make the current problem *more similar* to Sara's by renumbering the treatments: If I made treatment 1 [2] the bar [personal] darts, then the alternative would be reversed. Instead, I prefer to give you experience dealing with the alternative $<$.

Recall that the following summary statistics were given in Chapter 4:

$$n_1 = n_2 = 20; \bar{x} = 18.60 \text{ and } \bar{y} = 21.20, \text{ giving } u = 18.60 - 21.20 = -2.60.$$

Thus, the data are consistent with Doug's choice of alternative—his $\bar{x}$ being smaller than his $\bar{y}$ is in the same direction as the alternative $\mu_1 < \mu_2$.

In addition, in a Practice Problem beginning on page 134 in Chapter 6, we found that:

$$r_1 = 354, r_1/20 = 17.7, r_2 = 466.0 \text{ and } r_2/20 = 23.3.$$

Table 9.9: Decisions made by U and R_1 for the 10,000 data sets that were obtained in the computer simulation experiment on Doug's data under the assumption that the Skeptic is correct. The critical regions are $(U \leq -2.50)$ and $(R_1 \leq 348.5)$.

	R_1		
U	Fail to Reject H_0	Reject H_0	Total
Fail to reject H_0	9,416	60	9,476
Reject H_0	54	470	524
Total	$a + c =$ 9,470	$b + d =$ 530	$m =$ 10,000

Finally, I performed two simulation studies, each with 10,000 reps, to obtain approximate P-values for the two tests. My results were: 0.0426 for the test statistic U; and 0.0623 for the test statistic R_1. These P-values *suggest* that, for Doug's study, the test that compares means is better than the sum of ranks test. As we will see in this section, a test of power confirms the validity of this suggestion.

Our first task, of course, is to determine the critical regions for both tests. This begins with a specification of our target, which I will take to be 0.05. Unfortunately, the sampling distribution of U is pretty clumpy and the closest I could get to the target is revealed in Table 9.9.

First, note that the critical regions for the two tests are:

$$(U \leq -2.50) \text{ and } (R_1 \leq 348.5).$$

For Doug's actual data, his $u = -2.60$ falls in the critical region; thus, the test U rejects the null hypothesis. Also for Doug's actual data, his $r_1 = 354$ does not fall in the critical region; thus, the test R_1 fails to reject the null hypothesis. The significance levels corresponding to these two critical regions are: $\alpha = 0.0530$ for the sum of ranks test; and $\alpha = 0.0524$ for the test that compares means. As a result, when we discover below that U is more powerful than R_1 for the alternatives I examine, the superiority of U will be a bit more impressive because it has the disadvantage of its significance level being a bit smaller than the significance level for R_1.

There is one other feature of Table 9.9 that should be noted. Its values of $b = 60$ and $c = 54$ are much smaller than the values $b = c = 127$ in Table 9.7. This shows that in the current situation—Doug's data—the two tests have a greater agreement in how they assess the evidence in the data. It is my experience (this is not a theorem) that the better the tests agree, the better the relative performance of the test statistic U.

Next, we turn to the study of power for Doug's data via simulation experiments. Not surprisingly, I will concentrate on alternatives that reflect a constant treatment effect. Note that because Doug's alternative is $<$, the constant treatment effects I study must all be negative. In Table 9.10 I present the results of four simulation experiments, each with $m =$ 10,000 reps.

I will ask you to interpret the numbers in this table in Practice and Homework Problems below. For now, let me draw your attention to two obvious features revealed in this table.

1. For every constant treatment effect considered, the test that compares means is more powerful than the sum of ranks test.

Table 9.10: (Approximate) Power for U and R_1 from repeated simulation experiments for Doug's data. 'CTE' stands for the number, in rounds, of the constant treatment effect in each simulation. The critical regions for the tests are $(U \leq -2.50)$ and $(R_1 \leq 348.5)$.

CTE$= -1$	R_1		
U	Fail to Reject H_0	Reject H_0	Total
Fail to reject H_0	8,443	18	8,461
Reject H_0	211	1,328	**1,539**
Total	8,654	**1,346**	10,000

CTE$= -2$	R_1		
U	Fail to Reject H_0	Reject H_0	Total
Fail to reject H_0	6,193	33	6,226
Reject H_0	401	3,373	**3,774**
Total	6,594	**3,406**	10,000

CTE$= -3$	R_1		
U	Fail to Reject H_0	Reject H_0	Total
Fail to reject H_0	3,539	6	3,545
Reject H_0	758	5,697	**6,455**
Total	4,297	**5,703**	10,000

CTE$= -4$	R_1		
U	Fail to Reject H_0	Reject H_0	Total
Fail to reject H_0	1,490	2	1,492
Reject H_0	651	7,857	**8,508**
Total	2,141	**7,859**	10,000

2. As the constant treatment effect moves farther from 0, the power increases for both tests.

9.5 The Curious Incident ...

Inspector Gregory (IG): You consider that to be important?

Sherlock Holmes (SH): Exceedingly so.

IG: Is there any point to which you would wish to draw my attention?

SH: To the curious incident of the dog in the night-time.

IG: The dog did nothing in the night-time.

SH: That was the curious incident.

The above exchange, from the short story *Silver Blaze* by Sir Arthur Conan Doyle, is one of the most famous passages in detective fiction. (It also led to the title of the excellent novel, *The Curious Incident of the Dog in the Night-Time* by Mark Haddon, published in 2003.)

So, why am I mentioning this bit of literary trivia? Two reasons.

First, I have always found this passage to be extremely important in Statistics. Instead of despairing over the lack of data (the dog not barking), it is important to focus on what it means to have no data (why didn't the dog bark?). In the story, the dog's failure to bark suggests, to Holmes, that the dog knew the villain.

One obvious way this idea manifests itself in science is in medical studies: Why do people drop out of studies? Instead of viewing such subjects as contributing *no data*, one should seek to learn why they dropped out.

Second, after typing the last section (reminding me of my favorite literary quote: When asked to comment on Jack Kerouac's *On the Road*, Truman Capote reportedly said, "That's not writing, it's typing.") I was ready to proceed to the section immediately below, *Computing*. Then I thought: There are several obvious applications of power that I have not presented. Perhaps I should tell the reader why. I have three comments, which I will list below. **Note that all of the comments below, indeed all of the material in this section, is optional enrichment; you will be tested on none of this material!**

1. I make no mention of power for a dichotomous response because there is no natural analogue to the idea of a constant treatment effect alternative. The situation is better for population-based inference, covered in Part II of these *Course Notes*, but still troublesome.

2. I have not mentioned power for the alternative $\neq$. The reason for this omission can be seen most easily and most clearly if we reexamine Cathy's study of running. For the alternative $\neq$, the smallest possible value for α is 0.10 which is obtained by using the critical region $(|U| \geq 9.67)$. Next, consider the power for the alternative of a constant treatment effect of seven seconds; namely, the following probability, using Table 9.4:

$$\text{Power } = P(|U| \geq 9.67) = P(U \geq 9.67) + P(U \leq -9.67) = 0.30 + 0 = 0.30.$$

   In words, for a constant treatment effect that is a positive number, the power for the alternative $\neq$ and $\alpha = 0.10$ is equal to—or well approximated by—the power for the alternative $>$ and $\alpha = 0.05$.

   Why is this so? The mean of the sampling distribution of U will equal the value of the constant treatment effect—we saw this twice earlier for Cathy's data. As a result, the probability that the test statistic will be negative and far enough from 0 to be in the critical region is zero—as above—or very close to zero.

   Because of the above, if I want the power for the alternative $\neq$ for a given α and a positive [negative] value of a constant treatment effect, I simply obtain the the power for the alternative $>$ [$<$] and significance level equal to $\alpha/2$. This procedure usually provides me with a good approximation to the desired exact power.

3. The idea of a constant treatment effect is most natural for a measurement response. It *can* be of use for a count—see Doug's study above—but sometimes does not work well for counts—see the Homework Problem below concerning Dawn's study of her cat Bob.

9.6 Computing

I have found no websites that will perform the simulations we require for determining a critical region. And, given a critical region, I have found no websites that will perform the simulations we require for determining power. Thus, I am not expecting you to recreate any of the results presented in this chapter.

9.7 Summary

In the previous chapter, you learned how to find the critical region for a test for a dichotomous response. In the previous chapter and this chapter, you learned how to find the critical region for a numerical response for which the exact sampling distribution is available. This chapter also looks at approximate sampling distributions obtained by simulation experiments for two data sets and obtains critical regions for both U and R_1 for both sets of data, a total of four critical regions.

Once critical regions are obtained, we turn our attention to situations in which the alternative hypothesis is correct. Our investigation is quite limited, but it reveals some interesting results.

The most striking limitation is that we restrict attention to the constant treatment effect alternatives. Also, we limit attention to one-sided alternatives and the response being a number.

Make sure you are able to understand the information in Tables 9.8 and 9.10; see the Practice and Homework Problems if you need more experience with tables like these.

9.8 Practice Problems

1. Look at the *power table*—Table 9.8—for Sara's data. Consider the 2×2 table for a constant treatment effect of 12 yards. Match each of the nine counts in the table with its description below.

   (a) The total number of assignments examined.

   (b) The number of assignments for which: U failed to reject and R_1 rejected.

   (c) The number of assignments for which: R_1 rejected.

   (d) The number of assignments for which: U failed to reject.

   (e) The number of assignments for which: both U and R_1 failed to reject.

   (f) The number of assignments for which: U rejected.

   (g) The number of assignments for which: R_1 failed to reject.

   (h) The number of assignments for which: both U and R_1 rejected.

   (i) The number of assignments for which: R_1 failed to reject and U rejected.

2. Consider the 2×2 table for a constant treatment effect of 9 yards in Table 9.8.

   Calculate the nearly certain interval for $r_R - r_U$. (See Result 9.1.)

9.9 Solutions to Practice Problems

1. The answers to (a)–(i), respectively are: 10,000; 1,216; 6,930; 4,189; 2,973; 5,811; 3,070; 5,714; and 97.

2. From the table, we identify $b = 987$ and $c = 174$. These give us:

$$b + c = 987 + 174 = \text{ 1,161 and } b - c = 987 - 174 = 813.$$

   Thus, the nearly certain interval (Formula 9.1) is

$$(\frac{b-c}{m}) \pm (3/m)\sqrt{\frac{m(b+c) - (b-c)^2}{m-1}} = 0.0813 \pm 0.0003\sqrt{\frac{11{,}610{,}000 - (813)^2}{9{,}999}} =$$

$$0.0813 \pm 0.0099.$$

Table 9.11: Decisions made by U and R_1 for the 10,000 data sets that were obtained in the computer simulation experiment on Dawn's data. The critical regions are $(U \geq 1.80)$ and $(R_1 \geq 127.0)$.

Skeptic Correct	R_1		
U	Fail to Reject H_0	Reject H_0	Total
Fail to reject H_0	9,470	7	9,477
Reject H_0	10	513	523
Total	9,480	520	$m = 10{,}000$

CTE$= 1$	R_1		
U	Fail to Reject H_0	Reject H_0	Total
Fail to reject H_0	7,620	0	7,620
Reject H_0	287	2,093	2,380
Total	7,907	2,093	10,000

9.10 Homework Problems

1. I decided to perform a power comparison of U and R_1 for Dawn's study of her cat Bob with the alternative $>$. My results are presented Table 9.11. Use this table to answer the following questions.

   (a) I was unable to achieve the same value of α for both tests. Find my two α's and comment. The test with the larger value of α will have an advantage when the power study is performed. Do you think the advantage will invalidate the power study? Explain your answer briefly.

   (b) My lone simulation experiment on power is for a constant treatment effect of 1; why didn't I use other other values besides 1? (Hint: The answer, "You are lazy," is incorrect. Well, at least in the current case.) Here is a hint: The sorted data for each treatment is below. Think about what would happen for a fixed treatment effect of 2.

Chicken:	1	3	4	5	5	6	6	6	7	8
Tuna:	0	1	1	2	3	3	3	4	5	7

2. Refer to problem 1.

   (a) What is the power for U for a constant treatment effect of 1? What is the power for R_1 for a constant treatment effect of 1? Which test is better?

   (b) Calculate the nearly certain interval for for $r_U - r_R$.

Part II

Population-Based Inference

Chapter 10

Populations: Getting Started

You have now completed Part 1 of these notes, consisting of nine chapters. What have you learned? On the one hand, you could say that you have learned many things about the discipline of Statistics. I am quite sure that you have expended a great deal of time and effort to learn, perhaps master, the material in the first nine chapters. On the other hand, however, you could say, "I have learned more than I ever wanted to know about the Skeptic's Argument **and** not much else." I **hope** that you feel differently, but I cannot say this comment is totally lacking in merit.

So, why have we spent so much time on the Skeptic's Argument? First, because the idea of Occam's Razor is very important in science. It is important to be skeptical and not just *jump on the bandwagon of the newest idea*. For data-based conclusions, we should *give the benefit of the doubt* to the notion that nothing is happening and only conclude that, indeed, *something is happening* if the data tell us that the *nothing is happening* hypothesis is inadequate. The Skeptic's Argument is, in my opinion, the purest way to introduce you to how to use Statistics in science.

The analyses you have learned in the first nine chapters require you to make **decisions**: the choice of the components of a CRD; the choice of the alternative for a test of hypotheses; for numerical data, the choice of test statistic; for a power study, the choice of an alternative of interest. The analyses require you to take an **action**: you must randomize. But, and this is the key point, the analyses **make no assumptions.** The remainder of these notes will focus on population-based inference. Assumptions **are always necessary** in order to reach a conclusion on a population-based inference. The two most basic of these assumptions involve:

1. How do the units actually studied relate to the entire population of units?

2. What structure is assumed for the population?

By the way, if either (or both) of these questions makes no sense to you that is fine. We will learn about these questions and more later in these notes.

As we will see, in population-based inference, we never (some might say rarely; I don't want to quibble about this) know **with certainty** whether our assumptions are true. Indeed, we usually know that they are not true; in this situation, we spend time investigating *how much it matters* that our assumptions are not true. (In my experience, the reason why many—certainly not all, perhaps not even most—math teachers have so much trouble teaching Statistics is because they just don't

get the idea that an assumption can be wrong. If a mathematician says, "Assume we have a triangle or a rectangle or a continuous function" and I say, "How do you know the assumption is true," the mathematician will look at me and say, "Bob, you are hopeless!")

The above discussion raises an obvious question: If population-based inference techniques rely on assumptions that are not true, why learn them? Why not limit ourselves to studies for which we can examine the Skeptic's Argument? Well, as much as I love the Skeptic's Argument, I must acknowledge its fundamental weakness: It is concerned only with the units in the study; it has no opinion on the units that are not in the study. Here is an example of what I mean.

Suppose that a balanced CRD is performed on $n = 200$ persons suffering from colon cancer. There are two competing treatments, 1 and 2, and the data give a P-value of 0.0100 for the alternative $\neq$ with the data supporting the notion that treatment 1 is better. The Skeptic's Argument is, literally, concerned only with the $n = 200$ persons in the study. The Skeptic's Argument makes no claim as to how the treatments would work on any of the thousands of people with colon cancer who are not in the study. If you are a physician caring for one of these thousands you will need to decide which treatment you recommend. The Skeptic cannot tell you what to do. By contrast, with population-based inference a P-value equal to 0.0100 allows one to conclude that overall treatment 1 is better than treatment 2 for the entire population. By making more assumptions, population-based inference obtains a stronger conclusion. The difficulty, of course, is that the assumptions of the population-based inference might not be true and, if not true, might give a misleading conclusion.

Of course, there is another difficulty in my colon cancer example. As we saw in Case 3 in Table 5.3 on page 90 in Chapter 5, even if we conclude that treatment 1 is better than treatment 2 **overall**, this does **not** imply that treatment 1 is better than treatment 2 **for every subject**; this is true for the Skeptic's argument and it's true for population-based inference.

There is, of course, a second weakness of the methods we covered in Part 1 of these notes: They require the assignment of units to study factor levels by **randomization**. For many studies in science, randomization is either impossible or, if possible, highly unethical. For an example of the former, consider any study that compares the responses given by men and women. For an example of the latter, imagine a study that assigns persons, by randomization, to the *smokes three packs of cigarettes per day* treatment. As we will discuss often in the remainder of these notes, studies with randomization yield greater scientific validity—in a carefully explained way—than studies without randomization. This does not mean, however, that studies without randomization are inherently bad or are to be avoided.

One of the greatest strengths of population-based inference is that it allows a scientist to make *predictions* about future uncertain outcomes. The Skeptic's Argument cannot be made to do this. Predictions are important in real life and they give us a real-world measure of whether the answers we get from a statistical analysis have any validity.

Anyways, I have gotten very far ahead of myself. Thus, don't worry if much of the above is confusing. By the end of these notes, these issues will make sense to you.

In the next section we will begin a long and careful development of various ideas and methods of population-based inference.

10.1 The Population Box

In Chapter 1, we learned that there are two types of units in a study: trials and subjects. When the units are subjects, often the subjects are different people. The subjects could be anything from different automobiles to different aardvarks, but in my experience, my students are more comfortable with examples that have subjects that are people. Therefore, most of my examples of units as subjects will have the subjects be people.

When you are interested in subjects, you quickly realize that there are the subjects who are included in the study—i.e., subjects from whom we collect data—as well as potential subjects who are not included in the study. A key to population-based inference is that we care about **all of these subjects; those actually studied and those not studied.** Indeed, many statisticians describe their work as primarily using data from subjects in a study to draw conclusions about all subjects. For example, I might collect data from students in my class with the goal of drawing conclusions about all students at my university. The first term we need to do this is the idea of a **finite population.** In fact, let me give you four definitions at once:

Definition 10.1 *Below are four definitions we need to get started.*

1. *A* **finite population** *is a well-defined collection of individuals of interest to the researcher. Implicit in this definition is that each individual in the population has one or more features that are of interest to the researcher.*

2. *A* **census** *consists of the researcher obtaining the values of all features of interest from all members of the finite population.*

3. *Usually, it is not possible (for reasons of cost, logistics, authority) for a researcher to obtain a census of a finite population. A* **survey** *consists of the researcher obtaining the values of all features of interest from part, but not all, of the finite population.*

4. *The* **sample** *is comprised of the members of the population that are included in the survey.*

Here is a very quick—although not very interesting—example of the above ideas.

Bert teaches at a small college with an enrollment of exactly 1,000 students. These 1,000 students form the finite population of interest to Bert. For simplicity, suppose that Bert is interested in only one dichotomous feature per student—sex—which, of course, has possible values female and male. If Bert examined student records of all 1,000 members of his population he would be conducting a census and would know how many (what proportion; what percentage) of the students are female. If Bert did not have the authority to access student records, he could choose to conduct a survey of the population. If Bert were a lazy researcher, he might sample the 20 students enrolled in his Statistics class. With this choice of survey, Bert's sample would be the 20 students in his class. Such a sample is an example of what is called a **convenience sample**; the reason behind this name is rather obvious: the subjects selected for study were convenient to the researcher.

A convenience sample is an example of a **non-probability** sample. In the example of Bert's population above, undoubtedly, there were many *chance occurrences* that led to his particular 20 students being in his class. The point is that even though the sample is the result of chance,

it is **not** the result of chance that the researcher controls or understands in a way that can lead to a mathematical model. Hence, we call it a non-probability sample. Other examples of non-probability samples include: **volunteer samples** and **judgment samples**. I will not talk about non-probability samples in these notes; if you are interested in this topic, there are references on the internet.

Statisticians and scientists are more interested in **probability samples**. As you might guess, these are sampling procedures for which probabilities can be calculated. Examples of probability samples include: **systematic random samples; stratified random samples; and (simple) random samples.** In these notes we will consider only the last of these and the closely related notion of **i.i.d. random variables**. When we study units that are trials instead of subjects, we will see that assuming we have **i.i.d. trials** is equivalent to having i.i.d. random variables. (The abbreviation i.i.d. will be explained soon.)

The ideal for any (honest) researcher is to obtain a **representative sample**. A representative sample is a sample that *exactly matches the population on all features of interest*. With more than one feature of interest, this notion becomes complicated; we won't need the complication and for simplicity I will stick to my example above with Bert and one feature of interest—sex.

Suppose that Bert's class is comprised of 12 women and eight men. In other words, his sample is 60% women. **If the population has exactly 60% women, then his sample is representative.** If the population percentage of women is any number other than 60%, then his sample is not representative.

Being representative is a strange feature, for a number of reasons. First, a researcher will never know whether the sample at hand is representative; only one with perfect knowledge of the population can determine this. Second, a really lousy way of sampling (in some situations, I think convenience samples are the worst; in other situations, volunteer samples seem to be the worst) sometimes will yield a representative sample whereas a really good way of sampling might not. This brings us to the number one reason statisticians and scientists prefer probability samples, in particular simple random samples and i.i.d. random variables:

> We can calculate the probability, b, that a probability sample will be within c of being representative.

I admit, saying that we are *within c of being representative* is quite vague; keep working through these notes and this notion will become clear. Here is my point: If b is large—remember, saying that a probability is large means it is close to one—and c is small, then we can say—before data are actually collected—that it is very likely that we will obtain a sample that is close to being representative.

We will now begin a long exploration of probability samples. The obvious starting point is to tell you what a (simple) random sample is and show you how to calculate probabilities for a random sample.

It will be convenient to visualize a finite population as consisting of a box of cards. Each member of the population has exactly one card in this box, called the **population box**, and on its card is the value of the feature of interest. If there is more than one feature of interest, then the member's values of all features are on its card, but I will restrict attention now to one feature per population member. For example, if Lisa—a female—is a student at Bert's college, then one of

the 1,000 cards in the population box corresponds to her. On Lisa's card will be the word female, perhaps coded; for example, 1 for female and 0 for male. If Lisa is also in Bert's class then her card will be in his sample.

Suppose that we have a population box with one card for each member of the population. I will show you how to calculate probabilities if 1, 2, 3, ... cards are selected at random from the box. Now, however, we must face an important practical issue. In my experience, scientists usually are interested in large populations, sometimes populations that consist of millions of members; hence, the population box will have millions of cards in it. But I don't want to introduce you to this subject with a problem like the following one.

> I want to select three cards at random from a box containing 123,000,000 cards. Help me by writing down everything that could possibly happen.

As you can see, this problem would be no fun at all!

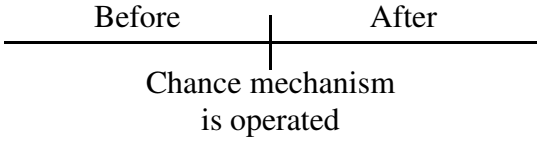
Therefore, I will introduce (several) important ideas with a population box that contains a very small number of cards. To that end, let N denote the number of cards in a population box. This means, of course, that the number of members of the population is N.

10.1.1 An Extended Example on a Very Small N

Consider a population box with $N = 5$ cards. The cards are numbered 1, 2, 3, 4 and 5, with one number per card. Consider the **chance mechanism** of selecting one card at random from this box. The expression *selecting one card at random from this box* is meant to imply that **before the chance mechanism is operated**, each card has the same likelihood of being selected.

It is necessary for me to introduce some notation. I do this with mixed feelings; as Robert DeNiro said in *Analyze This*, I am conflicted about it. Why am I conflicted? In my experience, few non-math majors say, "Wonderful! More notation!" Sadly, however, I can't figure out how to present this material without the notation below.

It is very important to think about the following time line when we talk about probabilities.



In all of our time lines, time advances from left to right. There is a point in time at which the chance mechanism is operated, yielding its outcome; in our case the identity of the selected card. To the **left** of that point is **before** the chance mechanism is operated. To the **right** of that point is **after** the chance mechanism is operated. Stating the obvious, before the chance mechanism is operated we don't know what the outcome will be; and after the chance mechanism is operated we know the outcome. It is appropriate to calculate probabilities before the chance mechanism is operated; it is not appropriate to calculate probabilities after the chance mechanism is operated. For example, once we have selected the card '3' from the box it is ridiculous to talk about the probability that the card **will be** '3' or '4' or any other number.

Define the random variable X_1 to denote the number on the card that will be selected. I say "will be" because you should think of X_1 as linked to the future; i.e., I am positioned, in time, before the chance mechanism is operated. There are five possibilities for the value of X_1: 1, 2, 3, 4 and 5. These five possibilities are equally likely to occur (which is the consequence of *selecting one card at random*), so we assign probability of $1/5 = 0.2$ to each of them, giving us the following five equations:

$$P(X_1 = 1) = 0.2, P(X_1 = 2) = 0.2, P(X_1 = 3) = 0.2, P(X_1 = 4) = 0.2, P(X_1 = 5) = 0.2.$$

We can write these equations more briefly as:

$$P(X_1 = x_1) = 0.2, \text{ for } x_1 = 1, 2, 3, 4, 5.$$

Note that, analogous to our notation for a test statistic, we use the lower case letter to denote the numerical possibilities for the upper case random variable. Either representation (a listing or the formula) of these five equations is referred to as the sampling (or probability) distribution of the random variable X_1. By the way, as you may have surmised, I put the subscript on X in anticipation of eventually sampling more than one card from the population box.

If we decide to select a random sample of size 1 from our population, then the sampling distribution of X_1 is all that we have. Obviously, a scientist will want to sample many members of a population, not just one. Well, the trip from *one* to *many* is easiest if we first visit *two*. Thus, suppose we want to have a random sample of size two from a population box. For the box of this subsection this means we select two cards at random from the cards 1, 2, 3, 4 and 5.

First I note that this problem is still manageable. With only five cards in the box, there are 10 possible samples of size two; they are:

1,2; 1,3; 1,4; 1,5; 2,3; 2,4; 2,5; 3,4; 3,5; and 4,5,

where, for example, by '2,4' I mean that the two cards selected are the cards numbered 2 and 4.

Some of you have no doubt studied probability. If so, you might remember that for many problems, a first step in the solution is to decide whether or not *order matters*. In the current problem, order does not matter. Let me be careful about this. If I reach into the box of five cards and *simultaneously* grab two cards at random, then, indeed, there is no notion of order. As we will see below, however, it is useful to reframe the notion of selecting two cards at random. Namely, it is mathematically equivalent to select one card at random, set it aside, and then select one card at random from the remaining cards. Literally, by introducing the idea of selecting the cards one-at-a-time I am introducing order into a problem in which order is not needed. I do this, as you will see below, because by making the problem apparently more difficult—by introducing order—I am, in fact, making it easier for us to study.

Henceforth, when I talk about a random sample I will refer to the first card selected and the second card selected and so on. I have previously defined X_1 to be the number on the first card selected. Not surprisingly, I define X_2 to be the number on the second card selected. My immediate goal is to show you how to calculate probabilities for the pair (X_1, X_2). Please refer to Table 10.1. The first feature of Table 10.1 to note is that it consists of three Tables: A, B and C. Five rows

Table 10.1: Three displays for the possible outcomes when selecting two cards at random, **without replacement,** from a box containing cards 1, 2, 3, 4 and 5.

Table A: All possible pairs of values on the two cards:

	X_2				
X_1	1	2	3	4	5
1	—	(1,2)	(1,3)	(1,4)	(1,5)
2	(2,1)	—	(2,3)	(2,4)	(2,5)
3	(3,1)	(3,2)	—	(3,4)	(3,5)
4	(4,1)	(4,2)	(4,3)	—	(4,5)
5	(5,1)	(5,2)	(5,3)	(5,4)	—

Table B: Joint probabilities for the values on the two cards:

	X_2				
X_1	1	2	3	4	5
1	0.00	0.05	0.05	0.05	0.05
2	0.05	0.00	0.05	0.05	0.05
3	0.05	0.05	0.00	0.05	0.05
4	0.05	0.05	0.05	0.00	0.05
5	0.05	0.05	0.05	0.05	0.00

Table C: Table B with marginal probabilities added:

	X_2					
X_1	1	2	3	4	5	Total
1	0.00	0.05	0.05	0.05	0.05	0.20
2	0.05	0.00	0.05	0.05	0.05	0.20
3	0.05	0.05	0.00	0.05	0.05	0.20
4	0.05	0.05	0.05	0.00	0.05	0.20
5	0.05	0.05	0.05	0.05	0.00	0.20
Total	0.20	0.20	0.20	0.20	0.20	1.00

[columns] of each of these three tables denote the five possible values for X_1 [X_2]. Five of the $5 \times 5 = 25$ cells in Table A are marked with '—,' denoting that they are impossible; if you select two cards at random from the box then you must obtain two different cards. In my experience people forget this feature of a random sample. **Thus, henceforth, I will sometimes refer to a random sample as selecting cards at random from the box without replacement**. Thus, for example, if the first card selected is '4' then the second card is selected at random from the remaining cards: 1, 2, 3 and 5.

Staying with Table A, the remaining 20 entries (excluding the '—' ones) correspond to the 20 possible outcomes. These are written as pairs, for example (3,5), the members of which denote the value of X_1 and then the value of X_2. Thus, for example, the pair (3,5) means that card 3 is selected first and card 5 is selected second. This might seem curious to you. The pairs (5,3) and (3,5) correspond to the same random sample, which is listed twice in each of the three tables in Table 10.1. This seems like extra work: our table has 20 possible cells for the 10 possible samples, with each sample appearing twice. **It is extra work**, but as we will see shortly, it will help us develop the material.

The idea of selecting a random sample of size two; or, equivalently, selecting two cards at random without replacement; or, equivalently, selecting one card at random, setting it aside and then selecting one card at random from the remaining cards; all of these ideas imply that the 20 **possible cells**—excluding the five impossible cells on the main diagonal—in Table A are equally likely to occur and, hence, each cell has probability $1/20 = 0.05$. (You can see why I selected a box with five cards; I like simple, short, nonrepeating decimals for my probabilities.) Table B presents the probabilities written within each of the 25 cells. Note that each of the five impossible cells has probability 0 and that each of the twenty possible cells has probability 0.05.

Finally, Table C supplements Table B by summing the probabilities across the rows and down the columns. The resulting probabilities are written in the margins (right and bottom) of the table; hence, they often are referred to as *marginal probabilities*. If we look at the entries in the extreme left and extreme right columns, we find the familiar sampling distribution for X_1:

$$P(X_1 = x_1) = 0.20, \text{ for } x_1 = 1, 2, 3, 4, 5.$$

If we look at the uppermost and lowermost rows, we find the sampling distribution for X_2:

$$P(X_2 = x_2) = 0.20, \text{ for } x_2 = 1, 2, 3, 4, 5.$$

Note that X_1 and X_2 have the same distributions: they both have possible values 1, 2, 3, 4 and 5, and their possible values are equally likely. The technical term for this is we say that X_1 and X_2 are **identically distributed**, abbreviated i.d. (Two-thirds of the initials in i.i.d.; one-half of the ideas, as we soon will learn.)

The 20 non-zero probabilities in the cells give us the **joint sampling distribution** of X_1 and X_2. We have the adjective *joint* to remind us that these probabilities are concerned with how X_1 and X_2 behave **together**. To avoid possible confusion, the distributions of either X_1 or X_2 alone are sometimes called their marginal sampling distributions. *Marginal* because they appear in the margins of our table above **and** because they are for a single random variable, ignoring the other.

There is another way of sampling, other than random sampling without replacement, that will be very important to us. I mentioned above that for a random sample of size two we may select the two cards at once *or* select the cards one-at-a-time, without replacement. The obvious question is: May we sample cards one-at-a-time **with replacement**? The obvious answer: Of course we may, we live in a free society! A more interesting question is: What happens if we select cards at random with replacement?

Before I turn to a computation of probabilities, I want you to develop some *feel* for what we are doing. First, I have good news for those of you who do not currently possess thousands of cards and a box to hold them. Our population box is simply an instructional device; a way to visualize the process of selecting a sample from a population. As a practicing statistician I always use an electronic device to select my sample, be it with or without replacement. In particular, recall the website:

`http://www.randomizer.org/form.htm`

that we introduced in Chapter 3 for the purpose of obtaining an assignment for a CRD. In Section 10.5 you will learn how this website can be used to select a sample from a population at random, either with or without replacement.

I used the website to obtain 10 random samples of size two, *with replacement*, from the box of this section. Below are the 10 samples I obtained:

Sample Number:	1	2	3	4	5	6	7	8	9	10
Sample Obtained:	1,1	3,2	1,4	5,1	2,1	4,5	4,3	2,3	4,4	3,5

In the above listing I am reporting the cards in the order in which they were selected. Thus, my second sample—3,2— consists of the same cards as my eight sample—2,3. Two samples consist of the same card being selected twice: the first—1,1—and the ninth—4,4.

Researchers *do not* select cards from a box because it is a *fun activity*; they do it to investigate what is in the box. For the purpose of learning, it is clearly a waste to sample the same card more than once. *Sampling with replacement* makes such a waste possible, while *sampling without replacement* makes such a waste impossible. For this reason, I sometimes refer to *sampling with replacement* as the **dumb** way to sample and *sampling without replacement* as the **smart** way to sample. The former is dumb because it is a waste (that is, dumb) to allow for the possibility of sampling the same card more than once. The latter is smart, well, because it isn't dumb!

Now I am going to do something strange, although perhaps—you be the judge—not out of character. I am going to give you several reasons why the dumb method of sampling is not such a bad thing.

First, as my 10 samples above suggest, sampling with replacement is potentially wasteful, not necessarily wasteful. Eight out of the 10 samples select two different cards; thus, they —speaking both practically and mathematically—provide the same information as would be obtained by sampling the smart way.

Second, selecting the same card more than once, while wasteful of effort, does not actually bias our results in any way. This fact is not obvious, but you will see why I say this later.

Additional reasons that dumb sampling can be good, beyond these two, will appear soon.

Table 10.2: Two displays for the possible outcomes when selecting two cards at random, **with replacement**, from a box containing cards 1, 2, 3, 4 and 5.

Table A: All possible pairs of values on the two cards:

	X_2				
X_1	1	2	3	4	5
1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)
2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)
3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)
4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)
5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)

Table B: Joint and marginal probabilities for the values on the two cards:

	X_2					
X_1	1	2	3	4	5	Total
1	0.04	0.04	0.04	0.04	0.04	0.20
2	0.04	0.04	0.04	0.04	0.04	0.20
3	0.04	0.04	0.04	0.04	0.04	0.20
4	0.04	0.04	0.04	0.04	0.04	0.20
5	0.04	0.04	0.04	0.04	0.04	0.20
Total	0.20	0.20	0.20	0.20	0.20	1.00

Table 10.2 addresses the issue of finding probabilities for X_1 and X_2 for selecting cards at random with replacement, the dumb way of sampling. As with our earlier table (Table 10.1), the current table is comprised of other tables, in this case two: Tables A and B. Table A presents the $5 \times 5 = 25$ possible outcomes from selecting two cards at random with replacement from our box. All 25 outcomes are equally likely to occur; thus, they all have the same probability: $1/25 = 0.04$, as presented in Table B. Table B also presents the marginal probabilities for both random variables.

The first thing to note about Table 10.2 is that, just as with the smart way of sampling, X_1 and X_2 have identical sampling distributions and, indeed, the same sampling distributions they had for the smart way of sampling. The difference between the smart and dumb methods of sampling appears in the joint distribution of X_1 and X_2.

Often we will be interested computing a probability that looks like:

$$P(X_1 = 3 \text{ and } X_2 = 5).$$

It is very tedious to write *and* inside a probability statement; thus, we adopt the following shorthand notation. We will write, for example,

$$P(X_1 = 3 \text{ and } X_2 = 5) \text{ as } P(X_1 = 3, X_2 = 5).$$

In words, a *comma* inside a probability statement is read as *and*.

The next thing to note is incredibly important. Look at the 25 joint probabilities in Table B of Table 10.2. **Every one** of the joint probabilities has the property that it is equal to the product of its row and column (marginals) probabilities. In particular, for every cell:

$$0.04 = 0.20 \times 0.20.$$

A similar equality is never true for Table C in Table 10.1. The product of the margins is again $0.20 \times 0.20 = 0.04$ which **never** appears as a joint probability. This observation leads us to the following definition.

Definition 10.2 (Two Independent Random Variables.) *Suppose that we have two random variables, denoted by X and Y. These random variables are said to be* **independent** *if, and only if, the following equation is true for all numbers x and y that are possible values of X and Y, respectively.*

$$P(X = x, Y = y) = P(X = x)P(Y = y). \tag{10.1}$$

Note: the restriction that x and y must be possible values of X and Y is not really needed, though some people find it comforting. It is not needed because if, say, $x = 2.5$ is not a possible value of X, then both sides of Equation 10.1 are 0 and, hence, equal.

In words, Equation 10.1 tells us that for independent random variables, the word *and* tells us to multiply. Hence, it is often referred to as the **multiplication rule for independent random variables**.

Let me carefully summarize what we have learned **for the box of this section**. If we select $n = 2$ cards at random:

- Without replacement—also called a (simple) random sample; also called (by me) the smart way to sample—then X_1 and X_2 are identically distributed, but are not independent.

- With replacement—also called (by me) the dumb way to sample—then X_1 and X_2 are independent as well as identically distributed.

We have spent a great deal of effort studying a very small and particular problem. This endeavor would be a waste of your time **if it weren't for the fact that the above results generalize in a huge way!** I will go through the important generalizations now. I won't prove these, although I will sometimes give an illustration. If you were working on a degree in Statistics, then we should spend more time on these matters, but you aren't, so we won't.

Still with two cards selected, the multiplication rule can be extended as follows. Let A_1 [A_2] be any event defined in terms of X_1 [X_2]. Then the probability that both A_1 and A_2 occur equals the product of their (individual or marginal) probabilities of occurring. For example, suppose that A_1 is the event that $X_1 \geq 3$ and suppose that A_2 is the event that X_2 is an even number (either 2 or 4). We can draw a picture of both of these events occurring:

	X_2				
X_1	1	2	3	4	5
1					
2					
3		X		X	
4		X		X	
5		X		X	

In the above display, the six cells marked with 'X' are the six cells for which both A_1 and A_2 occur. Each of these cells has probability of occurring equal to 0.04. Summing these we find that the probability that both A_1 and A_2 will occur is equal to $6(0.04) = 0.24$. Individually, $P(A_1) = 0.60$ and $P(A_2) = 0.40$. The product of these individual probabilities does, indeed, equal 0.24, the probability that both occur.

Here is our next generalization. The results about independence and identical distributions are true for any box, not just our favorite box with cards 1, 2, 3, 4 and 5.

Here is our next generalization. The results about independence and identical distributions are true for any number of cards selected at random, not just for two. For completeness, I will state these results below.

Result 10.1 (A summary of results on smart and dumb random sampling.) *For any population box, define the random variables*

$$X_1, X_2, X_3, \ldots, X_n,$$

as above. Namely, X_1 is the number on the first card selected; X_2 is the number on the second card selected; and so on. The following results are true.

1. *For both methods of sampling cards at random—smart and dumb—the random variables*

   $$X_1, X_2, X_3, \ldots, X_n$$

   are identically distributed. The common distribution is the same for dumb and smart sampling; moreover—because it does not depend on the method of random sampling—the common distribution is sometimes called the population probability distribution.

2. *For the dumb way of sampling, the random variables*

   $$X_1, X_2, X_3, \ldots, X_n$$

   are independent; for the smart way of sampling they are not independent, also called dependent.

I am afraid that I have made this material seem more difficult than necessary. Let me end this section with a brief example that, perhaps, will help.

I plan to select two cards at random from a population box with $N = 10$ cards. Six of the cards are marked '1' and four are marked '0.' Clearly,

$$P(X_1 = 0) = 0.40 \text{ and } P(X_1 = 1) = 0.60,$$

is the population distribution. I want to compute two probabilities:

$$P(X_1 = 1, X_2 = 1) \text{ and } P(X_1 = 0, X_2 = 1)$$

For the dumb method of random sampling, we use the multiplication rule and obtain:

$$P(X_1 = 1, X_2 = 1) = P(X_1 = 1)P(X_2 = 1) = 0.6(0.6) = 0.36 \text{ and}$$

$$P(X_1 = 0, X_2 = 1) = P(X_1 = 0)P(X_2 = 1) = 0.4(0.6) = 0.24.$$

These answers are **incorrect** for the smart way of sampling.

The correct answers for smart sampling, however, can be found using another version of the multiplication rule, which is called the **multiplication rule for dependent random variables**. For the smart way of sampling, I write $P(X_1 = 1, X_2 = 1)$ as

$$P(X_1 = 1)P(X_2 = 1|X_1 = 1),$$

where, the vertical line segment within a probability statement is short for *given that*. In particular, when I write

$$P(X_2 = 1|X_1 = 1),$$

I mean *the probability that the second card selected will be a 1, given that the first card selected is a 1*. Given this particular information, the box available when the second card is selected contains five cards marked '1' and four cards marked '0.' Thus,

$$P(X_1 = 1)P(X_2 = 1|X_1 = 1) = (6/10)(5/9) = 0.333, \text{ and, similarly,}$$

$$P(X_1 = 0, X_2 = 1) = P(X_1 = 0)P(X_2 = 1|(X_1 = 0) = (4/10)(6/9) = 0.267.$$

Thus, the **great thing about independence** is not that we have a multiplication rule, but rather that the things we multiply don't change based on what happened earlier!

10.2 Horseshoes . . . Meaning of Probability

I conjecture that most of you have heard the expression, *close counts in horseshoes and hand grenades*. In my experience, this is presented as a humorous statement, even though there is nothing funny about being close to a hand grenade! I will occasionally expand this homily in these notes; now I expand it to

> Close counts in horseshoes, hand grenades and probabilities.

I will explain what this means.

Consider a population box with 1,000 cards, numbered serially, 1, 2, 3, . . . , 1,000. This is an obvious generalization of our earlier box with $N = 5$ cards to a box with $N = 1{,}000$ cards. Next, suppose that we plan to select two cards at random from this box, either the dumb or the smart way. I am interested in calculating marginal and joint probabilities. Obviously, *actually drawing*

a table with 1,000 rows, 1,000 columns and 1,000,000 cells is not realistic. But because we are clever we can analyze this situation without drawing such huge tables.

Both marginal distributions are that each of the numbers 1, 2, 3, ..., 1,000, has probability 0.001 of occurring. With independence (dumb sampling) the probability of each of the one million cells is the product of its margins:

$$0.001 \times 0.001 = 0.000001.$$

With the smart method of sampling, the 1,000 cells on the main diagonal, where the row number equals the column number, are impossible and the remaining 1,000,000 − 1,000 = 999,000 cells are equally likely to occur. Thus, the probability of each of these cells is:

$$\frac{1}{999{,}000} = 0.000001001,$$

rounded off to the nearest billionth. In other words, the joint probabilities are very close to the product of the marginal probabilities for all one million cells. Thus, with two cards selected from this box of 1,000 cards, **if one is primarily interested in calculating probabilities** then it does not matter (approximately) whether one samples the smart way or the dumb way.

The above fact about our very specific box generalizes to all boxes, as follows. Let N denote the number of cards in the box. Let n denote the number of cards that will be selected at random—dumb or smart is open at this point—from the box. Let A be any event that is a function of some or all of the n cards selected. Let $P(A|\text{dumb})$ [$P(A|\text{smart})$] denote the probability that A will occur given the dumb [smart] way of sampling. We have the following result:

Provided that the ratio n/N is small,

$$P(A|\text{dumb}) \approx P(A|\text{smart}). \tag{10.2}$$

The above is, of course, a qualitative result: If the ratio is small, then the two probabilities are close. What do the words *small* and *close* signify? (It's a bit like the following statement, which is true and qualitative: If you stand really far away from us, I look like Brad Pitt. More accurately, if Brad and I are standing next to each other and you are very far away from us, you won't be able to tell who is who.) As we will see repeatedly in these notes, *close* is always tricky, so people focus on the *small*.

A popular general guideline is that if $n/N \leq 0.05$ (many people use 0.10 instead of 0.05 for the threshold to *smallness*) then the approximation is good. It's actually a bit funny that people argue about whether the threshold should be 0.05 or 0.10 or some other number. Here is why. I am typing this draft of Chapter 10 on October 4, 2012. It seems as if every day I read about a new poll concerned with who will win the presidential election in Wisconsin. I don't know how many people will vote next month, but in 2008, nearly 3 million people voted for president in Wisconsin. Even 1% (much smaller than either popular threshold) of 3 million is $n = 30{,}000$. I am quite certain that the polls I read about have sample sizes smaller than 30,000. In other words, in most surveys that I see in daily life, the ratio n/N is much smaller than 0.05.

Here is an important practical consequence of the above. Whether we sample the smart or the dumb way, when we calculate probabilities we may pretend that we sampled the dumb way because it makes computations easier. Our computations will be exactly correct if we sampled the dumb way and approximately correct if we sampled the smart way and n/N is small. Actually, as we shall see later with several examples, the biggest problem in sampling is **not** whether n/N is "small enough;" it is: What are the consequences of a sample that is **not** obtained by selecting cards at random from a box?

As I stated earlier, whenever we select cards from a box at random, with replacement (the dumb way), we end up with what we call independent random variables. Since each selection can be viewed as a trial (as we introduced these in Chapter 1) we sometimes say that we have **i.i.d. trials**. With the help of i.i.d. random variables (trials) I can now give an interpretation to probability.

10.2.1 The Law of Large Numbers

The level of mathematics in this subsection is much higher than anywhere else in these notes and, indeed, is higher than the prerequisite for taking this course. Therefore, please do not worry if you cannot follow all of the steps presented below.

I will give you a specific example and then state the result in somewhat general terms. Our result is called **the Law of Large Numbers** or **the long-run-relative-frequency interpretation of probability**.

Let's revisit my box with $N = 5$ cards numbered 1, 2, 3, 4 and 5. I plan to select n cards at random, with replacement, from this box, where n is going to be a really large number. Suppose that my favorite number is 5 and I will be really happy every time I draw the card marked '5' from the box. I define X_1, X_2, X_3, ... as before (i.e., X_i is the number on the card selected on draw number i, for all i). I know that the X_i's are identically distributed and that $P(X_1 = 5)$ is equal to 0.20. The question I pose is: How exactly should we interpret the statement: "The probability of selecting '5' is 0.20?"

Define $f_n(5)$ to be the frequency (f is for frequency) of occurrence of 5 in the first n draws. The Law of Large Numbers states that the limit, as n tends to infinity, of

$$\frac{f_n(5)}{n} \text{ is } 0.20.$$

Let me say a few words about this limiting result. First, if you have never studied calculus, the mathematical idea of limits can be strange and confusing. If you have studied calculus you might remember what has always seemed to me to be the simplest example of a limit:

$$\text{The limit as } n \text{ tends to infinity of } (1/n) = 0.$$

Let me make a couple of comments about this limiting result. First, n does not, literally, become infinity, nor does $(1/n)$ literally become zero. The real meaning (and usefulness) of the above limiting result is that it means that for n *really large* the value of $1/n$ becomes *really close* to 0. As a result, whenever n is *really large* it is a good approximation to say that $1/n$ is 0. This is such

a simple example because we can make precise the connection between n being really large and $1/n$ being really close to 0. For example, if n exceeds one billion, then $1/n$ is less than 1 divided by one billion and its distance from the limiting value, 0, is at most one in one billion. By contrast, in many applications of calculus the relationship between being *really large* and *really close* is not so easy to see. We won't be concerned with this issue.

In probability theory, limits—by necessity—have an *extra layer* of complexity. In particular, look at my limiting result above:

$$\text{The limit as } n \text{ tends to infinity of } \frac{f_n(5)}{n} = 0.20.$$

The object of our limiting, $f_n(5)/n$ is much more complicated than the object in my calculus example, $1/n$, *because* $f_n(5)$ is a random variable. For example, if $n = 1000$ we know that $1/n = 0.001$ but we don't know the value of $f_n(5)$; conceivably, it could be any integer value between 0 and 1000, inclusive. As a result, the Law of Large Numbers is, indeed, a very complicated math result. Here is what it means: For *any* specified (small) value of closeness and *any* specified (large, i.e., close to 1) value of probability, eventually for n large enough, the value of $f_n(5)/n$ will be within the specified closeness to 0.20 with probability equal to our greater than the specified target. This last sentence is quite complicated! Here is a concrete example.

I will specify closeness to being within 0.001 of 0.20. I specify my large probability to be 0.9999. Whereas we can never be **certain** about what the value of $f_n(5)/n$ will turn out to be, the Law of Large Numbers tells me that for n sufficiently large, the event

$$0.199 \leq f_n(5)/n \leq 0.201,$$

has probability of occurring of 0.9999 or more. How large must n be? We will not address this issue directly in these notes. (After we learn about confidence intervals, the interested reader will be able to investigate this issue, but the topic is not important in this course; it's more of a topic for a course in probability theory.) I will remark that the Law of Large Numbers is responsible for the thousands of gambling casinos in the world being profitable. (See my roulette example later in this chapter.)

I will now give a general version of the Law of Large Numbers. Here are the ingredients we need:

- We need a sequence of i.i.d. random variables X_1, X_2, X_3, ...
- We need a sequence of events A_1, A_2, A_3, ..., with the following properties:
  1. Whether or not the event A_i occurs depends only on the value of X_i, for all values of i.
  2. $P(A_i) = p$ is the same number for all values of i.

  For our use, the A_i's will all be the 'same' event. By this I mean they will be something like out example above where A_i was that $(X_i = 5)$.
- Define $f_n(A_i)$ to be the frequency of occurrence of the events A_i in opportunities $i = 1, 2, \ldots, n$.

The Law of Large Numbers states:

$$\text{The limit as } n \text{ tends to infinity of } \frac{f_n(A_i)}{n} = p.$$

The above presentation of the Law of Large Numbers is much more complicated (mathematically) than anything else in these notes. I made the above presentation in the spirit of intellectual honesty. **Here is what you really need to know about the Law of Large Numbers**. The probability of an event is equal to its long-run-relative-frequency of occurrence under the assumption that we have i.i.d. operations of the chance mechanism. As a result, if we have a **large number** of i.i.d. operations of a chance mechanism, then the relative frequency of occurrence of the event is approximately equal to its probability:

$$\text{Relative frequency of } A \text{ in } n \text{ trials } \approx P(A). \tag{10.3}$$

This approximation is actually **twice as exciting** as most people realize! I say this because it can be used in **two very different situations**.

1. If the numerical value of $P(A)$ is **known**, then **before we observe the trials** we can **accurately predict** the value of the relative frequency of occurrence of the event A for a large number of trials.

2. If the numerical value of $P(A)$ is **unknown**, then we **cannot predict**, in advance, the relative frequency of occurrence of A. We can, however, do the following. We can go ahead and **perform** or **observe** a large number of trials and then calculate the observed relative frequency of occurrence of A. This number is a reasonable approximation to the unknown $P(A)$.

10.3 Independent and Identically Distributed Trials

Chapter 1 introduced the idea of a unit, the entity from which we obtain a response. I said that sometimes a unit is a trial and sometimes it is a subject. Earlier in this chapter I introduced you to the population box as a model for a finite population of subjects. In this section I will argue that *sometimes* a box of cards can be used as part of a model for the outcomes of trials. In this chapter we will consider trials for which the response is either:

- a category, usually with two possible values; i.e., a dichotomy; or
- a count, for example, as in the example of Dawn's study of Bob's preferences for treats.

Trials with responses that are measurements (examples: time to run one mile; time to complete an ergometer workout; distance a hit golf ball travels) present special difficulties and will be handled later in these notes.

In my experience, in the current context of *populations*, students find trials to be conceptually more difficult than subjects. As a result, I am going to introduce this topic to you slowly with an extended familiar (I hope) example.

Beginning in my early childhood and extending well into my adulthood, I have played games that involved the throwing of one or more dice:

- Monopoly, Parchesi, Yahtzee, Skunk, and Risk, to name a few.

In these notes, unless otherwise stated, a die will be a cube with the numbers 1, 2, 3, 4, 5 and 6 on its faces, one number per face. The arrangement of the numbers on the faces follows a standard pattern (for example, opposite faces sum to 7), but we won't be interested in such features. If you want to learn about dice that possess some number of faces other than six, see the internet.

Suppose that I have a particular die that interests me. Define the chance mechanism to be a single cast of the die. The possible outcomes of this cast are the numbers: 1, 2, 3, 4, 5 and 6. The first issue I face is my answer to the question:

> Am I willing to assume that the six outcomes are equally likely to occur? Or, in the vernacular, is the die balanced and fair? (Not in the sense of Fox News.)

As we will see later in these notes, there have been dice in my life for which I am willing to assume balance, but there also have been dice in my life for which I am **not willing to assume balance**. For now, in order to proceed, let's assume that my answer is, "Yes, I am willing to assume that my die is balanced."

Now consider a box containing $N = 6$ cards numbered 1, 2, 3, 4, 5 and 6. Next, consider the chance mechanism of selecting one of these cards at random. Clearly, in terms of probabilities, selecting one card from this box is equivalent to one cast of a balanced die. What about repeated casts of a balanced die?

I argue that repeated casts of a balanced die is equivalent to repeatedly sampling of cards at random with replacement—the dumb method—from the above box. Why? At each draw (cast) the six possible outcomes are equally likely. Also, the result of any draw (cast) cannot *possibly influence* the outcome of some other draw (cast).

To be more precise, define X_i to be the number obtained on cast i of the die. The random variables X_1, X_2, X_3, ..., are i.i.d. random variables; as such, the Law of Large Numbers is true. Thus, for example, in the long run, the relative frequency of each of the six possible outcomes of a cast will equal one-sixth.

Here is another example. This example helps explain a claim I made earlier about why casinos do so well financially.

An American roulette wheel has 38 slots, each slot with a number and a color. For this example, I will focus on the color. Two slots are colored green, 18 are red and 18 are black. Red is a popular bet and the casino pays 'even money' to a winner.

If we assume that the 38 slots are equally likely to occur (i.e., that the wheel is fair), then the probability that a red bet wins is $18/38 = 0.4737$. But a gambler is primarily concerned with his/her relative frequency of winning. Suppose that the trials are independent—i.e., the wheel has no memory—and that a gambler places a very large number, n, of one dollar bets on red. By the Law of Large Numbers, the relative frequency of winning bets will be very close to 0.4737 and the relative frequency of losing bets will be very close to $1 - 0.4737 = 0.5263$. In simpler terms, in the long run, for every \$100 bet on red, the casino pays out $2(47.37) = 94.74$ dollars, for a net profit of \$5.26 for every \$100 bet.

As a side note, when a person goes to a casino, he/she can see that every table game has a range of allowable bets. For example, there might be a roulette wheel that states that the minimum

bet allowed is $1 and the maximum is $500. Well, a regular person likely pays no attention to the maximum, but it is very important to the casino. As a silly and extreme example, suppose Bill Gates or Warren Buffett or one of the Koch brothers walks into a casino and wants to place a $1 billion bet on red. No casino could/would accept the bet. (Why?) And, of course, I have no evidence that any of these men would want to place such a bet.

10.3.1 An Application to Genetics

A man with type AB blood and a woman with type AB blood will have a child. What will be the blood type of the child? This question cannot be answered with certainty; there are three possible blood types for the child: A, B and AB. These three types, however, are not equally likely to occur, as I will now argue. According to Mendelian inheritance (see the internet for more information) both the father and mother donate an allele to the child, with each parent donating either an A or a B, as displayed below:

The Child's Bloodtype:

Allele from Dad:	Allele from Mom: A	B
A	A	AB
B	AB	B

If we make three assumptions:

- The allele from Dad is equally likely to be A or B;
- The allele from Mom is equally likely to be A or B; and
- Mom's contribution is independent of Dad's contribution;

then the four cells above are equally likely to occur and, in the long run, the blood types A, AB and B will occur in the ratio 1:2:1.

If you have studied biology in the last 10 years your knowledge of Mendelian inheritance is, no doubt, greater than mine. The above ratio, 1:2:1, arises for traits other than human blood type. Other ratios that arise in Mendelian inheritance include: 1:1; 3:1; and 9:3:3:1. See the internet or a modern textbook on biology for more information.

10.3.2 Matryoshka (Matrushka) Dolls, Onions and Probabilities

Please excuse my uncertainty in spelling. Users of the English language seem to have difficulty making conversions from the Cyrillic alphabet. For example, the czar and the tsar were the same man. (Based on my two years of studying Russian, tsar is correct. Others may disagree with me.)

Anyways, a matryoshka doll is also called a Russian nesting doll. I conjecture that most of you have seen them or, at the very least, seen pictures of them. As you know, frequently in these notes I have referred you to *the internet* and, sometimes, more specifically, to Wikipedia. This, of

course, is risky because there is no guarantee that Wikipedia is, or will remain, accurate. I always overcome my nature to be lazy and actually check Wikipedia before typing my suggestion to visit the site. Imagine my happiness (delight is too strong) when I went to the matryoshka doll entry on Wikipedia and found exactly what I wanted to find:

> Matryoshkas are also used metaphorically, as a design paradigm, known as the *matryoshka principle* or *nested doll principle*. It denotes a recognizable relationship of *object-within-similar-object* that appears in the design of many other natural and man-made objects. ...
>
> The *onion metaphor* is of similar character. If the outer layer is peeled off an onion, a similar onion exists within. This structure is employed by designers in applications such as the layering of clothes or the design of tables, where a smaller table sits within a larger table and a yet smaller one within that.

My goal in this subsection is for you to realize that **many** (two or more) operations of a chance mechanism can be viewed as **one** operation of some different chance mechanism. A simple enough idea, but one that will be of great utility to us in these notes. Let me begin with a simple example. The following description is taken from Wikipedia.

The game of craps involves the simultaneous casting of two dice. It is easier to study if we imagine the dice being tossed one-at-a-time or somehow being distinguishable from each other. Each round of a game begins with the *come-out*. Three things can happen as a result of the come-out:

- An immediate *pass line win* if the dice total 7 or 11.
- An immediate *pass line loss* if the dice total 2, 3 or 12 (called *craps* or *crapping out*).
- No immediate win or loss, but the establishment of a *point* if the dice total 4, 5, 6, 8, 9 or 10.

My goal is to determine the probability of each of these three possible outcomes: win, loss and point.

I determine these probabilities by consider **two operations of the chance mechanism of i.i.d. casts of a balanced die**. To this end, I create the following table:

	X_2					
X_1	1	2	3	4	5	6
1	Loss	Loss	Point	Point	Point	Win
2	Loss	Point	Point	Point	Win	Point
3	Point	Point	Point	Win	Point	Point
4	Point	Point	Win	Point	Point	Point
5	Point	Win	Point	Point	Point	Win
6	Win	Point	Point	Point	Win	Loss

Based on my assumptions, the 36 cells in this table are equally likely to occur. Thus, by counting, I obtain the following probabilities:

$$P(\text{Win}) = 8/36 = 2/9; P(\text{Loss}) = 4/36 = 1/9; P(\text{Point}) = 24/36 = 2/3.$$

Now, define a new box containing $N = 9$ cards of which two are marked 'Win;' one is marked 'Loss;' and the remaining six are marked 'Point.' The chance mechanism is selecting one card at random from this new box. Clearly, selecting one card at random from this new box is equivalent to my two operations of the *balanced die box*. Admittedly, this example is easier because I don't care what the point is. If you are playing craps, you would prefer a point of 6 or 8 to a point of 4 or 10; i.e., not all points have the same consequences. I won't show you the details, but if you are interested in the point obtained, the probabilities become:

$$P(\text{Win}) = 8/36 = 2/9; P(\text{Loss}) = 4/36 = 1/9; P(\text{Point} = 4) = 3/36 = 1/12;$$

$$P(\text{Point} = 5) = 4/36 = 1/9; P(\text{Point} = 6) = 5/36; P(\text{Point} = 8) = 5/36;$$

$$P(\text{Point} = 9) = 4/36 = 1/9; \text{ and } P(\text{Point} = 10) = 3/36 = 1/12,$$

10.3.3 In Praise of Dumb Sampling

Suppose that I have a population of $N = 40$ college students and I want explore the question of how many of them are female. In addition, my resources allow me to select $n = 20$ students at random for my sample. Clearly, the smart way of sampling—which guarantees information from 20 different population members—is greatly preferred to the dumb way of sampling—which likely, it seems, will lead to my obtaining information from fewer than 20 distinct population members.

If I replace my population size of 40 by 40,000 and keep my sample size at 20, then, as argued earlier, the distinction between dumb and smart sampling becomes negligible. But even so, smart seems better; after all, it would be embarrassing to ask a student twice to state his/her sex. In this subsection I will show you a situation in which dumb sampling is actually much better than smart sampling. Indeed, we have made use of this fact many times in these notes.

Let's return to Dawn's study of Bob's eating habits, introduced in Chapter 1. In order to perform her study, Dawn needed an assignment: she needed to select 10 cards at random without replacement from a box containing 20 cards. Similar to my example on craps above, we could view Dawn's selection of 10 cards from her box as equivalent to selecting **one card** from a different box. Which different box? The box that contains all 184,756 possible assignments. Of course, for Dawn's purpose of performing her study, it was easier to create a box with 20 cards and then select 10 of them. (If the randomizer website had existed when she performed her study, it would have been easier still for Dawn to use it and not bother with locating a box and 20 cards.)

Let's now turn our attention to analyzing Dawn's data. Following our methodology, we wanted to know the sampling distribution of the test statistic, be it the difference of means, U, or the sum of ranks, R_1. To this end we prefer the box with 184,756 cards, one for each assignment. On each card is written the value of the test statistic of interest.

I did not give you the details (computer code) of our computer simulation experiment, but now I need to tell you a little bit about it. Each rep of the simulation consists of the computer selecting an assignment at random from the population of 184,756 possible assignments and recording the value of the test statistic for the selected assignment. In the language of this and the previous chapter, the program selected assignments at random **with replacement**; i.e., the program sampled in the dumb way. Why did I write a program that samples the dumb way?

The answer lies in our imagining what would be necessary to sample the smart way. As we will see very soon, the smart way would be a programming nightmare! It is important to remember/realize that, in reality, there is no box with 184,756 cards in it. If there were, we could pull out a card, look at it, and set it aside. This would make the smart way of sampling easy and, consequently, preferred to the dumb way. **But there is no such box of assignments!** Here is how the program operates. I tell the computer to select 10 trials (days) from the 20 trials (days) of Dawn's study. This selection tells the computer which responses are on treatment 1 and which are on treatment 2 and then the observed value of the test statistic is calculated. Here is the key point: There is no way to tell the computer, "Don't pick the same assignment again." (If you disagree, I challenge you to write the code; if you succeed, send it to me.) What I *could* tell the computer is,

> Write down the assignment you just used in file A. Every time you select a new assignment, before using it check to see whether it is in file A. If it is, don't use it; if not; use it and add it to file A.

I don't mean to be rude, but this would be one of the worst programs ever written! As we approach 10,000 reps, computer space would be wasted on storing file A and a huge amount of computer time would be spent checking the 'new' assignment against the ones previously used. Thus, we sample the dumb way.

Recall that in Part 1, I advocated using the relative frequencies from a computer simulation experiment as approximations to the unknown exact probabilities. According to the Law of Large Numbers, this is good advice; for a large number of reps, we can say that with large probability, the relative frequencies are **close** to the exact probabilities of interest. Indeed, our nearly certain interval did this. By nearly certain, I conveyed a large probability, which, as we will see, is approximately 99.74%. The nearly certain interval allowed us to compute **how close**, namely within

$$3\sqrt{\frac{\hat{r}(1-\hat{r})}{m}},$$

where $\hat{r}$ is our relative frequency approximation and m is our number of reps in the simulation experiment.

10.4 Some Practical Issues

We have learned about finite populations and two ways to sample from them: a random sample without replacement (smart method of sampling) and a random sample with replacement (dumb method of sampling). We have learned that sometimes I am willing to assume I have i.i.d. trials; my two examples of this were: repeated casts of a balanced die and the alleles contributed from two parents to child. With either i.i.d. trials or the dumb method of sampling, we have i.i.d. random variables.

All of the population-based methods presented in the remainder of these notes assume that we have some type of i.i.d. random variables. The same can be said of every introductory Statistics textbook that I have ever seen. Now I am going to surprise you. These various texts pay little or

no attention to the practical issues that result from such an assumption. They just repeatedly state the mantra: *Assume we have i.i.d. random variables* (or, in some books, *assume we have a random sample*). But let me be clear. I am not criticizing *teachers* of introductory Statistics; I suspect that many of them discuss this issue. Publishers want to keep textbooks short—to increase the profit margin, no doubt—and it takes time to present this material in lecture. Our medium—an online course—seems ideal for exploring this topic. I can make these notes longer without increasing any *production costs*—as my time is free—and I don't need to devote lecture time to this, because we have no lectures!

In any event, let me move away from these general comments and talk specific issues.

Let me begin with trials. Think of the studies of Part I that had units equal to trials. For example, let's consider Dawn's study of her cat Bob. Dawn concluded that Bob's consumption of chicken treats exceeded his consumption of tuna treats. Now let's consider the time right after Dawn concluded her analysis. Suppose that she decided to concentrate on chicken treats because she interpreted Bob's greater consumption of chicken as reflecting his preference for its taste. (One could argue that Bob ate more chicken because he required more of them to be satisfied, but I don't want to go down that road at this time.)

Thus, Dawn decided to offer Bob ten chicken treats every day for a large number of days, denoted by n. This gives rise to n random variables:

$$X_1, X_2, X_3, \ldots, X_n,$$

which correspond, naturally, to the number of treats he eats each day. Are these i.i.d. trials? Who knows? A better question is: Are we willing **to assume** that these are i.i.d. trials? According to Wikipedia, in 1966, psychologist Abraham Maslow wrote:

> I suppose it is tempting, if the only tool you have is a hammer, to treat everything as if it were a nail.

Maslow's Hammer is sometimes shortened to:

> If the only tool you have is a hammer, then everything looks like a nail.

Following Maslow, if one knows how to analyze data **only** under the assumption of i.i.d. trials, then one is very likely to make such an assumption.

Indeed, in my career as a statistician I have frequently begun an analysis by saying that I assume that I have i.i.d. trials. I comfort myself with the following facts.

- I state explicitly that my answers are dependent on my assumption; if my assumption is wrong, my answers might be misleading.
- I do a *mind experiment*. I think about the science involved and consider whether the assumption of i.i.d. trials seems reasonable. This is what I did for the die example: Does it make sense to think that a die remembers? Does it make sense to think that a die will change over time? Because my answer to both of these questions is *no*, my mind experiment tells me that I have i.i.d. trials.

- After collecting data, it is possible to critically examine the assumption of i.i.d. trials. This topic is explored briefly later in these notes.

I admit that the above list is unsatisfactory; the description of the mind experiment is particularly vague. Rather than try to expand on these vague notions, in the examples that follow in these *Course Notes* I will discuss the assumption of i.i.d. trials on a study-by-study basis.

For finite populations, all the methods we will learn in these *Course Notes* assume that the members of the sample have been selected at random from the population box, either with (dumb) or without (smart) replacement. In my experience, when the population is a well-defined collection of people, this assumption of random sampling is rarely, almost never, literally true. I will expand on this statement by telling you a number of stories based on my work as a statistician.

10.4.1 The Issue of Nonresponse

Many years ago I was helping a pharmacy professor analyze some of her data. I asked her if she had any other projects on which I might help. She replied, "No, my next project involves taking a census. Thus, I won't need a statistician." A few months later she called and asked for my help on this *census project*; what had gone wrong?

She had attended a national conference of **all**—numbering 1,000—members of some section of a professional association. She felt that 1,000 was a sufficiently small population size N to allow her to perform a census. She gave each of the 1,000 members a questionnaire to complete. A difficulty arose when only 300 questionnaires were returned to her. Her next plan was to treat her sample of 300 as a random sample from her population and was asking my advice because her ratio of n/N was $300/1000 = 0.30$ which is larger than the threshold of 0.05. I pointed out that she had a volunteer sample, not a random sample. Without making this story too long, I suggested that her best plan of action was to contact, say, 70 people—I can't remember the exact number—who had not respond and to politely and persistently—if needed—encourage them to respond. After obtaining the data from these 70, she could do various statistical analyses to see whether the responses from the original non-responders (she now has a smart random sample of 70 of these) were importantly different from the responses from the original 300 responders.

In my experience, expecting to complete a census is hopelessly naive. A better plan for the professor would have been to select, say, 200 people at random and understand that the 140 or so who chose not to respond would need to be tracked down and encouraged, to the extent possible and reasonable, to participate. If in the end, say, 20% of the original 200 absolutely refused to participate, then the analysis should highlight this fact, for example, by writing,

> Here is what we can say about the approximately 80% of the population that is willing to be surveyed. For the other 20% we have no idea what they think. Except, of course, that they didn't want to participate in my survey!

10.4.2 Drinking and Driving in Wisconsin

In 1981, the legislature in the State of Wisconsin enacted a comprehensive law that sharply increased the penalties for drinking and driving. Part of the law directed the State's Department of

Transportation (DOT) to:

- Educate the public about the new law and the dangers of driving after drinking.
- Measure the effectiveness of its educational programs as well as the drivers' attitudes and behavior concerning drinking and driving.

Eventually, the Wisconsin DOT conducted four large surveys of the population of licensed drivers in 1982, 1983, 1984 and 1986. After the data had been collected in 1983, DOT researchers contacted me for help in analyzing their data. I continued to work with the DOT and eventually analyzed all four surveys and submitted written reports on my findings.

Let me begin by describing how the DOT selected the samples for its surveys. (The same method was used each year.) First, the people at the DOT had the good sense **not** to attempt to obtain a random sample of licensed drivers. To see why I say this, let's imagine the steps involved in obtaining a random sample.

Even in the ancient days of 1982, I understand that the DOT had a computer file with the names of all one million licensed drivers in Wisconsin. (I will say one million drivers; I don't recall the actual number, but one million seems reasonable and, I suspect, a bit too small.) It would be easy enough to randomly select (smart method; they didn't want to question anyone twice), say, 1,500 names from the computer file and survey those selected. Two immediate difficulties come to mind:

1. How to **contact** the 1,500 members of the sample. Send them a questionnaire through the mail? Contact them by telephone? Have a researcher visit the home? All of these methods would be very time-consuming and expensive and all would suffer from the following difficulty.
2. What should be done about the drivers who choose **not to respond?** Any solution would be time-consuming and expensive and, in the end, in our society you cannot **force** people to respond to a survey.

Instead of selecting drivers at random, the DOT hit upon the following plan. I was **not** involved in selecting this plan, but I believe that it was a good plan.

First, **judgment** was used to select a variety of *driver's license exam stations* around the state. The goal was to obtain a mix of stations that reflect the rural/urban mix of Wisconsin as well as other demographic features of interest. (I can't remember what other features they considered.) Each selected station was sent a batch of questionnaires and told that over a specified period in late March and early April, every person who applied for a license renewal or a new license **was required** to complete the questionnaire and submit it before being served. Despite this rather draconian requirement, I understand that no complaints were reported and nobody left a station to avoid responding. (I guess that people in the 1980s really wanted their driver's licenses!)

The completed questionnaires—1,589 in 1982 and 1,072 in 1983—were sent to Madison and I was given the task of analyzing them. One of my main directives was to search for changes from the 1982 survey—conducted **before** the law took effect later in 1982—to the 1983 survey. Later in these notes I will report some of my findings; at this time, I am more concerned with the method of sampling that was used.

Let me state the obvious. Many of you are not very interested in these surveys that were conducted some 30 years ago. That is fair. Therefore, I will **not** use these data extensively, but rather I will use them primarily when they illustrate some *timeless difficulty* with survey research. In addition, because I was intimately involved in this research, I have the *inside story* on what happened. In my experience, it is difficult to convince a researcher to *share all* about the conduct of any research study.

Let me reiterate an important point. None of the driver surveys conducted by the Wisconsin DOT consisted of a *random sample* of drivers. Thus, a research purist could say, "Don't use any of the population-based inference methods on these data." I don't mean to be blunt, but if you totally agree with the purist's argument, then you have no future in research, unless you can carve out a niche as a professional contrarian! The purist ignores the sage comment by Voltaire:

> The perfect is the enemy of the good.

Admittedly, the DOT's samples were not random—hence, not perfect to a statistician—but were they good? I see the following advantages to the DOT's method (these were alluded to above):

1. A large amount of data was obtained at a very low cost of collection.

2. The issue of nonresponse was minor.

Regarding nonresponse, yes everyone did complete and submit a questionnaire, but some of the items on the questionnaire were ignored by some respondents. My recollection was that all or nearly all respondents took the activity seriously; i.e., I spent a great deal of time *looking at* the raw data and I don't recall any completely blank questionnaires. Rather, roughly 12% [6%] chose not to report their age [sex]; otherwise, subjects would occasionally leave an item blank, presumably because they were not sure about their knowledge, behavior or opinion. I reported the nonresponse rate on each questionnaire item, allowing the reader to make his or her own assessment of its importance.

The DOT's sampling method is an example of a **convenience sample**; it was convenient for the DOT to survey people who visited one of the stations selected for the study. A small change in procedure would have changed this convenience sample into a **volunteer sample**; can you think of what this change is? Well, there are several possible changes, but here is the one I am contemplating: Instead of *forcing* everyone who visits to complete a questionnaire, the questionnaires could have been placed on a table with a sign, "Please complete this questionnaire; there are no consequences for participating or not."

In the scenario described above, I think that the actual convenience sample is superior to my proposed volunteer sample, but we don't have time to spend on this topic. There is a more important point to consider.

As the analyst of the DOT data, I decided to make the **WTP** assumption, which I will now state.

Definition 10.3 (The Willing to Pretend (WTP) Assumption.) *Consider a survey for which the sample was selected in any manner other than a (smart or dumb) random sample. The WTP assumption means that, for the purpose of analyzing and interpreting the data, the data are assumed*

to be the result of selecting a random sample. In other words, a person who makes the WTP assumption is **willing to pretend** *that the data came from a random sample.*

In my experience the WTP assumption often is made only tacitly. For example—and this was not my finest hour—in my 1983 report for the DOT, I explained how the data were collected and then, without comment, proceeded to use various analysis methods that are based on the assumption of a random sample. In retrospect, I would feel better if I had explicitly stated my adoption of the WTP assumption. In my defense—a variation on the *all my friends have a later curfew* argument that you might know—it is common for researchers to suppress any mention of the WTP assumption.

As I hope is obvious, I cannot make general pronouncements about the validity of the WTP assumption, other than saying that the purist never makes the assumption and, alas, some researchers appear to always make the assumption.

For the DOT surveys, I cannot imagine any reason why people who visit a DOT station in late March to early April are different—in terms of attitudes, knowledge or behavior related to drinking and driving—than people who visit at other times of the year. Also, I cannot imagine any reason why people who visit the stations selected for study are different than those who visit other stations. I might, of course, be totally mistaken in my beliefs; thus, feel free to disagree. I believe in the principle that I should—to the extent possible—make all of my assumptions explicit for two reasons:

1. In my experience the act of making my assumptions explicit often has led to my making an important discovery about what is actually reasonable to believe; and

2. If I want other people to consider my work seriously, I should pay them the respect of being honest and explicit about my methods.

In the next subsection I will give additional examples of when I am willing to make the WTP assumption and when I am not.

10.4.3 Presidents and Birthdays

In 1968, I was 19 years-old and was not allowed to vote for President because the age requirement at that time was 21. In 1972, I voted in my first presidential election and have voted in every one of the subsequent ten presidential elections. In 1972, I was standing in line to vote, in a cold rain in Ann Arbor, Michigan. Next to me in line was Liv, a good friend of my wife. I commented on how miserable the weather was. Liv replied that she agreed, but it would be worth it once George McGovern was elected President. I was dumbfounded. "You don't really believe that McGovern will win, do you?"

"Of course I do," she replied, "Everyone I know is voting for him."

In the language of this chapter, Liv was willing to pretend that her circle of acquaintances could be viewed as a random sample from the population of voters. The fact that she would not have worded her process in this way, does not make it any less true.

Obviously, I remember this conversation well, even though it occurred 40 years ago. I remember it because I have seen variations on it many times over the years. The WTP assumption I held

and hold on the drivers' surveys, rarely holds for a *a sample* that consists of the following groups of people:

- family;
- friends and acquaintances;
- co-workers; or
- students in a class.

Regarding the last item in this list. Students in my classes: belong to a very narrow and young age group; are hard-workers; don't have much money; are smart; are highly educated; and so on. Several of these features are likely to be associated with many responses of interest to a researcher.

This brings us to birthdays. As I will argue below, I am willing to pretend that for the response of birth date, the students in my class can be viewed as a random sample from a population that is described below.

One of the most famous results in elementary probability theory is the **birthday problem**, also called the **birthday paradox**. I don't want to spend much time on it; interested readers are encouraged to use the internet or contact me. Among its probability calculations, the birthday problem shows that in a room with $n = 23$ [$n = 80$] persons, there is a 50.6% [99.99%] chance that at least two people will have the same date of birth—month and day; year is ignored by the birthday problem. This result is sometimes called a *paradox* because it seems surprising that with *only* 23 people, at least one match is more likely than no matches and that with only 80 people, at least one match is virtually certain.

As so often occurs in practice, a probability result is presented as an **iron-clad fact** when, indeed, it is based on assumptions. All probabilities are based on assumptions and the assumptions might be true, almost true or outrageously false. Let me describe the assumptions underlying the answers of 50.6% and 99.99% in the birthday problem.

Definition 10.4 (The Assumptions Needed for the Birthday Problem.) *There is population consisting of a large number of people. In the population box, a person's card contains the date of the person's birthday. All 366 days of the year (don't forget February 29) are equally represented in the box; i.e., if one card is selected from the box at random, then all 366 possible dates are equally likely to be selected. Twenty-three [Eighty] persons are selected at random, without replacement, from the population box.*

Let's look at the main ideas in these assumptions, including what happens if any assumption fails to be met.

1. We require the smart method of sampling because, frankly, it would not be noteworthy if we selected, say, Bert twice and found that his *birthdays* matched! The math argument, however, is much easier if we were to sample the dumb way. We get around this difficulty by assuming that the population box contains a large number of cards.

For example, suppose that the population consisted of 366,000 people with each of the 366 dates being the birthday of exactly 1,000 people. As I discuss earlier in this chapter, imagine that the 23—or 80 or, indeed, any number of—persons in the sample are selected one-at-a-time. By assumption, all 366 dates are equally likely to be the birthday of the first person selected. The exact probability that the second date selected matches the first date selected is:

$$999/365{,}999 = 1/366, \text{ approximately.}$$

This is the exact probability because after removing the first date selected, there are 999 cards (out of the 365,999 remaining cards) in the box that match the first date selected.

Thus, by having a large number of cards in the box—as well as being interested in a relatively small number of selections, 23 or 80—we can use the ideas of i.i.d. trials to simplify probability calculations, even though the smart method of sampling is used.

2. We assume that all 366 days are equally represented in the population box. This assumption is obviously false because February 29 occurs, at most, once every four years. (A little known fact and of even less importance to birthdays during our era: the years 1900, 2100, 2200, and others—xy00 where xy is not divisible by four—are **not leap years**.) This difficultly has been handled two ways:

   - Ignore the existence of February 29 and replace the 366 days in the computation by 365. The result is that the probability of at least one match for $n = 23$ increases from 50.6% to 50.7%; not even worthy of notice!
   - Use public health records of relative frequencies of dates of birth instead of the assumption of equally likely. If you do this, you find that 365 of the relative frequencies are very similar and one is a lot smaller. Using these numbers the probability of at least one match becomes *slightly larger than 50.6%* for $n = 23$ and even slightly larger than 99.99% for $n = 80$, but not enough to get excited.

3. I have been saving the big assumption for the last; namely, that the 23 or 80 persons are selected at random from the population box. First, it has **never** been the case that students are in my class because they were randomly selected from some population and forced to be in my class, although they are, in some ways, forced to take my class. A purist would say, "Bob, don't ever use the birthday problem in your class." I don't particularly mind this admonition because, frankly, the purist **never does anything, except solve problems in textbooks**. I am, however, very willing to adopt the WTP assumption. Indeed, I can't imagine any reason for date of birth being associated with taking my class.

Regarding the last assumption, for years I would survey my class to determine their dates of birth. The result of the birthday problem worked remarkably well; with samples of 80 students or more, I never failed to find at least one match on birthdays. When I subdivided my class into subgroups of size 23, about one-half of the groups had at least one match and about one-half of the groups had no match. (Sorry, I don't have exact data.)

I would challenge my class to think of a situation in which the WTP assumption would be unreasonable. They always quickly suggested the following three possibilities:

1. Twenty-three persons at a particular Madison bar that is well-known for giving free drinks to customers on their birthday.

2. Twenty-three persons in line to renew their driver's licenses. (In Wisconsin, licenses expire on one's birthday; hence, a person is likely to be in line because of the proximity of the current date with his/her birthday.)

   And, finally, the best possibility:

3. Twenty-three (brand new) persons in the new babies section of a hospital!

For each of these three possibilities, I am convinced that if one collected such data repeatedly, the relative frequency of occurrence of at least one match would be much larger than the 50.6% of the birthday problem. Indeed, for the last possibility, I would be amazed if the relative frequency was smaller than one!

10.5 Computing

This chapter is mostly about *ideas* and very little about computing answers. The main computing tool is the use of the randomizer website:

`http://www.randomizer.org/form.htm`

to generate a random sample of cards selected from a population box. I will illustrate the use of this website for a random sample (either smart or dumb) of size $n = 2$ from the box with $N = 5$ cards, numbered 1, 2, 3, 4 and 5. Homework Problem 3 will illustrate ways to use this site for a number of topics covered in this chapter.

Recall that in order to use the randomizer website, the user must respond to seven prompts. Below are the responses—in bold-face—we need for the above problem and the **smart**—without replacement—method of sampling. Note that I am asking the site to report the *order* of selection; sometimes, of course, we don't care about the order. Also, I am asking the site to give me six—my response to the first prompt—simulated samples of size $n = 2$.

6; **2**; From **1** To **5**; **Yes**; **No**; and **Place Markers Off**.

After specifying my options, I clicked on *Randomize Now!* and obtained the following output:

1,3; 1,2; 1,2; 3,5; 2,3; and 5,4.

Next, I repeated the above to obtain six simulated **dumb** samples of size $n = 2$. Only one of my responses—the fifth one—to the prompts was changed to shift from the smart to the dumb method. For completeness, my responses were:

6; **2**; From **1** To **5**; **No**; **No**; and **Place Markers Off**.

After specifying my options, I clicked on *Randomize Now!* and obtained the following output:

3,4; 3,5; 5,5; 1,4; 2,1; and 4,3.

Note that only once—the third sample—did the dumb method of sampling result in the same card being selected twice. As we saw in the notes, the probability of a repeat is 0.20; thus, a relative frequency of one out of six is hardly surprising.

10.6 Summary

In Part I of these notes you learned a great deal about the Skeptic's Argument. While it is obvious that I am a big fan of the Skeptic's Argument, I do acknowledge its main limitation: it is concerned only with the units in the study. It many studies, the researcher wants to generalize the findings beyond the units actually studied. Statisticians invent **populations** as the main instrument for generalizations. It is important to begin with a careful discussion of populations.

We begin with populations for subjects. The subjects could be automobiles or aardvarks, but in most of our examples, I will take subjects to be people or, sometimes, a family or a married couple or some other well-defined collection of people. As you will see a number of times in these notes, in population-based inference it is important to carefully define our subjects.

Whenever units are subjects, the finite population is a well-defined collection of all subjects of interest to the researcher. To lessen confusion, we say that the finite population is comprised of a finite number of **members**. The members from whom information is obtained are called the subjects in the study. It is convenient to visualize each member of the population having a card in the **population** box.

For a finite population, the goal of the researcher is to look at some of the cards in the population box and infer features of the population. These inferences will involve uncertainty; thus, we want to be able to use the discipline of probability to quantify the uncertainty. To this end, the researcher needs to assume a **probability sample** has been selected from the population, as opposed to a non-probability sample such as a judgment, convenience or volunteer sample. To this end, we study two types of probability samples:

- Selecting cards from the population box at random, **without replacement**—referred to as the **smart** random sample.
- Selecting cards from the population box at random, **with replacement**—referred to as the **dumb** random sample.

For a random sample of size n—smart or dumb; I will explicitly mention it whenever my results are true for one of these, but not the other—define n random variables, as follows:

- X_1 is the number on the first card selected;
- X_2 is the number on the second card selected;
- X_3 is the number on the third card selected; and so on until we get to
- X_n is the number on the n^{th} (last) card selected.

Following our earlier work with test statistics, the observed value of any of these random variables is denoted by the same letter and subscript, but lower case; e.g., x_3 is the observed value of X_3 and more generally x_i is the observed value of X_i, for any i that makes sense (i.e., any positive integer i for which $i \leq n$). Section 10.1.1 presents an extended example of computing probabilities for these random samples. The highlights of our findings are:

1. For both methods of random sampling, the random variables

   $$X_1, X_2, X_3, \ldots X_n,$$

   all have the same sampling/probability distribution; we say they are **identically distributed**, abbreviated **i.d.** Thus, for example $P(X_1 = 5) = P(X_3 = 5)$, and so on. This common distribution is called the population distribution and is the same regardless of whether the sampling is smart or dumb.

2. For the *dumb* method of random sampling, the random variables

   $$X_1, X_2, X_3, \ldots X_n,$$

   are statistically **independent**. This means we can use the multiplication rule; for example,

   $$P(X_1 = 5, X_2 = 7, X_3 = 9) = P(X_1 = 5)P(X_2 = 7)P(X_3 = 9).$$

   Thus, for the dumb method of sampling, the random variables

   $$X_1, X_2, X_3, \ldots X_n,$$

   are independent and identically distributed, abbreviated **i.i.d.**.

3. There is also a multiplication rule for the *smart* method of random sampling, but it is messier than the one above. For example, suppose we want to select three cards from a box with $N = 5$ cards, numbered 1, 2, 3, 4 and 5. Suppose further that I am interested in the event: $(X_1 = 3, X_2 = 2, X_3 = 5)$. Using the multiplication rule for conditional probabilities, I obtain:

   $$P(X_1 = 3, X_2 = 2, X_3 = 5) = P(X_1 = 3)P(X_2 = 2|X_1 = 3)P(X_3 = 5|X_1 = 3, X_2 = 2).$$

   This equation is intimidating in appearance, but quite easy to use. Indeed, you may use it, as I now describe, without thinking about how horrible it looks. We begin with

   $$P(X_1 = 3) = 1/5 = 0.20.$$

   So far, this is easy. Next, we tackle

   $$P(X_2 = 2|X_1 = 3).$$

   Given that the first card selected is the '3;' the remaining cards are 1, 2, 4 and 5. Thus,

   $$P(X_2 = 2|X_1 = 3) = 1/4 = 0.25.$$

Finally, given that the first two cards selected are '3' followed by '2,'

$$P(X_3 = 5|X_1 = 3, X_2 = 2) = 1/3 = 0.33.$$

Thus, we find

$$P(X_1 = 3, X_2 = 2, X_3 = 5) = (1/5)(1/4)(1/3) = (1/60) = 0.0167.$$

4. As illustrated in the previous two items in this list, it is much easier to calculate probabilities if we have independence. Thus, it often happens that a researcher samples the smart way, but computes probabilities **as if** the sample had been collected the dumb way. This is not cheating; it is an approximation. If the population size is N—known or unknown to the researcher—and the value of n/N is 0.05 or smaller, then the approximation is good.

Next, we consider the situation in which the units are trials. For example, suppose that each trial consists of my casting a die. It makes no sense to represent this activity as a finite population; for example, it makes no sense to say that I could cast a die N = 9,342 times, but not 9,343 times. As a result, for trials, some probabilists say that we have an infinite population. Sadly, this terminology can be confusing for the non-probabilist; because I am **not** an immortal, there is *some* limit to the number of times I could cast a die. It's just that there turns out to be no point in trying to specify that limit.

In fact, for a trial our attention is focused on **the process that generates the outcomes of the trials**. In particular, there are two main questions:

1. Is the process stable over time or does it change?

2. Is the process such that the particular outcome(s) of some trial(s) influence the outcome(s) of some different trial(s).

If we are willing to assume that the process is stable over time and there is no influence then we can model the process as being the same as dumb random sampling from a box. Because dumb sampling gives us i.i.d. random variables, we refer to this situation as having i.i.d. trials.

When we have i.i.d. random variables or trials, we have the Law of Large Numbers (LLN). The Law of Large Numbers gives us a qualitative link between the probability of an event and its long-run-relative-frequency of occurrence. In later chapters we will see how to make the Law of Large Numbers more quantitative.

After a number of topics: an application to Mendelian inheritance; the role of matryoshka dolls; and a homage to dumb sampling; Section 10.4 presents some important practical issues. This section does not settle the issue of how to deal with practical issues; rather, its ideas will be revisited throughout the remainder of the *Course Notes*. In particular, the *Willing to Pretend* (WTP) assumption, Definition 10.3, will be discussed many times.

10.7 Practice Problems

1. Consider a random sample without replacement (i.e., smart sampling) of size $n = 2$ from a population box with $N = 5$ cards, numbered 1, 2, 3, 5 and 5. Note that, unlike our example earlier in this chapter, two of the cards have the same response value. In order to calculate probabilities, it is helpful to pretend that we can distinguish between the two 5's in the box. To this end, I will represent one of the 5's by 5_a and the other by 5_b. With this set-up, we can immediately rewrite Table B in Table 10.1 as below:

	X_2					
X_1	1	2	3	5_a	5_b	Total
1	—	0.05	0.05	0.05	0.05	0.20
2	0.05	—	0.05	0.05	0.05	0.20
3	0.05	0.05	—	0.05	0.05	0.20
5_a	0.05	0.05	0.05	—	0.05	0.20
5_b	0.05	0.05	0.05	0.05	—	0.20
Total	0.20	0.20	0.20	0.20	0.20	1.00

Next, we combine the two rows [columns] corresponding to our two versions of 5 to obtain the following joint probability distribution for the two cards selected at random, without replacement, from our box.

	X_2				
X_1	1	2	3	5	Total
1	—	0.05	0.05	0.10	0.20
2	0.05	—	0.05	0.10	0.20
3	0.05	0.05	—	0.10	0.20
5	0.10	0.10	0.10	0.10	0.40
Total	0.20	0.20	0.20	0.40	1.00

(a) Calculate

$$P(X_1 \text{ is an odd number and } X_2 < 3),$$

and compare it to:

$$P(X_1 \text{ is an odd number})P(X_2 < 3).$$

(b) Define Y to equal the maximum of X_1 and X_2. Determine the sampling distribution of Y.

2. Refer to the previous problem. An alternative to creating a table to present the joint distribution of X_1 and X_2 is to use the multiplication rule for dependent random variables. For example,

$$P(X_1 = 1, X_2 = 5) = 0.10$$

from the table in problem 1. Alternatively,

$$P(X_1 = 1, X_2 = 5) = P(X_1 = 1)P(X_2 = 5|X_1 = 1) = (1/5)(2/4) = 0.10.$$

Use the multiplication rule for dependent random variables to calculate the following.

(a) $P(X_1 = 5, X_2 = 5)$.

(b) $P(X_1 = 2, X_2 = 3)$.

3. With the help of Minitab I performed the following simulation 10 times:

   Simulate 100,000 i.i.d. trials with $P(X_1 = 1) = 0.50$ and $P(X_1 = 0) = 0.50$.

   Let T_i denote the sum of the 100,000 numbers obtained on simulation i, for $i = 1, 2, 3, \ldots, 10$; T_i can also be interpreted as the number of trials in simulation i that yielded the value 1. My observed value of T_1 is $t_1 =$ 50,080. For all ten simulations, the sum of the observed values of the T_i's equals 500,372.

   Walt states, "The Law of Large Numbersstates that t_1 should be close to 50,000. It's not; it misses 50,000 by 80. **Worse yet**, for all 1,000,000 trials, the sum of the t_i should be close to 500,000. It's not and it misses 500,000 by 372 which is worse than it did for 100,000 trials! Explain why Walt is wrong.

4. Suppose that I am interested in the population of all married couples in Wisconsin. that have exactly two children. (I know; units other than *married couples* can have babies; but this is my pretend study, so I will choose the terms of it. Truly, there is no implied disrespect—or respect, for that matter—towards any of the myriad of other units that can and do have babies.) Let X denote the number of female children in a family chosen at random from this population. Possible values for X are, of course, 0, 1 and 2. I want to know the probability distribution for X.

   It is unlikely that I can find a listing of my population, all married couples in Wisconsin with exactly two children. In part, this is because babies have a way of just showing up; thus, a list of all married couples with exactly two children on any given date, will be inaccurate a few months later.

   Let's suppose, instead, that I have access to a listing of all married couples in Wisconsin. If resources were no issue, I would take a random sample of, say, 1,000 population members and ask each two questions:

   (a) As of today, right now, do you have exactly two children? If your answer is yes, please answer question (b); if your answer is no, you are finished.

   (b) How many of your two children are girls?

   This is a common technique. Take a random sample from a population that includes your population of interest and then disregard all subjects that are not in your population of interest. This is legitimate, but you won't know your sample size in advance. For example, above

all I know for sure is that my sample size of interest will be 1,000 or smaller; possibly a lot smaller.

My guess is that the best I could do in practice is to obtain a sample—not random—from the population of married couples and feel ok about making the WTP assumption, Definition 10.3.

I apologize for the lengthy narrative. I have included it for two reasons.

(a) To give you more exposure to my thought process when I plan a survey.

(b) To convince you that it really is a lot of work to learn about the composition of families of married couples with two children.

Please excuse a bit of a digression; it is important. I watched every episode of the television series *House*. If you are not familiar with the show, Dr. House is a brilliant diagnostician. Frequently, however, in addition to his vast store of medical knowledge, House must refer to his oft stated belief that "Everybody lies," in order to solve a particularly difficult medical problem. He does not believe, of course, that everybody lies all the time or even that everybody deserves the pejorative of liar; he simply believes that, on occasion, people lie and a diagnostician must take this into account.

So, why the digression to one of my favorite television shows? To transition to my belief about researchers: *Everybody is lazy.* I urge you to remember this in your roles as both consumer and creator of research results. As a consumer, so that you will possess a reasonable skepticism. Always saying, "You can prove anything with statistics," does **not** exhibit a reasonable skepticism. You need *reasons* for being skeptical; otherwise, you simply are exhibiting another form of laziness. As a researcher, so that you will not waste effort on flawed studies and not mislead the public.

I have seen many textbooks that claim that it is **easy** to determine the probability distribution of X, the number of female children in families of married couples with exactly two children. Their reasoning is quite simple. They refer to my table on blood types on page 233. Relabel what I call the Dad's [Mom's] allele as the sex of the first [second] child. Each child is equally likely to be female or male and the sexes of the two children are independent. While these assumptions might not be exactly true—identical twins will violate independence—they seem close enough to obtain reasonable answers. From this point of view, we get:

$$P(Y = 0) = 0.25, P(Y = 1) = 0.50 \text{ and } P(Y = 0) = 0.25.$$

I have actually seen the following statement in many texts:

> Of all families—marriage is really not an issue here— with two children, 25% have no girls, 25% have no boys and 50% have one boy and one girl.

Before you continue reading, stop and think about the above answer. Do you see anything wrong with it?

Immediately after giving the above 25%/50%/25% answer as **the probability distribution for all families with two children**, one textbook then had the following example, which I will call the **Jess model**.

> Suppose that in a city every married couple behaves as follows: They keep having children until they have a girl and then they stop. What is the distribution of the number of children per family?

Well, in this new scenario **every** couple that stops with exactly two children will have one boy (born first) and one girl. **Not the 50% that just moments earlier had been proclaimed to be the answer!** We also see a new difficulty, that perhaps you have noticed already. Selecting a couple that **currently** has two children is not the same as selecting a couple that **eventually has a total of two children**. This is a general difficulty, which we will return to later in these notes when we learn the difference between cross-sectional and longitudinal studies. Thus, even if the Jess model was true in a city—and it is, of course, ridiculous to assume that every couple has the same reproduction strategy—then with a cross-sectional study—which is what I describe above—in addition to sampling couples that have their one boy and one girl and have stopped reproducing, we would no doubt get quite a few families with two boys that are waiting for their next baby.

Thus, in conclusion, the 25%/50%/25% answer is wrong because it assumes that the choice of the number of children is unrelated to the sexes of the children. This assumption might be true, but in my experience, I don't believe it is even close to being true. **We should not build a probabilistic model because we are too lazy to collect data!**

10.8 Solutions to Practice Problems

1. (a) First, I identify the cells—by V's below—that satisfy the event of interest:

$$(X_1 \text{ is an odd number and } X_2 < 3).$$

	X_2			
X_1	1	2	3	5
1	V	V		
2				
3	V	V		
5	V	V		

To obtain our answer, we sum the probabilities of the six cells V-ed above:

$$0.00 + 0.05 + 0.05 + 0.05 + 0.10 + 0.10 = 0.35.$$

Looking at the margins, I find:

$$P(X_1 \text{ is an odd number}) = 0.80, P(X_2 < 3) = 0.40 \text{ and } 0.80(0.40) = 0.32.$$

(b) By inspecting the joint distribution table, we see that the possible values of Y are: 2, 3 and 5. There is no *easy* way to get the answer, we simply must plow through the table's information. The event $(Y = 2)$ will occur if the sample is (1,2) or (2,1). Thus,

$$P(Y = 2) = 0.05 + 0.05 = 0.10.$$

Similarly, the event $(Y = 3)$ is comprised of the samples (1,3), (2,3), (3,1) and (3,2). Thus,

$$P(Y = 3) = 4(0.05) = 0.20.$$

Finally, the total probability is 1; thus,

$$1 = P(Y = 2)+P(Y = 3)+P(Y = 5) = 0.10+0.20+P(Y = 5) = 0.30+P(Y = 5).$$

Thus, $P(Y = 5) = 1 - 0.30 = 0.70$.

2. (a) Write

$$P(X_1 = 5, X_2 = 5) = P(X_1 = 5)P(X_2 = 5|X_1 = 5) = (2/5)(1/4) = 0.10.$$

(b) Write

$$P(X_1 = 2, X_2 = 3) = P(X_1 = 2)P(X_2 = 3|X_1 = 2) = (1/5)(1/4) = 0.05.$$

3. For 100,000 trials the Law of Large Numbers states that $T_1/100{,}000$ will, with high probability, be close to 0.500000. Well, $t_1/100{,}000$ equals 0.500800. For 1,000,000 trials, the Law of Large Numbers states that the total of the T_i's divided by 1,000,000 will, with high probability, be close to 0.500000. For my simulations, the sum of the t_i's divided by 1,000,000 is 0.500372. Note also that 0.500372 is closer than 0.500800 to one-half by more than a factor of two. Thus, Walt's *worse yet* comment is a misinterpretation of the Law of Large Numbers. **Remember: The Law of Large Numbers is about the relative frequency, not the frequency!**

10.9 Homework Problems

1. Refer to Practice Problem 1. Consider random sample without replacement (i.e., smart sampling) of size $n = 2$ from a population box with $N = 5$ cards, numbered 2, 2, 3, 4 and 4.

   (a) Determine the correct probabilities for the table below.

	X_2			
X_1	2	3	4	Total
2				
3				
4				
Total				1.00

   (b) Use your answer from (a) to compute the following probability:

$$P(X_1 \text{ is an even number and } X_2 \geq 3).$$

   (c) Use your answer from (a) to compute the following probability:

$$P(X_1 \text{ is an odd number or } X_2 = 2).$$

   Recall that in math, or means and/or.

2. Consider random sample with replacement (i.e., dumb sampling) of size $n = 2$ from a population box with $N = 10$ cards, numbered 1, 2, 2, 3, 3, 3, 4, 4, 4 and 4.

   (a) Determine the correct probabilities for the table below.

	X_2				
X_1	1	2	3	4	Total
1					
2					
3					
4					
Total					1.00

   (b) Use your answer from (a) to compute the following probability:

$$P(X_1 \geq 3 \text{ and } X_2 < 4).$$

   (c) Use your answer from (a) to compute the following probability:

$$P(X_1 = 4 \text{ or } X_2 \leq 3).$$

   Recall that in math, or means and/or.

3. I have a population box with $n = 100$ cards, numbered 1, 2, ..., 100. A twist on this problem is that cards numbered 1, 2, ..., 60 are females and cards numbered 61, 62, ..., 100 are males. With the help of our website randomizer, I select 10 smart random samples, each with $n = 5$. The samples I obtained are listed below:

Sample	Cards Selected	Sample	Cards Selected
1:	6, 30, 31, 48, 70	2:	3, 21, 28, 37, 48
3:	15, 39, 52, 91, 95	4:	11, 34, 36, 56, 86
5:	71, 72, 76, 83, 84	6:	29, 37, 42, 75, 93
7:	27, 30, 34, 53, 89	8:	20, 44, 61, 72, 83
9:	13, 24, 65, 85, 99	10:	6, 21, 28, 66, 88

(a) What seven choices did I make on the randomizer website?

(b) Which sample(s) yielded zero females? One female? Two females? Three females? Four females? Five females?

(c) In regards to the feature sex, which samples are representative of the population?

Chapter 11

Bernoulli Trials

11.1 The Binomial Distribution

In the previous chapter, we learned about i.i.d. trials. Recall that there are three ways we can have i.i.d. trials:

1. Our units are trials and we have decided to assume that they are i.i.d.

2. We have a finite population and we will select our sample of its members at random with replacement—the dumb form of random sampling. The result is that we have i.i.d. random variables which means the same thing as having i.i.d. trials.

3. We have a finite population and we have selected our sample of its members at random without replacement—the smart form of random sampling. If n/N—the ratio of sample size to population size—is 0.05 or smaller, then we get a good approximation if we treat our random variables as i.i.d.

In this chapter, we study a very important special case of i.i.d. trials, called **Bernoulli trials**. If each trial has exactly two possible outcomes, then we have Bernoulli trials. For convenient reference, I will now explicitly state the assumptions of Bernoulli trials.

Definition 11.1 (The assumptions of Bernoulli trials.) *If we have a collection of trials that satisfy the three conditions below, then we say that we have Bernoulli trials.*

1. *Each trial results in one of two possible outcomes, denoted success (S) or failure (F).*

2. *The probability of a success remains constant from trial-to-trial and is denoted by p. Write $q = 1 - p$ for the constant probability of a failure.*

3. *The trials are independent.*

We will use the method described on page 168 of Chapter 8 to assign the labels success and failure. When we are involved in mathematical arguments, it will be convenient to represent a success by the number 1 and a failure by the number 0. Finally,

We are not interested in either of the *trivial cases* in which $p = 0$ or $p = 1$. Thus, we restrict attention to situations in which $0 < p < 1$.

One reason that Bernoulli trials are so important, is that if we have Bernoulli trials, we can calculate probabilities of a great many events. Our first tool for calculation is the multiplication rule that we learned in Chapter 10. For example, suppose that we have $n = 5$ Bernoulli trials with $p = 0.70$. The probability that the Bernoulli trials yield four successes followed by a failure is:

$$P(SSSSF) = ppppq = (0.70)^4(0.30) = 0.0720.$$

Our next tool is extremely powerful and very useful in science. It is the **binomial probability distribution**. Suppose that we plan to perform/observe n Bernoulli trials. Let X denote the total number of successes in the n trials. The probability distribution of X is given by the following equation.

$$P(X = x) = \frac{n!}{x!(n-x)!}p^x q^{n-x}, \text{ for } x = 0, 1, \ldots, n. \tag{11.1}$$

Equation 11.1 is called the **binomial probability distribution** with **parameters** n and p; it is denoted by the Bin(n, p) distribution. I will illustrate the use of this equation below, a compilation of the Bin(5,0.60) distribution. I replace n with 5, p with 0.60 and q with $(1 - p) = 0.40$. Below, I will evaluate Equation 11.1 six times, for $x = 0, 1, \ldots, 5$: You should check a couple of the following computations to make sure you are comfortable using Equation 11.1, but you don't need to verify all of them.

$$P(X = 0) = \frac{5!}{0!5!}(0.60)^0(0.40)^5 = 1(1)(0.01024) = 0.01024.$$

$$P(X = 1) = \frac{5!}{1!4!}(0.60)^1(0.40)^4 = 5(0.60)(0.0256) = 0.07680.$$

$$P(X = 2) = \frac{5!}{2!3!}(0.60)^2(0.40)^3 = 10(0.36)(0.064) = 0.23040.$$

$$P(X = 3) = \frac{5!}{3!2!}(0.60)^3(0.40)^2 = 10(0.216)(0.16) = 0.34560.$$

$$P(X = 4) = \frac{5!}{4!1!}(0.60)^4(0.40)^1 = 5(0.1296)(0.40) = 0.25920.$$

$$P(X = 5) = \frac{5!}{5!0!}(0.60)^5(0.40)^0 = 1(0.07776)(1) = 0.07776.$$

Whenever probabilities for a random variable X are given by Equation 11.1 we say that X has a binomial probability (sampling) distribution with parameters n and p and write this as $X \sim$ Bin(n, p).

There are a number of difficulties that arise when one attempts to use the binomial probability distribution. The most obvious is that each trial needs to give a dichotomous response. Sometimes it is obvious that we have a dichotomy: For example, if my trials are shooting free throws or attempting golf putts, then the natural response is that a trial results in a make or miss. Other

times, the *natural response* might not be a dichotomy, but the response of interest is. For example, in my example of the American roulette wheel in Chapter 10, the *natural response* is the winning number, but if I like to bet on red, then the response of interest has possible values *red* and *not red*. Similarly, in the game of craps, I might be primarily interested in whether or not my *come out* results in a *pass line win*, a dichotomy.

Thus, let's suppose that we have a dichotomous response. The next difficulty is that in order to calculate probabilities, we need to know the numerical values of n and p. Almost always, n is known to the researcher and if it is unknown, we might be able to salvage something by using the Poisson distribution, which you will learn about in Chapter 13. There are situations in which p is known, including the following:

- Mendelian inheritance; my roulette example above, *assuming the wheel is fair*; and my craps example, *assuming both dice are fair and they behave independently*.

In other words, sometimes I feel that the phenomenon under study is sufficiently well understood that I feel comfortable in my belief that I know the numerical value of p. Obviously, in many situations I won't know the numerical value of p. For shooting free throws or attempting golf putts I usually won't know exactly how skilled the player/golfer is. When my interest is in a finite population, typically I won't know the composition of the population; hence, I won't know the value of p.

Before I explore how to deal with the difficulty of p being unknown, let's perform a few computations when p is known.

Example 11.1 (Mendelian inheritance with the 3:1 ratio.) *If you are unfamiliar with Mendelian inheritance, you can find an explanation of the 1:2:1, 3:1 and 9:3:3:1 ratios at:*

`http://en.wikipedia.org/wiki/Mendelian_inheritance`.

Each trial—offspring—will possess either the dominant (success) or recessive (failure) phenotype. Mendelian inheritance tells us that these will occur in the ratio 3:1, which means that $p = 3q = 3(1-p)$. *Solving for* p*, this equation becomes* $4p = 3$ *and, finally,* $p = 0.75$. *Suppose that we will observe* $n = 8$ *independent trials, each with* $p = 0.75$. *Let* X *denote the total number of these eight offspring who will possess the dominant phenotype. We can calculate probabilities for* X *by using Equation 11.1 with* $n = 8$ *and* $p = 0.75$*; i.e., by using the Bin(8,0.75) probability distribution.*

Suppose, for example, that I want to calculate $P(X \geq 6)$ for X having the Bin(8,0.75) distribution. If I want to do this by hand, I must first rewrite my event of interest:

$$P(X \geq 6) = P(X = 6) + P(X = 7) + P(X = 8).$$

Next, I must evaluate Equation 11.1 *three times*; for $x = 6$, $x = 7$ and $x = 8$. Sorry, but this does not sound like fun!

Fortunately, there is a website that can help. Go to:

`http://stattrek.com/Tables/Binomial.aspx`

Table 11.1: The output, after rounding, from the binomial website for $p = 0.75$, $n = 8$ and $x = 6$.

Binomial Probability:	$P(X = 6) = 0.3115$
Cumulative Probability:	$P(X < 6) = 0.3215$
Cumulative Probability:	$P(X \leq 6) = 0.6329$
Cumulative Probability:	$P(X > 6) = 0.3671$
Cumulative Probability:	$P(X \geq 6) = 0.6785$

I will now show you how to use this website. The website requires you to enter three numbers:

- **Probability of success on a single trial**: Enter the value of p; for our current problem, I enter $p = 0.75$.
- **Number of trials**: Enter the value of n; for our current problem, I enter $n = 8$.
- **Number of successes (x)**: This is a bit tricky to explain explicitly, but once you see one example, you will understand how to do it. Because my event of interest, $(X \geq 6)$, involves the number 6, I enter $x = 6$.

After entering my values for p, n and x, I click on the *Calculate* box and obtain the output, rounded to four digits, printed in Table 11.1. The answer I want is the fifth entry in the list:

$$P(X \geq 6) = 0.6785.$$

Note that there is a great deal of redundancy in five answers in this list. Make sure you understand why the following identities are true:

- The third probability is the sum of the first two.
- The fifth probability is the sum of the first and the fourth.
- The sum of the second and the fifth probabilities equals 1.
- The sum of the third and the fourth probabilities equals 1.

For example, the third probability $P(X \leq 6)$ can be written as $P(X < 6) + P(X = 6)$. In the listing above, this becomes $0.6329 = 0.3115 + 0.3215$ which is correct except for round-off error.

The website is good for computing individual probabilities, but it is tedious to use it to generate an entire binomial distribution. For the latter objective, I use Minitab. In particular, with the help of Minitab, I obtained the Bin(8,0.75) distribution, displayed in Table 11.2. Literally, the first two columns of this table present the sampling distribution. It's easy to have the computer create the cumulative sums in the third and fourth columns, so I have included them. From the table we can find the five probabilities given by the website. For example,

$$P(X > 6) = P(X \geq 7) = 0.3671, \text{ from the fourth column.}$$

Table 11.2: The binomial distribution with $n = 8$ and $p = 0.75$.

x	$P(X = x)$	$P(X \leq x)$	$P(X \geq x)$
0	0.0000	0.0000	1.0000
1	0.0004	0.0004	1.0000
2	0.0038	0.0042	0.9996
3	0.0231	0.0273	0.9958
4	0.0865	0.1138	0.9727
5	0.2076	0.3215	0.8862
6	0.3115	0.6329	0.6785
7	0.2670	0.8999	0.3671
8	0.1001	1.0000	0.1001
Total	1.0000	—	—

11.1.1 Computational Difficulties

By trial-and-error, I discovered that if I go to Minitab and ask for the Bin(n,0.50) distribution with $n \geq 1023$, then I am given the following message:

> * ERROR * Completion of computation impossible.

If, however, $p = 0.50$ and $n \leq 1022$, then Minitab gives an answer. Similarly, Minitab reports its error message for Bin(n,0.60) if, and only if, $n \geq 1388$. The people who wrote Minitab are good programmers. I am not a very good programmer, but I *could* write a program that would handle at least some of the situations Minitab does not. How can this be? Well, as we will learn later, if n is large enough, then we can use a either a Normal curve or the Poisson distribution (see Chapter 13) to obtain good approximations to binomial probabilities. Thus, my inference is that the Minitab programmers were somewhat *casual* in writing their code because they knew that their users could opt for an approximation.

If you read through the exposition on the website

`http://stattrek.com/Tables/Binomial.aspx`,

near the bottom you will find the following:

> When the number of trials is greater than 1,000, the Binomial Calculator uses a Normal distribution to estimate the binomial probabilities.

Very soon I will give you the details of the Normal curve approximation.

Here is my advice for this course. You can trust the website's answers provided $n \leq 1000$. Do not use it for $n > 1000$ until you have read my discussion of the Normal approximation in the next section.

Figure 11.1: The Bin(100, 0.5) Distribution.

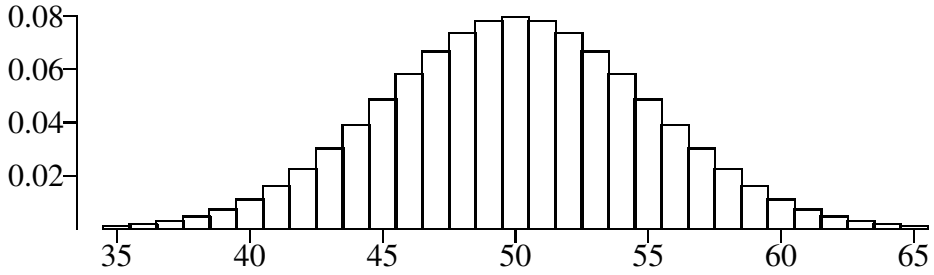


Figure 11.2: The Bin(100, 0.2) Distribution.

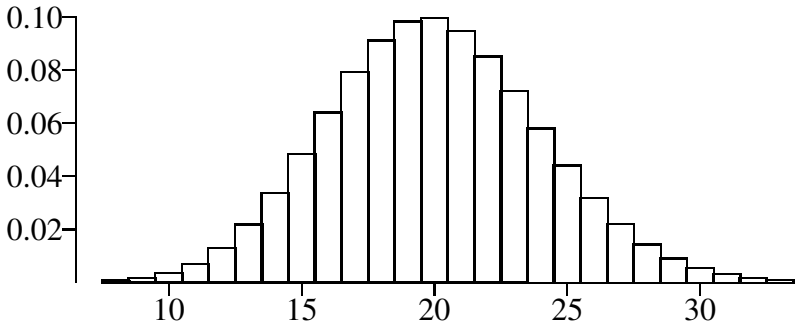


Figure 11.3: The Bin(25, 0.5) Distribution.

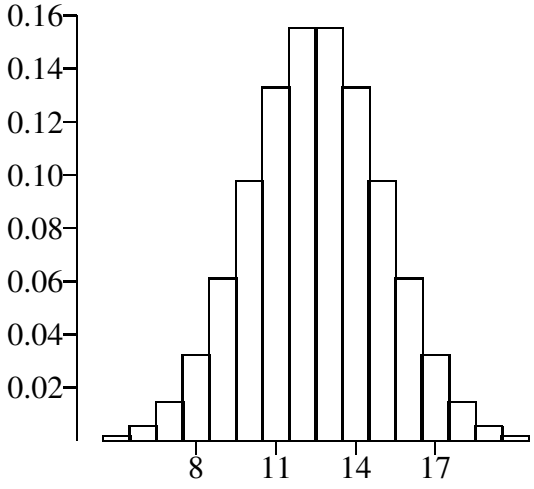
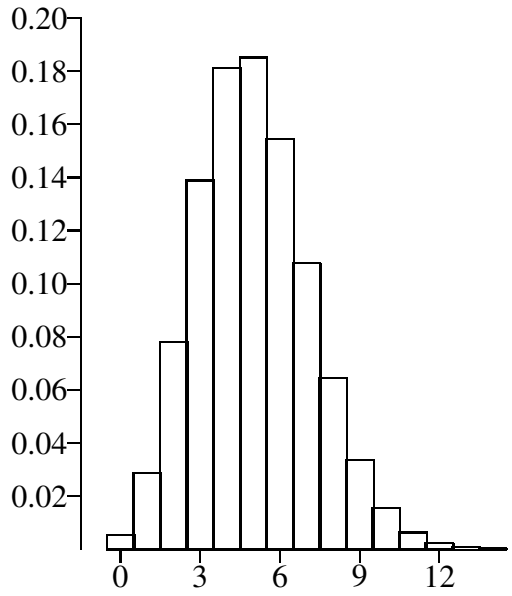


Figure 11.4: The Bin(50, 0.1) Distribution.



11.2 The Normal Curve Approximation to the Binomial

Recall that we learned how to draw a probability histogram on page 143 in Chapter 7. Figures 11.1–11.4 present **probability histograms** for several binomial probability distributions. Because $\delta = 1$ the area of each rectangle equals its height; thus, the probability of any integer value of x is the height of the rectangle centered at x.

As discussed in Chapter 7, a probability histogram allows us to 'see' a probability distribution. For example, for the four probability histograms that are presented above, the two with $p = 0.50$ are symmetric; the one with $n = 100$ and $p = 0.2$ is *almost* symmetric; and the one with $n = 50$ and $p = 0.1$ deviates a great deal from symmetry. Indeed, it can be shown that a binomial distribution is symmetric if, and only if, $p = 0.50$. Moreover, for $p \neq 0.5$, if both np and nq are *far from 0* then the binomial distribution is almost symmetric. A common guideline for *far from 0* is for both to be at least 25. We will return to this topic soon.

Below is a list of some other facts about binomial distributions.

1. The probability histogram for a binomial always has exactly one peak. The peak can be one or two rectangles wide, but never wider.

2. If np is an integer, then there is a one-rectangle wide peak located above np.

3. If np is not an integer, then the peak will occur either at the integer immediately below or above np; or, in some cases, at both of these integers.

4. If you move away from the peak in either direction, the heights of the rectangles become shorter. If the peak occurs at either 0 or n this fact is true in the one direction away from the

peak.

The following result is similar to Results 7.1–7.3 for the sum of ranks test in Chapter 7.

Result 11.1 (Mean and standard deviation of the binomial distribution.) *The mean and standard deviation of the Bin(n, p) distribution are:*

$$\mu = np \tag{11.2}$$

$$\sigma = \sqrt{npq} \tag{11.3}$$

Let's consider the Bin(100,0.50) distribution, pictured in Figure 11.1. From the above result, its mean and standard deviation are

$$\mu = np = 100(0.50) = 50 \text{ and } \sigma = \sqrt{npq} = \sqrt{100(0.50)(0.50)} = 5.$$

Suppose now that I want to compute $P(X \geq 55)$. I have three methods for obtaining this probability:

1. Because $n \leq 1000$ I can use the website

   `http://stattrek.com/Tables/Binomial.aspx`

   I go to the site, enter $p = 0.50$, $n = 100$ and $x = 55$; then I click on *Compute* and obtain the answer:

   $$P(X \geq 55) = 0.1841.$$

2. Because $n \leq 1022$, I can use Minitab. I did and obtained the answer 0.1841.

3. I can follow the method of Chapter 7 and obtain a Normal curve approximation. I go to the website:

   `http://davidmlane.com/hyperstat/z_table.html`

   I enter Mean $= 50$ and Sd $= 5$. Next to the option *Above* I enter 54.5—remember the continuity correction. The site tells me that the area under the N(50,5) curve to the right of 54.5 is 0.1841. To the nearest 0.0001, this approximation is exact!

Let me do another computational example. Consider the Bin(1200,0.60) distribution. I am interested in $P(X \leq 690)$. Again, I will try three methods for finding this probability.

1. Because $n \leq 1387$, I can use Minitab to find the exact probability. I did and obtained the answer 0.0414.

2. I calculate

   $$\mu = np = 1200(0.60) = 720 \text{ and } \sigma = \sqrt{npq} = \sqrt{1200(0.60)(0.40)} = 16.971.$$

   I go to the website:

`http://davidmlane.com/hyperstat/z_table.html`

I enter Mean $= 720$ and Sd $= 16.971$. Next to the option *Below* I enter 690.5—remember the continuity correction. The site tells me that the area under the N(720,16.971) curve to the left of 690.5 is 0.0411. If I trust Minitab, my approximate answer is too small by 0.0003. This is a very good approximation!

3. I can use the *binomial website*:

   `http://stattrek.com/Tables/Binomial.aspx`

   Because $n > 1000$, the website will give me an answer based on the Normal curve approximation. I entered $p = 0.60$, $n = 1200$, $x = 690$ and clicked on *Calculate*. The site gave me the answer 0.041. I would prefer the site to give me more digits of precision, but to the nearest 0.001, this answer agrees with my Normal curve approximation. And, of course, the binomial website is less work for me.

You might be wondering:

Is the Normal curve approximation always good?

I will give you a dramatic example of why the answer is **no!**

Let's consider the Bin(1001,0.001) distribution. I am choosing $n =$ 1,001 because this is the smallest value of n for which the *binomial website* uses the Normal curve to obtain approximate binomial probabilities. I am interested in $P(X = 0)$. Again, I will show you three methods.

1. Minitab gives me the exact probability, 0.3673.

2. I can calculate the exact probability using Equation 11.1:

$$P(X = 0) = \frac{1001!}{0!1001!}(0.001)^0(0.999)^{1001}.$$

   After canceling the factorials and noting that $(0.001)^0 = 1$, this answer reduces to

$$(0.999)^{1001}.$$

   Even the calculator on my cell phone can evaluate this! It obtains the correct answer, 0.3673.

3. I went to the *binomial website*, entered $p = 0.001$, $n =$ 1,001, $x = 0$ and clicked on *Calculate*. The website did not calculate an answer; instead, it printed the message:

   PROCESSING ERROR: Cannot complete computation.

   This is good; the program is smart enough not to use the Normal approximation. Sadly, however, I have bad news to report. Staying on the site, I calculated probabilities for $n =$

1,001 and $p = 0.5$; the site worked fine. Then I reentered $p = 0.001$ $x = 0$ and $n =$ 1,001; this time, the site gave me:

$$P(X = 0) = 0.025 \text{ and } P(X < 0) = 0.217.$$

Because X counts successes, it cannot be negative!

Thus, the kindest thing I can say is that for $n > 1000$ the site's behavior is erratic; do not use it unless you check that both np and nq equal or exceed 25.

In summary, here are two general guidelines I recommend you use.

1. **For the purpose of computing probabilities:** If $n \leq 1000$ use the binomial website. If $n > 1000$, $np \geq 25$ and $nq \geq 25$: you may use the binomial website or obtain the Normal curve approximation *by hand*.

2. **For the development of estimation and prediction intervals:** Use the Normal approximation to the binomial only if both np and nq are greater than or equal to 25.

Let me make a few comments on these guidelines. First, all statisticians agree that we need to consider the values of np and nq; not all would agree on my magic threshold of 25. Second, if $n > 1000$ and, say, $np < 25$ we can use the Poisson distribution to approximate the binomial; this material will be presented in Chapter 13.

Third, the second guideline is a bit odd. It implies, for example, that for $n = 50$ and $p = q = 0.50$ we **may use** the Normal curve approximation even though exact probabilities are readily available from the website. As you will learn in Chapter 12, being able to use the Normal curve approximation is very helpful for the development of general formulas.

11.3 Calculating Binomial Probabilities When p is Unknown

I could make this a very short section by simply remarking that if p is unknown then obviously neither the website, Minitab nor I can evaluate Equation 11.1. In addition, if p is unknown then we can calculate neither the mean nor standard deviation of the binomial, both of which are needed for the Normal curve approximation.

We do have this section, however, because I want to explore the idea of what if means to *know the value of* p. I am typing this in October, 2013. To date in his NBA (National Basketball Association) career, during the regular season Kobe Bryant has made 7,932 free throws out of 9,468 attempts. What can we do with these data in the context of what we have learned in this chapter?

I am always interested in computing probabilities; thus, when faced with a new situation I ask myself whether it is reasonable to assume a structure that will allow me to do so. Well, each free throw attempted by Bryant can be viewed as a trial, so I might assume that his 9,468 attempts were observations of i.i.d. trials. As a side note, let me state that years ago I was active in the *Statistics in Sports* section of the American Statistical Association. We had many vigorous debates—and many papers have been written—on the issue of whether the assumption of i.i.d. trials is reasonable in

sports in general, not just for free throws in basketball. In order to avoid a digression that could consume months, if not years, let's tentatively assume that we have i.i.d. trials for Bryant shooting free throws.

The next issue is the value of p for Bryant's trials. Bryant shooting a free throw was not as simplistic as, say, tossing a fair coin or casting a balanced die; nor was it as well-behaved as Mendelian inheritance. In short, there is no reason to believe that we **know** the value of p for Bryant or, indeed, any basketball player. But what do I mean by *know*? We know that p is strictly between 0 and 1. As any mathematician will tell you, *there are a lot of numbers between 0 and 1*! (The technical term is that there are an uncountable infinity of numbers between 0 and 1.) But we need to think more like a scientist and less like a mathematician. In particular, by scientist I mean someone who is—or strives to be—knowledgeable about basketball. A mathematician will (correctly) state that 0.7630948 and 0.7634392 are different numbers, but as a basketball fan, I don't see any *practical difference* between p equaling one or the other of these. In either situation, I would round to three digits and say, "The player's true ability, p, is that in the long-run he/she makes 76.3% of attempted free throws."

Bryant's data give us

$$\hat{p} = 7932/9468 = 0.838;$$

in words, during his career, to date, Bryant has made 83.8% of his free throws. As might be clear—and if not, we will revisit the topic in the next chapter—we may calculate the nearly certain interval (Formula 4.1 in Chapter 4) for p:

$$0.838 \pm 3\sqrt{\frac{0.838(0.162)}{9468}} = 0.838 \pm 0.011 = [0.827, 0.849].$$

Thus, given our assumption of i.i.d. trials, we don't know Bryant's p exactly, but every number in its nearly certain interval is quite close to his value of $\hat{p}$.

Thus, if I wanted to compute a probability for Bryant, I would be willing to use $p = 0.838$.

Let's consider the first $n = 50$ free throws that Bryant will attempt during the 2013–2014 NBA season. I am interested in the number of these free throws that he will make; call it X. Based on my discussion above I view X as having the Bin(50,0.838) distribution.

For example, I go to the website

`http://stattrek.com/Tables/Binomial.aspx`,

and enter $p = 0.838$, $n = 50$ and $x = 38$. I click on *Calculate* and obtain

$$P(X \geq 38) = 0.9482.$$

Thus, I believe that the probability that Kobe Bryant will make at least 38 of his first 50 free throws this season is just under 95%.

(**Optional enrichment for basketball fans.** It can be shown that the probability of the event I examine above, $(X \geq 38)$ is an increasing function of p. I found that for $p = 0.838$, this probability is 0.9482. If I use the lower bound of my nearly certain interval, $p = 0.827$, the website gives me 0.9201 for the probability of this event. If I use the upper bound of my nearly certain interval,

Table 11.3: The conference of the Super Bowl winner, by year. Year 1 denotes the 1966 season (game played in 1967) and year 47 denotes the 2012 season (game played in 2013). An NFC winner is labeled a success, denoted by 1, and an AFC winner is denoted by 0.

Year:	1	2	3	4	5	6	7	8	9	10	11	12
Winner:	**1**	**1**	0	0	0	**1**	0	0	0	0	0	**1**
Year:	13	14	15	16	17	18	19	20	21	22	23	24
Winner:	0	0	0	**1**	**1**	0	**1**	**1**	**1**	**1**	**1**	**1**
Year:	25	26	27	28	29	30	31	32	33	34	35	36
Winner:	**1**	**1**	**1**	**1**	**1**	**1**	**1**	0	0	**1**	0	0
Year:	37	38	39	40	41	42	43	44	45	46	47	
Winner:	**1**	0	0	0	0	**1**	0	**1**	**1**	**1**	0	

$p = 0.849$, the website gives me 0.9684 for the probability of this event. Thus, another approach is for me to say that I am nearly certain that the probability of the event $(X \geq 38)$ is between 0.9201 and 0.9684.)

11.4 Runs in Dichotomous Trials

There have been 47 Super Bowls played. Each game resulted in a winner from (using current names) the National Football Conference (NFC), which I will call a success, or the American Football Conference (AFC), which I will call a failure. The 47 outcomes, in time-order, are presented in Table 11.3. The outcome, *conference of the Super Bowl winner* is a dichotomous response. I am **not** assuming that they are the outcomes of 47 Bernoulli trials. In fact, later in this section I will conclude that they don't seem to be the result of observing Bernoulli trials. I am getting ahead of myself.

Our Super Bowl data contains 18 runs which are detailed (and implicitly defined) in Table 11.4. There is a great deal of information in this table. You don't need to study it exhaustively, but you should understand how it was constructed, which I will now explain.

The responses for years 1 and 2 are successes, but year 3 is a failure. Thus, the data begins with a run of successes of length 2 covering years 1 and 2. Next, the data has a run of failures of length 3 covering years 3–5. And so on.

In addition to noting that there are 18 runs in the Super Bowl data, it's difficult to miss the existence of the very long run of successes, 13, spanning years 19–31.

In general, any or all of of the following statistics may be used to investigate the issue of whether the data are the result of observing Bernoulli trials.

- The number of runs;

Table 11.4: The 18 runs for the data in Table 11.3.

Run:	1	2	3	4	5	6	7	8	9
Year(s):	1–2	3–5	6	7–11	12	13–15	16–17	18	19–31
Length:	2	3	1	5	1	3	2	1	13
Type:	S	F	S	F	S	F	S	F	S
Run:	10	11	12	13	14	15	16	17	18
Year(s):	32–33	34	35-36	37	38–41	42	43	44–46	47
Length:	2	1	2	1	4	1	1	3	1
Type:	F	S	F	S	F	S	F	S	F

- The length of the longest run of successes; and
- The length of the longest run of failures.

In particular:

- Does it appear that there is a constant probability of success from trial-to-trial?
- Do the trials appear to be independent?

This topic is frustrating for a number of reasons. In part because it is so frustrating, this topic typically is not presented in an introductory Statistics class. I will, however, discuss these issues in this section briefly for the following reasons.

1. I feel that I am doing you a disservice if I state assumptions without giving any idea of how to investigate their validity.
2. I feel that I am doing you a disservice if I present you with a *sanitized view* of Statistics; a view in which there are no controversies and no confusion about how to analyze data.
3. In my experience, this is one of the topics in Statistics that non-statisticians find very interesting. Especially the occurrence of long runs of successes (or failures) are interesting to people.

Let me now define some ideas rather formally. I plan to observe a total of n dichotomous trials. I want to investigate whether the dichotomous trials satisfy the assumptions of Bernoulli trials. I decide to pursue this by conducting a test of hypotheses. My null hypothesis is that the trials are Bernoulli trials. This is in line with the principle of Occam's Razor. Also, usually a researcher *wants* to have Bernoulli trials because Bernoulli trials allow the computation of many answers. Of course, there are some researchers—and I often find myself in this camp—who sometimes are hoping **not** to have Bernoulli trials because sometimes it is nice to live in a world that is a bit more complicated.

Please allow me a brief digression. Many textbooks state that the *alternative always represents what the researcher is trying to prove*. Often times, this is a valid view of the hypotheses, **but not always**. For example, in this current section, I must assume that the null is Bernoulli trials because otherwise there is no way to find a sampling distribution and so on; it doesn't really matter what I prefer to be true!

Wait a minute. You might be thinking—even though we don't yet have a test statistic:

> We need to be able to determine the sampling distribution of the test statistic **under the assumption that the null hypothesis is true**. This is hopeless! The null hypothesis does not specify the value of p for the Bernoulli trials; thus, it will be impossible to calculate a unique set of probabilities!

If these are your thoughts, then you are correct. Well, almost correct. The trick is that we use a *conditional* test statistic. Let me explain.

Let's look at the Super Bowl example again. The total number of trials in the data set is $n = 47$. Before collecting data I didn't know what the value of X, the total number of successes in the 47 trials, would be. After I collect the data I know that the observed value of X is $x = 25$. The trick is that I condition on X eventually being 25. Let me make the following points.

1. Most (nearly all?) statisticians feel fine about this conditioning; here is why. Given that $X = 25$—and therefore that the number of failures, $n - X = 47 - X$ is 22—what have I learned? I have learned that over the course of the data collection, neither conference had a huge advantage over the other. **But**, and this is the key point, knowing that $X = 25$ gives me no information about whether p is changing or whether the trials are independent. In other words, knowing that $X = 25$ gives me no information about whether the assumptions of Bernoulli trials are reasonable.

2. This point is a little more esoteric. I have always extolled you to remember: probabilities are calculated before we collect data. Conditioning on $X = 25$ and then calculating probabilities—as I will soon do—looks a lot like I am violating my directive. But I am not. Actually, before collecting data I can imagine 48 different computations, one for each of the 48 possible values of X (0, 1, 2, ..., 47). If I were to perform all of these computations, after I collect data I would find that my computations conditional on $X = x$ would be irrelevant for all $x \neq 25$. Why should I perform computations that I won't use?

I will now explain why conditioning is so useful mathematically for our problem. Consider the Super Bowl data again. Conditional on knowing $X = 25$, we know that the data will consist of an arrangement of 25 1's and 22 0's. The number of such arrangements is:

$$\frac{47!}{25!22!} = 1.483 \times 10^{13},$$

almost 15 trillion. It can be shown that, on the assumption that the null hypothesis is true, these arrangements are equally likely to occur, regardless of the value of p! Thus, if we choose our test statistic to be a function of the arrangement of 1's and 0's, then we can compute its sampling distribution **without knowing the value of** p.

Table 11.5: All possible arrangements of three 1's and two 0's. For each arrangement, the observed values of R, V and W are given, where R is the number of runs, V is the longest run of successes, and W is the longest run of failures. On the assumption that the data come from Bernoulli trials, conditional on obtaining three successes, the 10 arrangements are equally likely to occur.

Arrangement	r	v	w	Arrangement	r	v	w
11100	2	3	2	10011	3	2	2
11010	4	2	1	01110	3	3	1
11001	3	2	2	01101	4	2	1
10110	4	2	1	01011	4	2	1
10101	5	1	1	00111	2	3	2

Table 11.6: The sampling distributions of R, V and W for the 10 equally likely arrangements in Table 11.5.

r:	2	3	4	5	Total
$P(R = r)$:	0.2	0.3	0.4	0.1	1.0

v:	1	2	3	Total
$P(V = v)$	0.1	0.6	0.3	1.0

w:	1	2	Total
$P(W = w)$	0.6	0.4	1.0

Admittedly, 15 trillion is a lot of arrangements to visualize! To help make this result more concrete, I will do a quick example with $n = 5$ and conditioning on $X = 3$; see Table 11.5. I will now explain the information in this table.

With five trials, conditional on a total of three successes and two failures, the number of possible arrangements is:

$$\frac{5!}{3!2!} = \frac{5(4)}{2(1)} = 10;$$

the 10 arrangements are listed in the table. For each arrangement I determine the observed values of: the number of runs; the length of the longest run of successes; and the length of the longest run of failures. These test statistics are denoted by R, V and W, respectively. (I am sorry that V and W are not evocative of the successes and failures, but I feared confusion if I used S and F.) Once we have the results in Table 11.5, it is easy to obtain the sampling distributions of R, V and W. These sampling distributions are shown in Table 11.6.

We have seen that for a small number of trials it is possible to find the exact sampling distribution of any of the test statistics R, V and W. We will not find **exact** distributions for any practical example because there are too many arrangements to consider. It is quite easy to perform a sim-

ulation experiment for any of these test statistics; indeed, you might have noticed how similar the current situation is to selecting assignments for a CRD at random. (If you don't see the connection, no worries.)

There is a fancy math approximation for the sampling distribution of R and I will discuss it briefly in the following subsection.

11.4.1 The Runs Test

The null hypothesis is that the trials are Bernoulli trials. The test statistic is R. Pause for a moment. What is missing? Correct; I have not specified the alternative hypothesis. For the math level I want to maintain in this course, a careful presentation of the alternative is not possible. Instead, I will proceed by examples.

But first, I want to give you the formulas for the mean and standard deviation of the sampling distribution of R.

Result 11.2 (The mean and standard deviation of the sampling distribution of R.) *Conditional on the number of successes, x, and number of failures, $n - x$, in a sequence of n Bernoulli trials, the mean and standard deviation of the number of runs, R, are given by the equations below. First compute*

$$c = 2x(n - x); \text{ then} \tag{11.4}$$

$$\mu = 1 + \frac{c}{n}; \text{ and} \tag{11.5}$$

$$\sigma = \sqrt{\frac{c(c - n)}{n^2(n - 1)}}. \tag{11.6}$$

Recall my artificial small example—with $n = 5$, $x = 3$ and $n - x = 2$. First, I calculate $c = 2(3)(2) = 12$. The mean and standard deviation are:

$$\mu = 1 + 12/5 = 3.4 \text{ and } \sigma = \sqrt{\frac{12(7)}{5^2(4)}} = \sqrt{0.84} = 0.916.$$

(If you look at the 10 possible arrangements in Table 11.5, and sum the corresponding 10 values of r, you will find that the sum is 34 and the mean is $34/10 = 3.4$, in agreement with the above use of Equation 11.5.)

For my Super Bowl data, $n = 47$, $x = 25$ and $n - x = 22$. Thus, $c = 2(25)(22) = 1100$. The mean and standard deviation are:

$$\mu = 1 + 1100/47 = 24.404 \text{ and } \sigma = \sqrt{\frac{1100(1053)}{47^2(46)}} = \sqrt{11.399} = 3.376.$$

Note that the observed number of runs, 18, is smaller than the mean number under the null hypothesis. This will be relevant very soon.

I need to talk about the alternative, but first I need to present an artificial example. Imagine that I have $n = 50$ dichotomous trials with $x = n - x = 25$. From Result 11.2, $c = 2(25)(25) = 1250$. The mean and standard deviation are:

$$\mu = 1 + 1250/50 = 26 \text{ and } \sigma = \sqrt{\frac{1250(1200)}{50^2(49)}} = \sqrt{12.245} = 3.499.$$

Based on my intuition, there are two arrangements that clearly provide very strong evidence against Bernoulli trials. (I know. We are not supposed to talk about evidence against the null; please bear with me.) The first is a perfect alternating arrangement:

10101 01010 10101 01010 10101 01010 10101 01010 10101 01010

The second is 25 successes followed by 25 failures:

11111 11111 11111 11111 11111 00000 00000 00000 00000 00000

For the first of these arrangements, $R = 50$ and $V = W = 1$. For the second arrangement, $R = 2$ and $V = W = 25$. It is easy to see that the first arrangement gives the largest possible value for R—after all, there cannot be more runs than trials! Similarly, the first arrangement gives the smallest possible value for both V and W. Also, the second arrangement gives the smallest possible value of R and the largest possible value for both V and W. If we consider all possible arrangements it makes sense that large [small] values of R tend to be matched with small [large] values of both V and W. The important consequence of this tendency is that when I do talk about alternatives, the $<$ alternative for R will correspond to the $>$ alternative for V or W.

Note that everything I say about the first arrangement is also true for the arrangement:

01010 10101 01010 10101 01010 10101 01010 10101 01010 10101

Similarly, everything I say about the second arrangement is also true for the arrangement:

00000 00000 00000 00000 00000 11111 11111 11111 11111 11111

Let's suppose that I have convinced you that both of these arrangements—or, if you prefer all four—provide convincing evidence that we do **not** have Bernoulli trials. I now look at the question: Which assumption of Bernoulli trials is being violated? Here are two possible interpretations of the data in the second arrangement which, recall, consists of 25 successes followed by 25 failures:

1. There is almost—but not quite—perfect positive dependence. A success [failure] is almost always followed by a success [failure]. In fact, the only wrong prediction in the *predict the response to remain the same* paradigm occurs at trials 25 to 26.

2. The value of p is 1 for the first 25 trials and is 0 for last 25 trials.

Thus, we don't have Bernoulli trials, but is it because of dependence or because p changes? Note that I am talking about *what we know from the data*; it's possible that your knowledge of the science behind the study will lead you to discard one of my two explanations.

I will **not** give a careful statement of the alternative for the test based on R. Instead I will say that by $<$ I mean that the true model tends to give fewer runs than the (null) Bernoulli trials model. By $>$ I mean that the true model tends to give more runs than the (null) Bernoulli trials model. As usual, by $\neq$ I mean that the true model tends to give either fewer or more runs than the the (null) Bernoulli trials model. Below are the rules for finding the P-value.

Result 11.3 (The P-value for the runs test.) *In each of the equations below, r denotes the observed value of the test statistic R.*

1. *For the alternative $>$, the P-value equals*

$$P(R \geq r) \tag{11.7}$$

2. *For the alternative $<$, the P-value equals*

$$P(R \leq r) \tag{11.8}$$

3. *For the alternative $\neq$, the P-value is the smallest of the following three numbers: 1; twice the P-value for $>$; and twice the P-value for $<$.*

In the interest of intellectual honesty, I mention that there is some minor controversy over my rule for the P-value for $\neq$. But—as best I can tell—only a small proportion of professional statisticians care about it. In general, for many tests of hypotheses, the math arguments for the alternative $\neq$ can be messy. Also, in the examples in this course, we will use a Normal curve approximation, with the continuity correction of 0.5, to obtain approximate P-values for the runs test. I have not been able to locate a good general guideline for using the Normal curve approximation. Thus, I will (somewhat, ok, quite) arbitrarily state that you should use the Normal curve approximation only if the null standard deviation of R is 3 or larger. Thus, we would not use the Normal curve approximation if $n = 5$ and $x = 3$, because its standard deviation is 0.916. We can, however, use the Normal curve approximation for the Super Bowl data because its standard deviation, 3.376, is larger than 3.

I will now illustrate the use of Result 11.3 with the Super Bowl data. Recall that $r = 18$, $\mu = 24.404$ and $\sigma = 3.376$. Because R takes on integer values only, I will use the continuity correction. I go to the Normal curve website:

`http://davidmlane.com/hyperstat/z_table.html`,

enter 24.404 for the *Mean* and 3.376 for the *Sd*.

1. For $>$, I want $P(R \geq 18)$. I find that the area *above* 17.5 is 0.9796.

2. For $<$, I want $P(R \leq 18)$. I find that the area *below* 18.5 is 0.0402.

3. For $\neq$, the approximate P-value is $2(0.0402) = 0.0804$.

Table 11.7: The partial approximate sampling distribution of V, the longest run of successes, in 47 Bernoulli trials, conditional on a total of 25 successes.

v:	9	10	11	12	13	14
Rel. Freq. $(V \geq v)$	0.0355	0.0134	0.0049	0.0017	0.0004	0.0001

Let me note that I have been unable to find a website that allows us to enter our data for a runs test. As always, please let me know if you find one. Minitab will perform a runs test. It uses the Normal curve approximation, but it has two curious features:

1. Minitab does not use the continuity correction. Thus, its P-value for $\neq$ is 0.0578, substantially smaller than the answer—0.0804—obtained with the continuity correction.

2. Minitab gives the P-value only for the alternative $\neq$.

Regarding this second item; one could, of course, halve Minitab's answer—0.0578—to obtain 0.0289 as the approximate P-value for the alternative supported by the data, in this case $<$. I suspect that Minitab's creators are expressing the belief—which has merit—that when we check assumptions we should not be overly restrictive in our choice of alternative.

11.4.2 The Test Statistics V and W

The presentation in this section will be very brief. The null hypothesis is that the trials are Bernoulli trials. I will consider two possible test statistics: V [W] the length of the longest run of successes [failures]. The only alternative I will explore is $>$, that the true model tends to give larger values of V [W] than the (null) Bernoulli trials model. The P-value for the alternative $>$ and the test statistic V is

$$P(V \geq v), \tag{11.9}$$

where v is the observed value of V. The P-value for the alternative $>$ and the test statistic W is

$$P(W \geq w), \tag{11.10}$$

where w is the observed value of W.

I can obtain the exact sampling distributions of V and W only for very small values of n. I am unaware of the existence of any accurate fancy math approximation to either of these sampling distributions. Therefore, we will obtain approximations to these sampling distributions by using a computer simulation experiment.

I performed a simulation experiment with 10,000 reps on the Super Bowl data; my approximate distribution for V is in Table 11.7. Recall that the observed value of V for the Super Bowl data is $v = 13$. Thus, the approximate P-value for the alternative $>$ is 0.0004, a very small number. I am quite convinced that the Super Bowl data did not come from Bernoulli trials.

The observed value of W is 5 for the Super Bowl data; the AFC's longest run of victories occurred in years 7–11. I anticipated that the P-value for W would not be small; thus, I performed

Table 11.8: Katie's data on day 7. She obtained a total of 68 successes and 32 failures. The observed values of the various test statistics are $r = 42$, $v = 19$ and $w = 4$.

SS F SS F S FFF SSSSS FF SSSSSSSS F S F SS F SS F SS FF SSSSS FF SS FF SS F SSS F S FFFF SSSSSSSSSSSSSSSSSS FF SSS F SS FF S F SS F SS F S F

a simulation experiment with only 1,000 reps. The relative frequency of occurrence of $W \geq 5$ was 0.422; thus, this is my approximate P-value.

I end this section with a small piece of a large study. Katie Voigt was a starting shooting guard on the women's basketball team at the University of Wisconsin–Madison. A few years later she was kind enough to collect a large amount of data and allow me to analyze it. For each of 20 days, after warming-up, Katie would attempt 100 shots from a favorite spot behind the three-point line. In the following example I will tell you about Katie's data on day 7. The last Practice Problem will look at her data from day 3 and the last Homework Problem will look at her data from day 8.

Example 11.2 (Day 7 of Katie's Study of Shooting.) *On day 7, Katie made 68 of her 100 attempts. The data are presented in Table 11.8. I do* **not** *want you to verify any of the numbers in the caption of the table. Indeed, the table's construction is* **not** *conducive to verifying calculations. Rather, I want you to* look at *the data and form an impression of it.*

I will now find the approximate P-values for Katie's day 7 data, for each of our three test statistics. I will use the alternative $<$ for R for illustration because a small value of R tends to go with large values of V and W, as I discussed earlier.

First, I need to use Result 11.2 to find the (null) mean and standard deviation of R. First, $c = 2x(n - x) = 2(68)(32) = 4{,}352$. The mean and the standard deviation are:

$$\mu = 1 + c/n = 1 + (4352/100) = 44.52 \text{ and } \sigma = \sqrt{\frac{c(c-n)}{n^2(n-1)}} = \sqrt{\frac{4352(4252)}{100^2(99)}} = 4.323.$$

For the alternative $<$ the P-value equals $P(R \leq r) = P(R \leq 42)$. I go to the website:

`http://davidmlane.com/hyperstat/z_table.html`

I enter Mean $= 44.52$ and Sd $= 4.323$. Next to the option *Below* I enter 42.5—remember the continuity correction. The site tells me that the approximate P-value is 0.3202. Thus, the runs test detects only weak evidence in support of the alternative.

I performed two simulation experiments; with 10,000 reps for V and 1,000 reps for W. I obtained the following relative frequencies:

$$\text{Rel. Freq. } (V \geq 19) = 0.0074 \text{ and Rel. Freq. } (W \geq 4) = 0.525$$

Thus, V is very sensitive to the evidence in the data, while W is not.

11.5 Summary

If the response is a dichotomy, then i.i.d. trials are called Bernoulli trials. One possible response is denoted a success and the other is a failure. The probability of a success on any particular trial is denoted by p and the probability of a failure on any particular trial is denoted by $q = 1 - p$. We restrict attention to p's that satisfy $0 < p < 1$; i.e., we are not interested in the cases in which $p = 0$ or $p = 1$.

Let X denote the total number of successes in n Bernoulli trials. The sampling (probability) distribution of X is given by Equation 11.1:

$$P(X = x) = \frac{n!}{x!(n-x)!} p^x q^{n-x}, \text{ for } x = 0, 1, \ldots, n.$$

The above is really a *family* of equations, because we get a different equation by changing n and/or p. It is referred to as a binomial distribution with parameters n and p and a particular member of this family is denoted by the Bin(n, p) distribution.

Except for very small values of n it is tedious to calculate a binomial probability by hand. If $n \leq$ 1,000 the website:

`http://stattrek.com/Tables/Binomial.aspx`

provides exact probabilities. If $n >$ 1,000 and both np and nq are 25 or larger, then we can obtain good approximations to the binomial by using the Normal curve with

$$\mu = np \text{ and } \sigma = \sqrt{npq}.$$

As you learned in Chapter 7, the website:

`http://davidmlane.com/hyperstat/z_table.html`

can be used to obtain areas under a Normal curve—remember to use the continuity correction.

Instead of obtaining the Normal approximation *by hand*—by which I mean calculating the values of μ and σ by hand—if $n >$ 1,000 and both np and nq are 25 or larger, then the website

`http://stattrek.com/Tables/Binomial.aspx`

will do the work for you. **But be careful!** This website **does not always check the values of** np and nq and if either of these is small, then the approximate answer from the website can be very bad.

Stating the obvious, we cannot compute a binomial probability—exact or approximate—without knowledge of the numerical value of p. In Section 11.3, I argue that if I have data from a huge number of Bernoulli trials—huge in this context means, to me, approximately 10,000 or more, provided p is not too close to either 0 or 1—then I am willing to replace p by the proportion of successes in the data, $\hat{p}$. I will return to this topic in the next chapter.

It is more than a bit unsatisfactory to be forced to simply assume or fail to assume that one has Bernoulli trials. One might, for example, want to perform a test of hypotheses with the null hypothesis that the assumptions of Bernoulli trials are correct. In Section 11.4, I provide an approach

to this test of hypotheses with three possible test statistics: the number or runs, R; the length of the longest run of successes, V; and the length of the longest run of failures, W.

Apart from the issue of a test of hypotheses, sports fans have long been interested in long runs of successes; for example, Joe DiMaggio's record of hitting safely in 56 consecutive major league baseball games and Micheal Williams's record of 97 consecutive free throws made during NBA games. Knowing the sampling distribution of V—in addition to its use in finding a P-value—can be used to gauge how *remarkable* a record might be.

11.6 Practice Problems

1. Suppose that we have Bernoulli trials with $p = 0.8$.

   (a) Calculate the probability that the first four trials yield: two successes, then a failure and then a success.

   (b) Calculate, by hand, the probability that the first six trials yield a total of five or more successes.

   (c) Use the website:

   `http://stattrek.com/Tables/Binomial.aspx`

   to verify your answer to (b).

2. Let X denote the total number of successes in $n = 900$ Bernoulli trials for which $p = 0.7$.

   (a) Use the website

   `http://stattrek.com/Tables/Binomial.aspx`

   to find the exact value of: $P(X \geq 600)$; and $P(X \leq 645)$.

   (b) Calculate the mean and standard deviation of X.

   (c) Find the Normal curve approximation of: $P(X \geq 600)$; and $P(X \leq 645)$; remember to use the continuity correction.

   (d) Compare your answers in (a) and (c)

3. I programmed Minitab to generate 40 Bernoulli trials with $p = 0.6$. Below is the result I was given. (The spaces have no meaning; I typed them to make the counting in part (a) easier.)

   1 00 1111 00 1 00 1 0 1 0 1111111 00 1 00 1 00 111111 0 1 0

   (a) Determine the observed values of X, $n - X$, R, V and W.

   (b) Conditional on the observed value of X calculate the mean and standard deviation of R

   (c) Use the Normal approximation to compute the P-value for the test statistic R and the alternative $\neq$.

   (d) I performed a simulation experiment with 1,000 reps and obtained the following frequencies:

   $$\text{Frequency } (V \geq v) = 305 \text{ and Frequency } (W \geq w) = 887,$$

   where the values of v and w were determined in part (a). Comment on these results.

4. I examined the outcomes of the last 100 World Series in baseball, for the years 1912–2012—there was no Series in 1994. The American League team won (a success) 58 Series. The observed values of our three test statistics are $r = 58$, $v = 7$ and $w = 4$.

Table 11.9: The partial approximate sampling distribution of V, the longest run of successes, in 100 Bernoulli trials, conditional on a total of 66 successes.

v:	11	12	13	14
Relative Freq. $(V \geq v)$	0.2491	0.1536	0.0951	0.0587

Table 11.10: The partial approximate sampling distribution of W, the longest run of failures, in 100 Bernoulli trials, conditional on a total of 34 failures.

w:	7	8	9	10
Relative Freq. $(W \geq w)$	0.0219	0.0072	0.0019	0.0008

(a) Calculate the mean and standard deviation of R, conditional on $X = 58$ successes.

(b) Use the Normal curve approximation to find an approximate P-value for the runs test with alternative $>$.

(c) A simulation experiment with 1,000 reps yielded:

$$\text{Frequency } (V \geq 7) = 595 \text{ and Frequency } (W \geq 4) = 875.$$

Comment on the meaning of these results.

5. Recall day 7 of Katie's study of shooting, Example 11.2 and the description immediately before it. In this problem we will analyze Katie's data from day 3.

Here are some important statistics from Katie's 100 trials on day 3:

$$x = 66, r = 31, v = 12 \text{ and } w = 9.$$

(a) Use the Normal curve approximation to obtain the P-value for the runs test and the alternative $<$. Comment.

(b) I performed a simulation study with 10,000 reps. Partial results for the test statistic V are in Table 11.9. Use this information to find the approximate P-value for the test statistic V and the alternative $>$.

(c) I performed a simulation study with 10,000 reps. Partial results for the test statistic W are in Table 11.10. Use this information to find the approximate P-value for the test statistic W and the alternative $>$.

11.7 Solutions to Practice Problems

1. (a) The probability of the sequence SSFS is obtained by using the multiplication rule:

$$P(SSFS) = ppqp = (0.8)^3(0.2) = 0.1024.$$

   (b) The probability of interest is

$$P(X \geq 5) = P(X = 5) + P(X = 6) = \frac{6!}{5!1!}(0.8)^5(0.2) + (0.8)^6 =$$

$$0.3932 + 0.2621 = 0.6553.$$

   (c) I entered the values $p = 0.8$, $n = 6$ and $x = 5$; then I clicked on *Calculate* and obtained the answer:

$$P(X \geq 5) = 0.6554,$$

   which equals my answer, except for (my) round-off error.

2. (a) I entered the values $p = 0.7$, $n = 900$ and $x = 600$; then I clicked on *Calculate* and obtained the answer:

$$P(X \geq 600) = 0.9861.$$

   I changed my x to 645, then I clicked on *Calculate* and obtained the answer:

$$P(X \leq 645) = 0.8705.$$

   (b) The values are:

$$\mu = 900(0.7) = 630 \text{ and } \sigma = \sqrt{900(0.7)(0.3)} = 13.75.$$

   (c) For $P(X \geq 600)$: The area under the N(630,13.75) curve to the right of 599.5 is 0.9867.

   For $P(X \leq 645)$: The area under the N(630,13.75) curve to the left of 645.5 is 0.8702.

   (d) The first approximation is too large by 0.0006; the second is too small by 0.0003. Both approximations are quite good.

3. (a) By counting, I obtain:

$$X = 24, n - X = 16, R = 20, V = 7 \text{ and } W = 2.$$

   (b) First, $c = 2(24)(16) = 768$. Thus,

$$\mu = 1 + 768/40 = 20.2 \text{ and } \sigma = \sqrt{\frac{768(728)}{40^2(39)}} = 2.993.$$

(c) Note that the approximating Normal curve has mean equal to 20.2. For the alternative $<$, we need the area under the Normal curve to the left of 20.5. This area must be larger than 0.5 because $20.5 > 20.2$. For the alternative $>$, we need the area under the Normal curve to the right of 19.5. This area must be larger than 0.5 because $19.5 < 20.2$. The P-value for $\neq$ is the minimum of three numbers: 1 and two numbers that are both larger than 1. Hence, the P-value is 1.

(d) The smallest possible value of W is 1 and we obtained $W = 2$. Thus, it is no surprise that its approximate P-value is huge, 0.887. A run of seven successes seems, to my intuition, to be pretty long, but the relatively large approximate P-value, 0.305, means that the data's support for $>$ is not very strong.

4. (a) First, $c = 2(58)(42) = 4{,}872$. Thus,

$$\mu = 1 + 4872/100 = 49.72 \text{ and } \sigma = \sqrt{\frac{4872(4772)}{100^2(99)}} = 4.846.$$

(b) I need the area under the N(49.72,4.846) curve to the right of 57.5. This area equals 0.0542. The large number of runs is not quite statistically significant, but it indicates that the data set includes a great deal of switching between 0's and 1's.

(c) Because of the large amount of switching back-and-forth noted in part (b), I don't expect the lengths of runs will be very noteworthy. Indeed, my approximate P-value for $>$ for V [W] is 0.595 [0.875]. Thus, neither test statistic provides much evidence in support of the alternative.

5. (a) I need to obtain the values of μ and σ. First, $c = 2(66)(34) = 4{,}488$. Thus,

$$\mu = 1 + 4488/100 = 45.88 \text{ and } \sigma = \sqrt{\frac{4488(4388)}{100^2(99)}} = 4.460.$$

I need the area under the N(45.88,4.460) curve to the left of 31.5. This area is 0.0006, a very small P-value. Katie obtained many fewer runs than expected under the null hypothesis.

(b) The approximate P-value for the alternative $>$ and the test statistic V is 0.1536. There is evidence in support of the alternative, but it is not convincing.

(c) The approximate P-value for the alternative $>$ and the test statistic W is 0.0019. There is very strong evidence in support of the alternative; nine is a very long run of failures in a sequence of 100 trials, of which 66 are successes.

Table 11.11: The partial approximate sampling distribution of V, the longest run of successes, in 100 Bernoulli trials, conditional on a total of 59 successes.

v:	12	13	14	15
Relative Freq. $(V \geq v)$	0.0457	0.0242	0.0134	0.0064

Table 11.12: The partial approximate sampling distribution of W, the longest run of failures, in 100 Bernoulli trials, conditional on a total of 41 failures.

w:	3	4	5	6	7	8
Relative Freq. $(W \geq w)$	0.9982	0.8559	0.4890	0.2062	0.0770	0.0274

11.8 Homework Problems

1. We plan to observe $n = 600$ Bernoulli trials with $p = 0.40$. Let X denote the total number of successes in the 600 trials.

   (a) Calculate the probability that the first four trials yield: a success, two failures and a success, in that order.

   (b) Calculate the probability that the first four trials yield a total of exactly two successes.

   (c) Use the website

   `http://stattrek.com/Tables/Binomial.aspx`

   To obtain:

   $$P(X \leq 255), P(X \leq 230) \text{ and } P(231 \leq X \leq 255).$$

   (d) Use the Normal curve to obtain approximate answers for the three probabilities of interest in part (c).

2. Mimic what I did in Tables 11.5 and 11.6 for the 15 possible arrangements of four 1's and two 0's.

3. Recall day 7 of Katie's study of shooting, Example 11.2 and the description immediately before it. In this problem we will analyze Katie's data from day 8.

   Here are some important statistics from Katie's 100 trials on day 8:

   $$x = 59, r = 42, v = 13 \text{ and } w = 4.$$

   (a) Use the Normal curve approximation to obtain the P-value for the runs test and the alternative $<$. Comment.

(b) I performed a simulation study with 10,000 reps. Partial results for the test statistic V are in Table 11.11. Use this information to find the approximate P-value for the test statistic V and the alternative $>$.

(c) I performed a simulation study with 10,000 reps. Partial results for the test statistic W are in Table 11.12. Use this information to find the approximate P-value for the test statistic W and the alternative $>$.

Chapter 12

Inference for a Binomial p

In Part 1 of these *Course Notes* you learned a great deal about statistical tests of hypotheses. These tests explore the unknowable; in particular, whether or not the Skeptic's Argument is true. In Section 11.4, I briefly introduced you to three tests that explore whether or not a sequence of dichotomous trials are Bernoulli trials. In this chapter, we will assume that we have Bernoulli trials and turn our attention to the value of the parameter p. Later in this chapter we will explore a statistical test of hypotheses concerning the value of p. First, however, I will introduce you to the inference procedure called **estimation**. I will point out that for Bernoulli trials, estimation is inherently much more interesting than testing.

The estimation methods in this chapter are relatively straightforward. This does not mean, however, that the material will be *easy*; you will be exposed to several new ways of thinking about *things* and this will prove challenging.

After you complete this chapter, however, you will have a solid understanding of the two *types* of inference that are used by scientists and statisticians: testing and estimation. Most of the remainder of the material in these *Course Notes* will focus on introducing you to new scientific scenarios, and then learning how to test and estimate in these scenarios. (In some scenarios you also will learn about the closely related topic of prediction.) Thus, for the most part, after this chapter you will have been exposed to the major ideas of this course, and your remaining work, being familiar, should be easier to master.

12.1 Point and Interval Estimates of p

Suppose that we plan to observe n Bernoulli Trials. More accurately, we plan to observe n *dichotomous* trials and we are willing to assume—for the moment, at least—that the assumptions of Bernoulli trials are met. Throughout these *Course Notes*, unless I state otherwise, we always will assume that the researcher knows the value of n.

Before we observe the n Bernoulli trials, if we know the numerical value of p, then we **can compute probabilities** for X, the total number of successes that will be observed. If we do not know that numerical value of p, then we **cannot compute probabilities** for X. I would argue—not everyone agrees with me—that there is a *gray area* between these extremes; refer to my example

concerning basketball player Kobe Bryant on page 264 of Chapter 11; i.e., if I have a massive amount of previous data from the process that generates my future Bernoulli trials, then I might be willing to use the proportion of successes in the massive data set as an approximate value of p.

Still assuming that the numerical value of p is unknown to the researcher, **after** n Bernoulli trials are observed, if one is willing to condition on the total number of successes, then one can critically examine the assumption of Bernoulli trials using the methods presented in Section 11.4. **Alternatively**, we can use the data we collect—the observed value x of X—to make an inference about the unknown numerical value of p. Such inferences will always involve some uncertainty. To summarize, if the value of p is unknown a researcher will attempt to infer its value by looking at the data. It is convenient to create *Nature*—introduced in Chapter 8 in the discussion of Table 8.8—who knows the value of p.

The simplest inference possible involves the idea of a point estimate/estimator, as defined below.

Definition 12.1 (Point estimate/estimator.) *A researcher observes n Bernoulli trials, counts the number of successes, x and calculates $\hat{p} = x/n$. This proportion, $\hat{p}$, is called the* **point estimate** *of p. It is the observed value of the random variable $\hat{P} = X/n$, which is called the* **point estimator** *of p. For convenience, we write $\hat{q} = 1 - \hat{p}$, for the proportion of failures in the data; $\hat{q}$ is the observed value of the random variable $\hat{Q} = 1 - \hat{P}$.*

Before we collect data, we focus on the random variable, the point **estimator**. **After** we collect data, we compute the value of the point **estimate**, which is, of course, the observed value of the point estimator.

I don't like the technical term, *point estimate/estimator*. More precisely, I don't like half of it. I like the word *point* because we are talking about a single number. (I recall the lesson I learned in math years ago, "Every number is a point on the number line and every point on the number line is a number.") I *don't particularly like* the use of the word estimate/estimator. If I become tsar of the Statistics world, I might change the terminology. I say *might* instead of *will* because, frankly, I can't actually suggest an improvement on estimate/estimator. I recommend that you simply remember that estimate/estimator is a word statisticians use whenever they take observed data and try to infer a feature of a population.

It is trivially easy to calculate $\hat{p} = x/n$; thus, based on experiences in previous math courses, you might expect that we will move along to the next topic. But we won't. In a Statistics course we *evaluate the behavior* of a procedure. What does this mean? Statisticians evaluate procedures by seeing how they perform *in the long run*.

We say that the point estimate $\hat{p}$ is **correct** if, and only if, $\hat{p} = p$. Obviously, any honest researcher wants the point estimate to be correct. As we will see now, whereas having a correct point estimate is desirable, the concept has some serious difficulties.

Let's suppose that a researcher observes $n = 100$ Bernoulli trials and counts a total of $x = 55$ successes. Thus, $\hat{p} = 55/100 = 0.55$ and this point estimate is correct if, and only if, $p = 0.55$. This leads us to the first *difficulty* with the concept of being correct.

- Nature knows whether $\hat{p}$ is correct; the researcher never knows.

The above example takes place **after** the data have been collected. We can see this because we are told that a total of $x = 55$ successes were counted. Now let's go back in time to **before** the data are collected **and** let's take on the role of Nature. I will change the scenario a bit to avoid confusing this current example with what I just did. As Nature, I am aware that a researcher plans to observe $n = 200$ Bernoulli trials. I also know that $p = 0.600$, but the researcher does not know this. In addition, after collecting the data, the researcher will calculate the point estimate of p. What will happen? I don't know what will happen—I don't make Nature omniscient; it just knows the value of p! When I don't know what will happen, I resort to calculating probabilities. In particular, as Nature, I know that $\hat{p}$ will be correct if, and only if, the total number of successes turns out to be 120—making $\hat{p} = 120/200 = 0.600$. Thus, back in time before data are collected, I want to calculate

$$P(X = 120) \text{ given that } X \sim \text{ Bin}(200, 0.600).$$

I can obtain this exact probability quite easily from the website

`http://stattrek.com/Tables/Binomial.aspx`.

I used this website and obtained $P(X = 120) = 0.0575$. Thus, in addition to the fact that only Nature knows whether a point estimate is correct, we see that

- The probability that the point estimator will be correct can be very small and, indeed, can be calculated by Nature, but not the researcher.

I don't want to dwell on this too much, but we need something better than point estimation!

In Section 10.2 I extended the saying,

> Close counts in horseshoes and hand grenades

to

> Close counts in horseshoes, hand grenades and probabilities.

I want to extend it again; this time to

> **Close counts in horse shoes, hand grenades, probabilities and estimation.**

Let's revisit my last example: a researcher plans to observe $n = 200$ Bernoulli trials; the researcher does not know the value of p; and Nature knows that $p = 0.600$. The researcher plans to compute the point estimate of p. We saw above that the probability that the point estimator will be correct—i.e., that $\hat{p}$ will be exactly equal to $p = 0.600$ is small; indeed only 0.0575. **Suppose** now that the researcher thinks, "In order for me to be happy, I don't really need to have my point estimate be exactly equal to p; all I need is for $\hat{p}$ to be *close* to p." In order to proceed, the researcher needs to specify *how close* is required for happiness. I will look at two examples.

1. The researcher decides that *within 0.04* is close enough for happiness. Thus, Nature knows that the event $(0.560 \leq \hat{P} \leq 0.640)$ is the event that the researcher will be happy. (Paradoxically, of course, the researcher won't know that this is the *happiness event*!) Nature, being good at algebra, writes

$$P(0.560 \leq \hat{P} \leq 0.640) = P(112 \leq X \leq 128), \text{ for } X \sim \text{ Bin}(200, 0.600).$$

Next,

$$P(112 \leq X \leq 128) = P(X \leq 128) - P(X \leq 111).$$

With the help of the website

`http://stattrek.com/Tables/Binomial.aspx`,

this probability equals

$$0.8906 - 0.1103 = 0.7803.$$

2. The researcher decides that *within 0.07* is close enough for happiness. Thus, Nature knows that the event $(0.530 \leq \hat{P} \leq 0.670)$ is the event that the researcher will be happy. Nature writes

$$P(0.530 \leq \hat{P} \leq 0.670) = P(106 \leq X \leq 134), \text{ for } X \sim \text{ Bin}(200, 0.600).$$

Next,

$$P(106 \leq X \leq 134) = P(X \leq 134) - P(X \leq 105).$$

With the help of the website

`http://stattrek.com/Tables/Binomial.aspx`,

this probability equals

$$0.9827 - 0.0188 = 0.9639.$$

The above ideas lead to the following definition.

Definition 12.2 (Interval estimate/estimator.) *A researcher observes n Bernoulli trials and counts the number of successes, x. An interval estimate of p is a closed interval with endpoints l (for lower bound) and u (for upper bound), written $[l, u]$. Although the dependence is often suppressed, both l and u are functions of x. Thus, more properly, an interval estimate should be written as $[l(x), u(x)]$. An interval estimate is the observed value of the interval estimator: $[l(X), u(X)]$.*

Below is an example of an interval estimate of p. It is called a **fixed-width interval estimate** because, as you will see, its width is constant; i.e., its width is not a function of the random variable X.

- Define the interval estimate to be $[\hat{p} - 0.04, \hat{p} + 0.04]$. Note that

$$l(x) = \hat{p} - 0.04 = x/n - 0.04 \text{ and } u(x) = \hat{p} + 0.04 = x/n + 0.04;$$

thus, this is a bona fide interval estimate. The width of this interval is

$$u - l = x/n + 0.04 - (x/n - 0.04) = 0.08.$$

Thus, this is a fixed-width interval estimate with width equal to 0.08. Usually, however, statisticians refer to this as a fixed-width interval estimate with **half-width** equal to 0.04.

Recall, we say that the point estimate $\hat{p}$ is **correct** if, and only if, p is equal to $\hat{p}$. Similarly, we say that an interval estimate is **correct** if, and only if, p lies in the interval; i.e., if, and only if, $l \leq p \leq u$.

Let's look at this notion of correctness with our fixed-width interval estimate with half-width equal to 0.04. The interval estimate is correct if, and only if,

$$\hat{p} - 0.04 \leq p \leq \hat{p} + 0.04.$$

I will need to rearrange the terms in this expression which contains two *inequality signs*. If you are good at this activity, you will find my efforts below to be a bit tedious because I will break the above into two pieces; analyze the pieces separately; and then put the pieces back together. In particular, let's start with

$$p \leq \hat{p} + 0.04 \text{ which becomes } p - 0.04 \leq \hat{p}.$$

Similarly,

$$\hat{p} - 0.04 \leq p \text{ becomes } \hat{p} \leq p + 0.04.$$

Combining these inequalities, we obtain

$$p - 0.04 \leq \hat{p} \leq p + 0.04.$$

This last expression, in words, means that $\hat{p}$ is within 0.04 of p. This implies that a researcher who would be happy to have the point estimate be within 0.04 of p, should estimate p with interval estimate with half-width equal to 0.04; the interval is correct if, and only if, the researcher is happy.

Sadly, fixed-width interval estimates have a serious weakness for statistical inference; details will be given in one of the Practice Problems for this chapter. At this time we turn our attention to the type of interval estimate that is very useful in inference and science.

12.2 The (Approximate) 95% Confidence Interval Estimate

In this section you will be introduced to a particular type of interval estimate of p, called a **confidence interval estimate**.

This is a tricky topic. I want to derive the confidence interval for you, but experience has taught me that the topic is very confusing if I **begin** with the derivation. Thus, instead I will give you the formula and use it twice before I derive it.

Recall the definition of an interval estimate presented in Definition 12.2. In order to specify an interval estimate, I must give you formulas for l and u, the lower and upper bounds of the interval.

Result 12.1 (The (approximate) 95% confidence interval estimate of p.) *The lower and upper bounds of the (approximate) 95% confidence interval estimate of p are*

$$l(x) = \hat{p} - 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}} \text{ and } u(x) = \hat{p} + 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}}. \tag{12.1}$$

Because of the similarity between the formulas for l and u, we usually combine the above into one formula. The (approximate) 95% confidence interval estimate of p is

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}} \tag{12.2}$$

Let me make a few comments about this definition.

1. It will be convenient to give a symbol for the half-width of an interval estimate. We will use h. With this notation, the half-width of the (approximate) 95% confidence interval estimate of p is

   $$h = 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}}.$$

   Note that this is **not** a constant half-width; the value of h depends on x through the value of $\hat{p}$ (remembering that $\hat{q} = 1 - \hat{p}$).

2. The confidence interval is centered at $\hat{p}$. It is correct—includes p—if, and only if, $\hat{p}$ is within h of p.

3. The formula for the confidence interval is mysterious. The derivation I give later will clear up the mystery, especially the presence of the magic number 1.96. As you might have guessed, the appearance of 1.96 in the formula is tied to the specification of 95% confidence. Note that if I replace 1.96 in Formula 12.2 by the number 3, we get the nearly certain interval introduced in Chapter 4.

I will now illustrate the computation of the 95% confidence interval for two data sets similar to my earlier example with Kobe Bryant.

1. In his NBA career, Karl Malone attempted 13,188 free throws during games and made 9,787 of them. On the assumption that Malone's game free throws were Bernoulli trials, calculate the 95% confidence interval for his p.

   Solution: Note that unless $\hat{p}$ is close to zero, my convention in these *Course Notes* is to round $\hat{p}$ to three digits. We compute $\hat{p} = 9{,}787/13{,}188 = 0.742$ and $\hat{q} = 1 - 0.742 = 0.258$. Thus,

   $$h = 1.96\sqrt{\frac{0.742(0.258)}{13{,}188}} = 0.007.$$

   Thus, the approximate 95% confidence interval estimate of p is

   $$0.742 \pm 0.007 = [0.735, 0.749].$$

   This interval is very narrow. If it is indeed correct, then we have a very precise notion of the value of p.

2. In his NBA career, Shaquille O'Neal attempted 11,252 free throws during games and made 5,935 of them. On the assumption that O'Neal's game free throws were Bernoulli trials, calculate the 95% confidence interval for his p.

   Solution: We compute $\hat{p} = 5{,}935/11{,}252 = 0.527$ and $\hat{q} = 1 - 0.527 = 0.473$. Thus,

$$h = 1.96\sqrt{\frac{0.527(0.473)}{11{,}252}} = 0.009.$$

   Thus, the approximate 95% confidence interval estimate of p is

$$0.527 \pm 0.009 = [0.518, 0.536].$$

   This interval is very narrow. If it is indeed correct, then we have a very precise notion of the value of p. Note that O'Neal's interval is a bit wider than Malone's; as we will see later, this difference is due to: O'Neal's n is smaller than Malone's; and O'Neal's $\hat{p}$ is closer to 0.5 than Malone's.

12.2.1 Derivation of the Approximate 95% Confidence Interval Estimate of p

Our derivation involves the computation of probabilities; thus, we go back in time to before data are collected. Our basic random variable of interest is X, the number of successes that will be obtained in the n future Bernoulli trials. We know that the sampling distribution of X is Bin(n, p). Of course, we don't know the value of p, but let's not worry about that. Much of algebra, as you no doubt remember, involves manipulating unknown quantities!

The reason our confidence interval includes the modifier approximate is that we are not going to work with exact binomial probabilities; instead, we will approximate the Bin(n, p) distribution by using the Normal curve with $\mu = np$ and $\sigma = \sqrt{npq}$. I want to obtain an answer that is true for a variety of values of n and p; as a result, it will prove messy to constantly have our approximating curve change; e.g., if I change from $n = 100$ to $n = 200$, then the μ and σ of the approximating Normal curve will change. It is more convenient to instead **standardize** the random variable X, as now described. Define a new random variable, denoted by Z, which we call the **standardized version** of X. For a general X—i.e., not just binomial—we define

$$Z = \frac{X - \mu}{\sigma}, \tag{12.3}$$

where μ is the mean and σ is the standard deviation of the random variable X. In this chapter—i.e., because X has a binomial distribution—Equation 12.3 becomes

$$Z = \frac{X - np}{\sqrt{npq}}. \tag{12.4}$$

It can be shown mathematically—although I won't demonstrate it—that the Normal curve with $\mu = np$ and $\sigma = \sqrt{npq}$ approximates the binomial distribution of X exactly the same as the N(0,1)

curve approximates the distribution of Z. In other words, the conditions for the first approximation to be good are exactly the conditions for the second approximation to be good. Recall also that the general guideline I gave you for using a Normal curve to approximate a binomial is that both np and nq should equal or exceed 25. Finally, recall that whereas everybody agrees that the values of np and nq are critical, not everybody agrees with my threshold of 25.

It turns out that for the goal of interval estimation, the unknown p (and $q = 1 - p$) in the denominator of Z creates a major difficulty. Thanks, however, to an important result of Eugen Slutsky (1925) (called *Slutsky's Theorem*) probabilities for Z',

$$Z' = \frac{(X - np)}{\sqrt{n\hat{P}\hat{Q}}},$$

can be well approximated by the N(0,1) curve, provided n is reasonably large; p is not too close to 0 or 1; and $0 < \hat{P} < 1$ (we don't want to divide by zero). Note that Z' is obtained by replacing the unknown p and q in the denominator of Z with the random variables $\hat{P}$ and $\hat{Q}$, both of which will be replaced by their observed values once the data are collected.

Here is the derivation. Suppose that we want to calculate $P(-1.96 \leq Z' \leq 1.96)$. Because of Slutsky's result, we can approximate this probability with the area under the N(0,1) curve between -1.96 and 1.96. Using the website,

`http://davidmlane.com/hyperstat/z_table.html`

you can verify that this area equals 0.95. Next, dividing the numerator and denominator of Z' by n gives

$$Z' = \frac{\hat{P} - p}{\sqrt{\hat{P}\hat{Q}/n}}.$$

Thus,

$$-1.96 \leq Z' \leq 1.96 \text{ becomes } -1.96 \leq \frac{\hat{P} - p}{\sqrt{\hat{P}\hat{Q}/n}} \leq 1.96;$$

rearranging terms, this last inequality becomes

$$\hat{P} - 1.96\sqrt{\hat{P}\hat{Q}/n} \leq p \leq \hat{P} + 1.96\sqrt{\hat{P}\hat{Q}/n}.$$

Examine this last expression. Once we replace the random variables $\hat{P}$ and $\hat{Q}$ by their observed values $\hat{p}$ and $\hat{q}$, the above inequality becomes

$$\hat{p} - 1.96\sqrt{\hat{p}\hat{q}/n} \leq p \leq \hat{p} + 1.96\sqrt{\hat{p}\hat{q}/n}.$$

In other words,

$$l \leq p \leq u.$$

Thus, we have shown that, before we collect data, the probability that we will obtain a correct confidence interval estimate is (approximately) 95% and that this is true for all values of p! Well,

all values of p for which the Normal curve and Slutsky approximations are good. We will return to the question of the quality of the approximation soon.

Let me say a bit about the use of the word *confidence* in the technical expression *confidence interval*. First and foremost, remember that I use *confidence* as a technical term. Thus, whatever the word *confidence* means to you in every day life is **not necessarily relevant**. Let's look at the 95% confidence interval I calculated for Karl Malone's p for free throw shooting. I am 95% confident that

$$0.735 \leq p \leq 0.749.$$

Literally, this is a statement about the value of p. This statement might be correct or it might be incorrect; only my imaginary creature Nature knows. Here is the key point. I *assign 95% confidence* to this statement *because of the method I used to derive it*. **Before** I collected Malone's data I knew that I would calculate the 95% confidence interval for p. **Before** I collected data I knew that the probability I would obtain a correct confidence interval is (approximately) 95%. By appealing to the Law of Large Numbers (Subsection 10.2.1), I know that as I go through life, observing Bernoulli trials and calculating the approximate 95% confidence interval from each set of said data, in the long run approximately 95% of these intervals will be correct. Thus, a particular interval—such as mine for Malone—might be correct or it might be incorrect. Because, in the long run, 95% of such intervals will be correct, I am 95% **confident** that my particular interval is correct.

12.2.2 The Accuracy of the 95% Approximation

Later, I will give you a specific general guideline for when to use the approximate confidence interval as well as an alternative method to be used if the guideline is not satisfied. In the current subsection I will focus on *how we assess the accuracy of the approximation*. I will do this with several examples. Essentially, we must specify values of both n and p and then see how the formula performs.

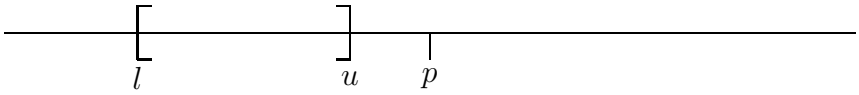
Before I get to my first example, it is convenient to have a not-so-brief digression. I want to introduce you to what I call the **Goldilocks metaphor**, a device that repeatedly will prove useful in these notes.

According to Wikipedia,

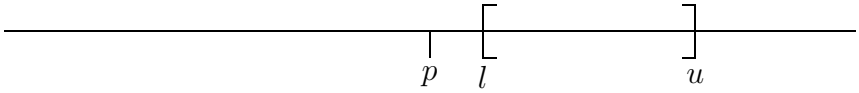
> *The Story of the Three Bears* (sometimes known as *The Three Bears*, *Goldilocks and the Three Bears* or, simply, *Goldilocks*) is a fairy tale first recorded in narrative form by British author and poet Robert Southey, and first published anonymously in a volume of his writings in 1837. The same year, British writer George Nicol published a version in rhyme based upon Southey's prose tale, with Southey approving the attempt to bring the story more exposure. Both versions tell of three bears and an old woman who trespasses upon their property. . . . Southey's intrusive old woman became an intrusive little girl in 1849, who was given various names referring to her hair until Goldilocks was settled upon in the early 20th century. Southey's three bachelor bears evolved into Father, Mother, and Baby Bear over the course of several years. What was originally a fearsome oral tale became a cozy family story with only a hint of menace.

Figure 12.1: The Goldilocks metaphor for confidence intervals.

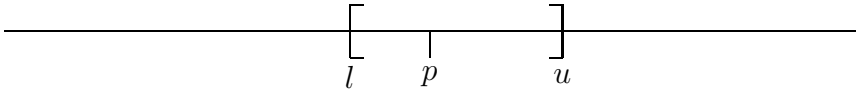
The CI is too small, $u < p$:



The CI is too large, $p < l$:



The CI is correct, $l \leq p \leq u$:



In my opinion, Goldilocks, a juvenile delinquent specializing in home invasion, gets her comeuppance when she stumbles into the wrong house. In any event, Goldilocks is well-known for complaining that something was *too hot* or *too cold*, before setting on something that was *just right*.

So, what does any of this have to do with confidence intervals? Only that it is useful to realize that a confidence interval can be *too small*, *too large* or *correct* (just right!). Perhaps a picture will help. Figure 12.1 presents the three possibilities for the relationship between a confidence interval and the p it is estimating. Let's look at the three pictured possibilities.

1. The confidence interval could be too small. This means that p is larger than every number in the confidence interval. It will be convenient to note that a confidence interval is too small if, and only if, $u < p$.

2. The confidence interval could be too large. This means that p is smaller than every number in the confidence interval. It will be convenient to note that a confidence interval is too large if, and only if, $p < l$.

3. The confidence interval could be correct. This means that $l \leq p \leq u$.

The main message of these three observations is: In general, it is easier to determine whether a confidence interval is too small or too large rather than correct. This is because determining either of the former requires checking one inequality, whereas determining the latter requires checking two inequalities.

For my first example, I will take $n = 200$ and $p = 0.500$. I anticipate that the interval should perform well, because both $np = 200(0.5) = 100$ and $nq = 200(0.5) = 100$ are much larger than our Chapter 11 guideline threshold of 25 for using a Normal curve to approximate binomial probabilities. We have a very specific criterion that we want to examine. We want to determine the exact probability that the 95% confidence interval will be correct. If you desire, you may verify the following facts, but you don't need to; i.e., I will never ask you to perform such an activity on an exam.

- The event *the confidence interval is too small* is the event $(X \leq 86)$; i.e., for any $x \leq 86$, the value of u is less than 0.500.

- The event *the confidence interval is too large* is the event $(X \geq 114)$; i.e., for any $x \geq 114$, the value of l is greater than 0.500.

- In view of the previous two items, the event *the confidence interval is correct* is the event $(87 \leq X \leq 113)$.

With the help of the website

`http://stattrek.com/Tables/Binomial.aspx`,

we find

$$P(X \leq 86) = 0.0280 \text{ and } P(X \geq 114) = 0.0280; \text{ thus,}$$

$$P(87 \leq X \leq 113) = 1 - 2(0.0280) = 1 - 0.0560 = 0.9440.$$

Actually, I am a bit disappointed in this approximation. In the limit (long run), 94.4%, not the advertised 95.0%, of the confidence intervals will be correct.

For my second example, I will take n = 1,000 and $p = 0.600$. For this example, $np = 1000(0.6) = 600$ and $nq = 1000(0.4) = 400$ are both substantially larger than the threshold value of 25. If you desire, you may verify the following facts:

- The event *the confidence interval is too small* is the event $(X \leq 568)$; i.e., for any $x \leq 568$, the value of u is less than 0.600.

- The event *the confidence interval is too large* is the event $(X \geq 631)$; i.e., for any $x \geq 631$, the value of l is greater than 0.600.

- In view of the previous two items, the event *the confidence interval is correct* is the event $(569 \leq X \leq 630)$.

With the help of the website

`http://stattrek.com/Tables/Binomial.aspx`,

we find

$$P(X \leq 568) = 0.0213 \text{ and } P(X \geq 631) = 0.0241; \text{ thus,}$$

$$P(569 \leq X \leq 630) = 1 - (0.0213 + 0.0241) = 1 - 0.0454 = 0.9546.$$

In this example, in the limit (long run), the nominal 95% confidence interval is correct a bit more often than promised.

For my third and final example, I will take $n = 100$ and $p = 0.020$. For this example, $np = 100(0.02) = 2$, which is far below the threshold value of 25. Thus, I anticipate that our approximate confidence interval will not perform as advertised. If you desire, you may verify the following facts:

- The event *the confidence interval is too small* is the event $(X = 0)$; i.e., for $x = 0$, the value of u is less than 0.02. In fact, for $x = 0$ the confidence interval is $[0, 0]$, a single point! Also, for $(1 \leq x \leq 3)$ the lower bound, l, of the confidence interval is a negative number! Whenever an interval reduces to a single number or a nonnegative quantity (in the current set-up p) is stated to be possibly negative, it's a good indication that the formula being used can't be trusted!

- The event *the confidence interval is too large* is the event $(X \geq 9)$; i.e., for any $x \geq 9$, the value of l is greater than 0.020.

- In view of the previous two items, the event *the confidence interval is correct* is the event $(1 \leq X \leq 8)$.

With the help of the website

`http://stattrek.com/Tables/Binomial.aspx`,

we find

$$P(X = 0) = 0.1326 \text{ and } P(X \geq 9) = 0.0001; \text{ thus,}$$

$$P(1 \leq X \leq 8) = 1 - (0.1326 + 0.0001) = 1 - 0.1327 = 0.8673.$$

In this example, in the limit (long run), the nominal 95% gives way too many incorrect intervals.

We will revisit my third example in Section 12.3 and you will learn a method that performs as it promises.

In summary, the approximate 95% confidence interval for p (Formula 12.2) is one of the most useful results in Statistics. For its computation, we don't need access to the internet; we don't need a fancy calculator; all we need is a calculator that can compute square roots. If both np and nq are 25 or larger, then the actual probability that the confidence interval will be correct is indeed reasonably close to the advertised (nominal) value of 95%. Admittedly, this last sentence is quite vague, but it will suffice for a first course in introductory Statistics.

You may have noticed a flaw in the part of my advice that requires both np and nq to be 25 or larger. Do you see it? The whole point of estimation is that we don't know the value of p or q and, thus, we can't actually check the values of np and nq. There are two ways we handle this.

1. Sometimes n is so large that even though I can't literally check the values of np and nq I am quite sure that they both exceed 25. For example, I don't know who will be running for President of the United States in 2016, but if I have a random sample of $n = 1{,}000$ voters, and the dichotomy is vote Democrat or Republican, I am quite sure that both $np = 1000p$ and $nq = 1000q$ will be much larger than 25.

2. If you aren't sure that the above item applies, a popular—and valuable—guideline is to use Formula 12.2 provided that both x and $(n - x)$ equal or exceed 35. Note that 35 is my personal choice; other statisticians might consider me to be too cautious (they advocate a smaller threshold) or too reckless (they advocate a larger threshold).

12.2.3 Other Confidence Levels

In our 95% confidence interval, the number 95% is called the **confidence level** of the interval. The obvious question is: Can we use some other confidence level? The answer is "Yes, and I will show you how in this short subsection."

If you think back to my derivation of the 95% confidence interval formula, you will recall that my choice of 95% gave us the magic number of 1.96. In particular, the Normal curve website told us that the area under the N(0,1) curve between the numbers -1.96 and $+1.96$ is equal to 0.95. You can verify that the area under the N(0,1) curve between the numbers -1.645 and $+1.645$ is equal to 0.90. Thus, we immediately know that the approximate 90% confidence interval estimate of p is

$$\hat{p} \pm 1.645\sqrt{\frac{\hat{p}\hat{q}}{n}}.$$

We can summarize our two confidence intervals—95% and 90%—by writing them as

$$\hat{p} \pm z^*\sqrt{\frac{\hat{p}\hat{q}}{n}}, \tag{12.5}$$

with the understanding that for 95% [90%] confidence we substitute 1.96 [1.645] for z^*. Let me make two comments on Formula 12.5.

1. There is no need to restrict ourselves to 95% or 90%. The most popular choices for confidence level—and their corresponding values of z^* are provided in Table 12.1. Note that you now know that the nearly certain interval for p is, indeed, the 99.73% confidence interval estimate of p.

2. Many texts—I hate to be rude, but frankly— are sadistic in their presentation of the general confidence interval, Formula 12.5. In particular, instead of our user-friendly z^*, they write something like

   $$z^*_{\alpha/2}$$

   (usually without the asterisk) and refer to the result as the $100(1 - \alpha)\%$ confidence interval estimate of p. I prefer my method; seeing z^* reminds you that you need to make a choice of confidence level and that the number you use, z^*, depends on your choice.

Table 12.1: Popular choices for the confidence level and their corresponding values of z^* for the general approximate confidence interval estimate of p, Formula 12.5.

Confidence Level	80%	90%	95%	98%	99%	99.73%
z^*:	1.282	1.645	1.960	2.326	2.576	3.000

I will discuss the choice of confidence level in Section 12.4. Let me just state that earlier I examined the quality of the approximation and gave guidelines on whether or not to use the 95% confidence interval formula. All of my results are *qualitatively the same* for other confidence levels. In particular, my guideline for using Formula 12.5 is the same as it was for the 95% confidence level: if both x and $(n-x)$ equal or exceed 35, I recommend using it. This will be a general pattern in our ongoing studies of confidence intervals. When I explore properties of the intervals, I will focus on 95% and the results will **always be qualitatively the same** for other confidence levels.

12.3 The Clopper and Pearson "Exact" Confidence Interval Estimate of p

In Section 12.4, I will explain why I put the word *Exact* in quotes in the title of this section.

In line with the last paragraph of the previous section, let me summarize the facts about the approximate 95% confidence interval estimate of p; remembering that similar comments are true for other choices of the confidence level.

Formula 12.2 has the following property. **For every possible value of** p the probability that the researcher will obtain a correct confidence interval is approximately 95%. The fact that the formula works for every possible p is pretty amazing; there are, after all, an infinite number of possibilities for p! The word *approximately* is, however, troublesome. We saw by example, that sometimes the approximation can be bad. In particular, if either np or nq is smaller than 25 then I recommend that you do not use Formula 12.2. Because the values of p and q are unknown, my more useful recommendation is that if, after you collect data, you notice that either x or $(n-x)$ is smaller than 35, then I recommend that you do not use Formula 12.2.

Let's look at the origin of Formula 12.2 again. The approach was as follows. For any fixed value of p, we found numbers b and c such that $P(b \leq X \leq c)$ is approximately 95%, where the approximation is based on using a Normal curve. Next, because both b and c are functions of p we were able to manipulate the event $(b \leq X \leq c)$ into a confidence interval for p. This last part, recall, required help from Slutsky's theorem too. Thus, our confidence interval is based on two approximations: using a Normal curve to approximate binomial probabilities and then using Slutsky's theorem.

There is an obvious alternative approach. Let's find the numbers b and c without using an approximating Normal curve; obtain them by using exact binomial probabilities. **If** we can *invert* this collection of inequalities—a big if because there are an infinite number of inequalities—then we will have a confidence interval for p that does not involve any approximations. In 1934, Clopper

Table 12.2: The Clopper and Pearson (CP) 95% confidence intervals for $n = 20$.

x	$[l(x), u(x)]$	x	$[l(x), u(x)]$	x	$[l(x), u(x)]$	x	$[l(x), u(x)]$
0:	$[0, 0.168]$	6:	$[0.119, 0.543]$	11:	$[0.315, 0.769]$	16:	$[0.563, 0.943]$
1:	$[0.001, 0.249]$	7:	$[0.154, 0.592]$	12:	$[0.360, 0.809]$	17:	$[0.621, 0.968]$
2:	$[0.012, 0.317]$	8:	$[0.191, 0.640]$	13:	$[0.408, 0.846]$	18:	$[0.683, 0.988]$
3:	$[0.032, 0.379]$	9:	$[0.231, 0.685]$	14:	$[0.457, 0.881]$	19:	$[0.751, 0.999]$
4:	$[0.057, 0.437]$	10:	$[0.272, 0.728]$	15:	$[0.509, 0.913]$	20:	$[0.832, 1]$
5:	$[0.087, 0.491]$						

and Pearson managed to make this alternative approach work. (See *Binomial proportion confidence interval* in Wikipedia for more information.) Well, it almost works. The main sticking problem is that because of the discrete nature of the binomial distribution, for a given p we cannot, in general, find numbers b and c so that the binomial $P(b \leq X \leq c)$ is *exactly* equal to 0.95. Instead, they settled on finding numbers b and c so that

$$P(b \leq X \leq c) \geq 0.95, \text{ for every value of } p.$$

(Historical note: Working prior to the computer-age, the accomplishment of Clopper and Pearson was quite amazing. Their work has been improved in recent years because while their choices of b and c were good for inverting the infinite number of inequalities, for many values of p, their exact $P(b \leq X \leq c)$ is much larger than 0.95. As we will see later, this means that their intervals were wider—and, hence, less informative—than necessary. I won't show you the modern improvement on Clopper and Pearson because it is not easily accessible computationally.)

As you probably know, there was no internet in 1934; in fact, as best I can tell there were no computers in 1934. Thus, Clopper and Pearson distributed their work by creating lots of tables. An example of a Clopper and Pearson table is given in Table 12.2. This table presents all of the Clopper and Pearson (CP) 95% confidence intervals for $n = 20$. Let's look at a few of the entries in Table 12.2.

If we observe a total of $x = 7$ successes in 20 Bernoulli trials, then the CP 95% confidence interval estimate of p is: $[0.154, 0.592]$. If $x = 15$, the confidence interval is $[0.509, 0.913]$. Note that for all 21 possible values of x, the CP 95% confidence intervals are very wide; in short, we don't learn much about the value of p with only 20 Bernoulli trials.

Next, let's do a couple of computations to verify that the probability that a CP 95% confidence interval will be correct is at least 95%. Let's consider $p = 0.500$. From Table 12.2, we can quickly ascertain the following facts and you should be able to verify these.

- The CP confidence interval is too small ($u < 0.500$) if, and only if, $(X \leq 5)$.
- The CP confidence interval is too large ($0.500 < l$) if, and only if, $(X \geq 15)$.
- The CP confidence interval is correct ($l \leq 0.500 \leq u$) if, and only if, $(6 \leq X \leq 14)$.

With the help of the website

`http://stattrek.com/Tables/Binomial.aspx`,

we find that for $n = 20$ and $p = 0.500$,

$P(X \leq 5) = 0.0207, P(X \geq 15) = 0.0207$ and, thus, $P(6 \leq X \leq 14) = 1-2(0.0207) = 0.9586$.

The probability that the CP interval will be correct does, indeed, achieve the promised minimum of 0.95.

Let's do one more example, for $n = 20$ and $p = 0.200$. From Table 12.2, we can quickly ascertain the following facts:

- The CP confidence interval is too small ($u < 0.200$) if, and only if, $(X = 0)$.
- The CP confidence interval is too large ($0.200 < l$) if, and only if, $(X \geq 9)$.
- The CP confidence interval is correct ($l \leq 0.200 \leq u$) if, and only if, $(1 \leq X \leq 8)$.

With the help of the website

`http://stattrek.com/Tables/Binomial.aspx`,

we find that for $n = 20$ and $p = 0.200$,

$$P(X = 0) = 0.0115, P(X \geq 9) = 0.0100 \text{ and, thus,}$$

$$P(1 \leq X \leq 8) = 1 - (0.0115 + 0.0100) = 0.9785.$$

The probability that the CP interval will be correct does, indeed, achieve the promised minimum of 0.95. In fact, the probability of being correct is quite a bit larger than the nominal 95%.

The obvious question is: Suppose you want to obtain a CP confidence interval for p, but your number of trials n is not 20. Before the internet you would have needed to find a CP table for your value of n. The method we use now is introduced after the next example.

Example 12.1 (Mahjong solitaire online.) *My friend Bert loves to play mahjong solitaire online. (See Wikipedia if you want details of the game.) Each game ends with Bert winning or losing. He played $n = 100$ games, winning a total of 29 of the games.*

Let's analyze Bert's data. Clearly, the trials yield a dichotomous response: a win (success) or a loss (failure). Are we willing to assume that they are Bernoulli trials? I have two remarks to make on this issue:

1. The game claims that it randomly selects an arrangement of tiles for each game. (Pieces in mahjong are called tiles; they look like dominoes. Well, more accurately, online tiles look like pictures of dominoes.) Of course, there might be something about the way Bert performs that violates the assumptions of Bernoulli trials: perhaps he improved with practice; perhaps his skills declined from boredom; perhaps he had streaks of better or worse skill.

2. I looked for patterns in Bert's data: his 100 trials contained 41 runs; his longest run of successes [failures] had length 3 [14]. In the first [last] 50 games he won 16 [13] times. We will examine these statistics together in a Practice Problem. For now, let's assume that we have Bernoulli trials.

There exists a website that will give us the CP 95% confidence interval estimate of p; it is:

`http://statpages.org/confint.html`

I will now explain how to use this site.

First of all, **do not scroll down this page**. A bit later we will learn the benefits of scrolling down, but don't do it yet! You will see a box next to **Numerator (x)**; enter the total number of successes in this box—for Bert's data, enter 29. Next, you will se a box next to **Denominator (N)**; enter the value of n in this box—for Bert's data, enter 100. Click on the box labeled *Compute*. The site produces three numbers for us, the value of $\hat{p}$ and the lower and upper bounds of the CP interval:

- **Proportion (x/N):** For Bert's data, we get $\hat{p} = 29/100 = 0.29$.
- **Exact Confidence Interval around Proportion:** For Bert's data we get 0.2036 to 0.3893.

For comparison, let's see the answer we obtain if we use the 95% confidence interval based on the Normal curve approximation and Slutsky's theorem:

$$0.2900 \pm 1.96\sqrt{\frac{0.29(0.71)}{100}} = 0.2900 \pm 0.0889 = [0.2011, 0.3789].$$

These two confidence intervals are very similar. As a practical matter, I cannot think of a scientific problem in which I would find these answers to be importantly different.

Recall that on page 294 we looked at an example with $n = 100$ and $p = 0.020$. We found that the approximate 95% confidence interval was correct—included $p = 0.020$—if, and only if, $(1 \leq x \leq 8)$. We further found that

$$P(1 \leq X \leq 8 | p = 0.020) = 0.8673,$$

which is much smaller than the target of 0.95. Thus, for this combination of n and p the approximate 95% confidence interval performs poorly and should not be used. Let's see what happens if we use the CP 95% confidence interval.

The long answer is for me to create a table of all CP 95% confidence intervals for $n = 100$, as I reported in Table 12.2 for $n = 20$. If I do that, I obtain the following results. The CP interval is never too small; it is too large if, and only if, $x \geq 6$; and it is correct if, and only if, $x \leq 5$. With the help of

`http://stattrek.com/Tables/Binomial.aspx`,

I find that

$$P(X \leq 5 | n = 100 \text{ and } p = 0.020) = 0.9845.$$

Thus, the CP interval performs as advertised—this probability actually exceeds the target 0.95—in the same situation in which the approximate confidence interval performs very poorly.

12.3.1 Other Confidence Levels for the CP Intervals

Let's return to the site

`http://statpages.org/confint.html`

and now let's scroll down. Scroll past the section headed **Poisson Confidence Intervals** all the way to the section headed **Setting Confidence Levels**, below which you will see the following display:

Confidence Level:	95
% Area in Upper Tail:	2.5
% Area in Lower Tail:	2.5

The three numbers above—95, 2.5 and 2.5—are the default values for the confidence level. The first number, 95, tells us that the default confidence level for the site is 95%. It is important to note that the site does not want the % sign, nor does it want 95% written as a decimal. It wants 95. Similarly, the complement of 95% is 5%; equally divided 5 gives 2.5 twice; these numbers appear in the *Upper* and *Lower* rows.

If you want, say, 90% confidence instead of the default 95%, no worries. The easiest way to accomplish this is to replace the default 95 by 90 (not 90%, not 0.90) and click on the compute box. When you do this you will note that the site automatically changes both the *Upper* and *Lower* rows entries to 5. If you now scroll back up the page to the **Binomial Confidence Intervals** section, you will see that your entries 29 and 100 have not changed. If you now click on the box *Compute* you will be given a new confidence interval: 0.2159 to 0.3737—this is the CP 90% confidence interval estimate of p.

12.3.2 The One-sided CP Confidence Intervals

Both the approximate and CP confidence intervals of this chapter are two-sided. They provide both an upper and a lower bound on the value of p. Sometimes a scientist wants only one bound; the bound can be either upper or lower and there are approximate methods as well as methods derived from the work of Clopper and Pearson. A one-semester class cannot possibly present an *exhaustive* view of introductory Statistics; thus, I will limit the presentation to the upper confidence bound that can be obtained using the Clopper and Pearson method.

Before I turn to a website for answers, I want to create a table that is analogous to our Table 12.2, the two-sided CP confidence intervals for $n = 20$. Table 12.3 presents the Clopper and Pearson 95% upper confidence bounds for p for $n = 20$. Let's compare the one- and two-sided 95% intervals for p for $n = 20$ and a couple of values of x.

- For $x = 19$, the two-sided interval states $0.751 \leq p \leq 0.999$; and the one-sided interval states $p \leq 0.997$
- For $x = 1$, the two-sided interval states $0.001 \leq p \leq 0.249$; and the one-sided interval states $p \leq 0.216$

Table 12.3: The Clopper and Pearson (CP) 95% upper confidence bounds for p for $n = 20$.

x	$[l(x), u(x)]$	x	$[l(x), u(x)]$	x	$[l(x), u(x)]$	x	$[l(x), u(x)]$
0:	$[0, 0.139]$	6:	$[0, 0.508]$	11:	$[0, 0.741]$	16:	$[0, 0.929]$
1:	$[0, 0.216]$	7:	$[0, 0.558]$	12:	$[0, 0.783]$	17:	$[0, 0.958]$
2:	$[0, 0.283]$	8:	$[0, 0.606]$	13:	$[0, 0.823]$	18:	$[0, 0.982]$
3:	$[0, 0.344]$	9:	$[0, 0.653]$	14:	$[0, 0.860]$	19:	$[0, 0.997]$
4:	$[0, 0.401]$	10:	$[0, 0.698]$	15:	$[0, 0.896]$	20:	$[0, 1]$
5:	$[0, 0.456]$						

The most obvious thing to note is that for $x = 19$—which is fairly likely to occur if p is close to 1—then computing a one-sided upper bound for p is ridiculous. For $x = 1$, however, the one-sided upper bound might well be preferred to the two-sided interval. The two-side interval rules out the possibility that $p < 0.001$, but at the cost of having an upper bound that is $0.249/0.216 = 1.15$ times as large as the upper bound for the one-sided interval.

The computations above are insightful, but what does *science* tell me to do? In my experience, sometimes what we call a success can be a very nasty outcome. For example, a success might be that a biological item is infected; that a patient dies; or that an asteroid crashes into the Earth. In such situations, we are really hoping that p—if not zero—will be very small. When we estimate p it might well be more important to have a **sharp** upper bound on p rather than have a scientifically rather uninteresting lower bound on p.

In any event, I will now show you how to use the website

`http://statpages.org/confint.html`

to obtain the CP one-sided 95% upper confidence bound for p. Scroll down to the **Setting Confidence Levels** section. Enter 5 in the *Upper* box and 0 in the *Lower* box and click on *Compute*. The entries in the *Upper* and *Lower* boxes—i.e., 5 and 0, respectively—will remain unchanged, but the entry in the *Confidence Level* box will become 95. Similarly, if you want the one-sided 90% upper confidence bound for p, repeat the steps above, but put 10 in the *Upper* box.

After you have made your entries in the **Setting Confidence Levels** section, scroll back up to the top, enter your data and click on compute. You should practice this activity a few times by making sure you can obtain my answers in Table 12.3.

12.4 Which Should You Use? Approximate or CP?

In this section I will *tie-up various loose ends* related to confidence interval estimation and eventually give you my answer to the question in its title.

The obvious question is:

> Given that the probability of obtaining a correct interval when using the CP 95% confidence interval always equals or exceeds 0.95. Given that the approximate method

cannot make this claim, why do people ever use the approximate method?

The CP method is a *black box*, as discussed in Chapter 3; it gives us answers, but little or no insight into the answers. In a particular problem, the CP interval we obtain is a function of three values: n; x or $\hat{p}$; and the confidence level. We could, literally, vary these values and obtain hundreds of CP intervals and **not see** how the answers are related. As I stated in Chapter 7, when introducing you to fancy math approximations:

> Being educated is **not** about acquiring lots and lots of facts. It is more about seeing how lots and lots of facts *relate to each other* or *reveal an elegant structure in the world*. Computer simulations are very good at helping us acquire *facts*, whereas fancy math helps us see how these facts fit together.

The above sentiment is relevant in this chapter, if we replace *computer simulations* by *CP intervals*. Indeed, the approximate confidence intervals of this chapter are *fancy math* solutions. Thus, I will now turn to a discussion of what the approximate confidence intervals reveal.

The approximate confidence intervals are centered at the value of $\hat{p}$. Another way to say this is that these intervals are symmetric around $\hat{p}$. This symmetry is a direct result of the fact that the approximating curve we use—a Normal curve—is symmetric. By contrast, if you look at the 95% CP intervals for $n = 20$ that are presented in Table 12.2, you see that they are **not** symmetric around $\hat{p} = x/20$ **except** when $x = 10$, giving $\hat{p} = 0.50$. In fact, as x moves away from 10, in either direction, the CP intervals become more asymmetrical. This is a direct result of the fact that for $p \neq 0.50$ the binomial distribution is not symmetric and it becomes more skewed as p moves farther away from 0.50.

Because the approximate confidence intervals are symmetric around the value $\hat{p}$, we learn *how they behave* by looking at the half-width,

$$h = z^* \sqrt{\frac{\hat{p}\hat{q}}{n}}. \tag{12.6}$$

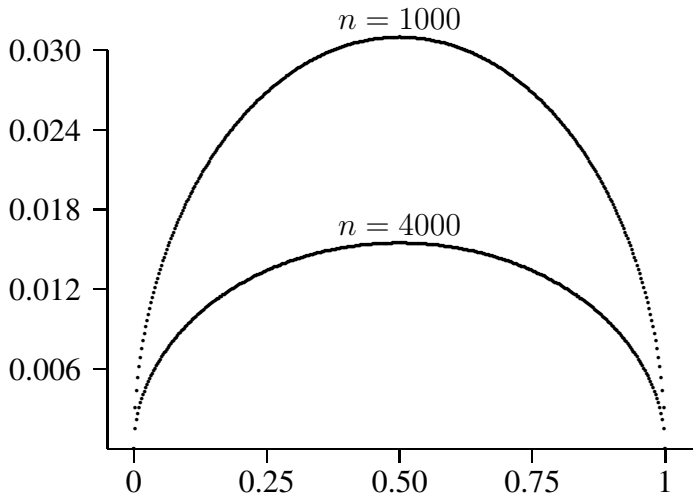
There are three numbers in this formula for h that we can vary: z^*, which is determined by the choice of confidence level; n; and $\hat{p}$, remembering that $\hat{q} = 1 - \hat{p}$. My first effort is to fix the confidence level at 95%—i.e. $z^* = 1.96$—and see how h varies as a function of n and $\hat{p}$. We can—and will—investigate this issue analytically—i.e., by doing algebra—but it is helpful to first draw a picture, which I have done in Figure 12.2, and which I will now discuss.

There are two curves in this figure; one for $n = 1{,}000$ and one for $n = 4{,}000$. I will begin by focusing on either of these curves; i.e., by fixing the value of n and looking at the effect of $\hat{p}$ on h. Analytically, I note that $\hat{p}$ affects h through the term

$$\sqrt{\hat{p}\hat{q}} \text{ which I will write as } \sqrt{\hat{p}(1-\hat{p})}.$$

Visually, the figure shows me that the curve of h values to the right of 0.50 is the mirror image of the curve of h values to the left of 0.50; i.e., the curve of h values is symmetric around 0.50. This fact is obvious analytically because whether we replace $\hat{p}$ by b, $0 < b < 1$ or by $(1 - b)$, the value of $\sqrt{\hat{p}(1-\hat{p})}$ is the same.

Figure 12.2: Plot of the half-width, h in Equation 12.6, versus $\hat{p}$ for the approximate 95% confidence interval estimate of p for $n = 1{,}000$ and $n = 4{,}000$.



The figure does a good job of showing us that the curve of h values is rather flat for $\hat{p}$ close to 0.50. For example, we have the following values:

$\sqrt{\hat{p}(1-\hat{p})}$	0.458	0.490	0.500	0.490	0.458
$\hat{p}$	0.30	0.40	0.50	0.60	0.70

This table—and the figure—indicates that for $\hat{p}$ between 0.40 and 0.60, the actual value of $\hat{p}$ will affect h by at most 2% (0.490 is 2% smaller than 0.500). This fact is very useful for most sample surveys—there is not much interest in asking questions unless there is a good controversy; i.e., unless there is a roughly equal split in the population between the possible responses. Thus, before performing a survey, for a large n a researcher might be quite sure that $\hat{p}$ will take on a value between 0.40 and 0.60; if the surety comes to pass it means that before collecting data the researcher has a good idea what the half-width—and, hence, the usefulness—of the interval will be.

In science, we are often interested in estimating p's that are very different from 0.500 and, hence, we often obtain values for $\hat{p}$ that are outside the range of 0.400 to 0.600. Thus, the central flatness of the curve of h values is not as interesting.

Next, let's look at the effect of n on the half-width.

- For **any** fixed value of $\hat{p}$, changing n from 1,000 to 4,000 results in h being halved.

This fact is true for any n, as can be seen analytically. In general, consider two possibilities for n: m and $4m$; i.e., the smaller sample size can be anything and the larger is four times as large. For $n = 4m$, we find

$$h = 1.96\sqrt{\frac{\hat{p}\hat{q}}{4m}} = \frac{1.96}{2}\sqrt{\frac{\hat{p}\hat{q}}{m}},$$

because when we factor a 4 out of a square root sign we get $\sqrt{4} = 2$. We can see that this argument remains true for any fixed value of z^*, not just $z^* = 1.96$. I am tempted to say,

> If we quadruple the amount of data we collect, the half-width of the approximate confidence interval is halved.

I cannot literally say this because if I quadruple the amount of data collected, the value of $\hat{p}$ will likely change. If, however, I believe that both of my $\hat{p}$'s will be in the interval $[0.400, 0.600]$, then, from our earlier work, we know that a change in $\hat{p}$ will have only a minor affect on h. Thus, the statement

> If we quadruple the amount of data we collect, the half-width of the approximate confidence interval is halved

will be reasonably accurate.

Finally, let's look at the effect of changing the confidence level, i.e., changing the value of z^* in the formula for the half-width. Our guidance for this issue is contained in our table relating confidence level to z^*, namely Table 12.1 on page 296. First, we see that as the confidence level increases, z^* increases and, thus, the half-width h increases. This makes sense: In order to increase the probability of obtaining a correct confidence interval, we must make the interval wider; i.e., a more general statement about p has a better chance of being correct. There is a striking feature about this relationship that can be overlooked. Namely, if we take the ratio of two h's; one for 99% confidence and the other for 80% confidence, we get

$$2.576/1.282 = 2.01,$$

because the other terms in h cancel. Thus, in words, the price we pay for increasing confidence from 80% to 99% is that the half-width increases by a factor of approximately 2. This can be counteracted—sort of, see above—by quadrupling the sample size.

I want to end this section with a comment about the label *Exact* that is popularly affixed to the CP confidence intervals. This label is actually quite misleading. The CP 95% intervals are usually referred to as the **exact** 95% confidence intervals for p. Indeed, the title across the top of the website we use claims that it provides **Exact Binomial and Poisson Confidence Intervals**. Based on **everything** we did in Part I of this book, indeed also based on all that we have done so far in Part II, **exact should mean that the probability that the researcher will obtain a correct confidence interval is exactly equal to 0.95.** What is weird about these being called exact intervals is that statisticians have a perfectly good technical term to describe the truth about the CP intervals: We say that the CP 95% intervals are **conservative**. By *conservative* I don't mean that these are the preferred intervals of Michele Bachman—although they might be, I am not familiar with Ms. Bachman's views on Statistics—nor am I trying to conjure up memories of Ronald Reagan. Saying that the CP 95% intervals are **conservative** conveys that the target probability is 95% and, *no matter what the value of* p, the true probability will be at least 95%. We saw, by example, that even when an approximate confidence interval performs well, its performance is not necessarily conservative. For example, if the approximate method gives an actual probability of 94.9% for a particular p, then in my opinion the approximation is very good, but not conservative. Similarly, if a CP 95%

interval has true probability of 99% I would not be happy, but it is conservative. I would not be happy because if the true probability is 99%, the interval must be wider than if we could somehow make the true probability closer to 95%.

Here is the idea I would like you to remember. We might choose 95% confidence because *everybody else does*, but we should remember what it means. When we select 95% confidence, we are telling the world that we have decided that we are happy with intervals that are incorrect about one time in every 20 intervals. If, instead, the intervals are incorrect about one time for every 100 intervals, then we are seriously out-of-touch with the actual performance of our conclusions; this cannot be a good thing!

Finally, the intervals are called *exact* not because they give exact probabilities (or confidence levels) but because they use exact binomial probabilities for their derivation.

12.5 A Test of Hypotheses for a Binomial p

.

We immediately have a problem. How do we specify the null hypothesis? Let me explain why this is a problem. In Part I of these notes we had an obvious choice for the null hypothesis: the Skeptic is correct. I say that this choice was obvious for two reasons.

1. If a scientist is comparing two treatments by performing a CRD, it is natural to *at least wonder* whether the Skeptic is correct.

2. Following the principle of Occam's Razor, given that we wonder about the Skeptic being correct, it should be the null hypothesis.

I **cannot** create a similar argument for a study of a binomial p. Here is what I can do. Suppose that out of all the possible values of p—i.e., all numbers between 0 and 1 exclusive—there is one possible value for p for which I have a *special interest*. I will denote this special value of interest by the symbol p_0. (The reason for a subscript of zero will soon be apparent.)

Let's be clear about this. The symbol p with no subscript represents the true probability of success for the Bernoulli trials. It is unknown to the researcher, but known to Nature. By contrast, p_0 is a known number; indeed, it is specified by the researcher as being the singular value of p that is special. How does a researcher decide on this *specialness*? Be patient, please.

The null hypothesis specifies that p is equal to the special value of interest; i.e.,

$$H_0 : p = p_0.$$

Our test in this section allows three possibilities for the alternative hypothesis:

$$H_1 : p > p_0; H_1 : p < p_0 \text{ or } H_1 : p \neq p_0.$$

As in Part I of these notes, you could use the Inconceivable Paradigm to select the alternative. Actually, for the applications in this section, I will take the alternative to be the one-sided alternative of most interest to the researcher.

If you go back to the beginning of this chapter, the last sentence in the first paragraph reads:

I will point out that for Bernoulli trials, estimation is inherently much more interesting than testing.

Now I can say why or *point out* why. If I am a researcher and I don't know the value of p then I will **always** be interested in obtaining a confidence interval estimate of p. I will, however, be interested in a test of hypotheses **only if there exists in my mind a special value of interest** for p. In my experience, it is somewhat unusual for a researcher to have a special value of interest for p.

Let me digress for a moment before I show you the details of the test of hypotheses of this section. In many ways, the test is almost scientifically useless. Almost, but not quite. Thus, there is a little bit of value in your knowing it. The value of the test is not sufficient for your valuable time *except* that it provides a relatively painless introduction to tests of hypotheses for population-based inference. You need to see this introduction at some point in these notes, so it might as well be now.

I will introduce the remainder of the test very mechanically and then end with the only applications of it that I consider worthwhile.

12.5.1 The Test Statistic, its Sampling Distribution and the P-value

The only random variable we have is X, the total number of successes in the n Bernoulli trials; thus, it is our test statistic. The sampling distribution of X is Bin(n, p_0) because if the null hypothesis is true, then $p = p_0$. The three rules for computing the exact P-value are given in the following result.

Result 12.2 *In the formulas below, $X \sim$ Bin(n, p_0) and x is the actual observed value of X.*

1. *For the alternative $p > p_0$, the exact P-value equals*

$$P(X \geq x) \tag{12.7}$$

2. *For the alternative $p < p_0$, the exact P-value equals*

$$P(X \leq x) \tag{12.8}$$

3. *For the alternative $p \neq p_0$, the exact P-value is a bit tricky.*

   - *If $x = np_0$, then the exact P-value equals one.*
   - *If $x > np_0$, then the exact P-value equals*

$$P(X \geq x) + P(X \leq 2np_0 - x) \tag{12.9}$$

   - *If $x < np_0$, then the exact P-value equals*

$$P(X \leq x) + P(X \geq 2np_0 - x) \tag{12.10}$$

The above result is all we need *provided* $n \leq 1000$ *and* we have access to the website

`http://stattrek.com/Tables/Binomial.aspx`.

Another approach is to use the Normal curve approximation, as detailed in the following result.

Result 12.3 *Let $q_0 = 1 - p_0$. Assume that both np_0 and nq_0 equal or exceed 25. In the rules below, when I say* **area to the right [left] of***, I am referring to areas under the Normal curve with mean $\mu = np_0$ and standard deviation $\sigma = \sqrt{np_0q_0}$. Also, x is the actual observed value of X.*

1. *For the alternative $p > p_0$, the Normal curve approximate P-value equals the area to the right of $(x - 0.5)$.*
2. *For the alternative $p < p_0$, the Normal curve approximate P-value equals the area to the left of $(x + 0.5)$.*
3. *For the alternative $p \neq p_0$, the situation is a bit tricky.*
   - *If $x = np_0$, then the exact P-value equals one.*
   - *If $x \neq np_0$:*
     - *Calculate the area to the right of $(x - 0.5)$; call it b.*
     - *Calculate the area to the left of $(x + 0.5)$; call it c.*

     The Normal curve approximate P-value is the minimum of the three numbers: $2b$, $2c$ and 1.

Let's now turn to the question: How does a researcher choose the value p_0. The textbooks I have seen claim that there are three possible scenarios for choosing the special value of interest; they are: history; theory; and contractual or legal. I will consider each of these possibilities in turn.

12.5.2 History as the Source of p_0

I won't be able to hide my contempt; so I won't even try. History as the source of p_0 is almost always dumb or dangerous. (Indeed, **every example I have seen** of this type is bad. I am being generous by allowing for the possibility that there *could* be a good example.)

The basic idea of the history justification goes as follows. Let's say that we are interested in a finite population, for example all 28 year-old men currently living in the United States. For some reason, we are interested in the proportion of these men, p, who are married. We don't know what p equals, but somehow we know the proportion, p_0, of 28 year-old men living in the United States in 1980 who were married! The goal of the research is to compare the current men with the same age group on 1980. Thus, we would be interested in the null hypothesis that $p = p_0$. I have seen numerous textbooks that have problems just like this one. I am amazed that an author could type such a ludicrous scenario! Do you really believe that someone conducted a **census** of all 28 year-old men in the United States in 1980? (Note: After typing this it occurred to me that the United States did conduct a census in 1980. The problem, however, is that the US census suffers from an undercount—how could it not? The idealized census as used in these notes is perfect in that it samples every population member.)

My final example of this subsection is one that I have seen in many textbooks. It is not only dumb, but dangerous. It is dangerous because it promotes a really bad way to do science. A textbook problem reads as follows. The current treatment for disease B will cure 40% of the persons to whom it is given. Researchers have a new treatment. The researchers select $n = 100$ persons at random from the population of those who suffer from disease B and give the new treatment to each of these persons. Let p denote the proportion of the population that would be cured with the new treatment. The researcher want to use the data they will collect to test the null hypothesis that $p = 0.40$. Think about this problem for a few moments. Do you see anything wrong with it?

The first thing I note is that it's a total fantasy to say that we ever know exactly what percentage of people will be cured with a particular treatment. But suppose you disagree with me; suppose you think I am way too cynical and feel that while the cure rate for the existing treatment might not be exactly 40%, pretending that it is 40% seems relatively harmless. Even if you are correct and I am wrong, this is still a horribly designed study! Why do I say this?

The key is in the statement:

> The researchers select $n = 100$ persons at random from the population of those who suffer from disease B.

I opine that this statement has never been literally true in any medical study. (Can you explain why?) It is very possible that the actual method used by the researchers to select subjects for study resulted in either *better than average* patients—which would skew the results in the new treatment's favor— or *worse than average* patients—which would skew the results in favor of the existing treatment. Even if the researchers *got lucky* and obtained *average* patients, good luck to them in trying to convince the scientific community to believe it!

Phrasing the medical situation as a *one population problem* is bad science. It would be better to take the 100 subjects—better yet, have 200 subjects—and divide them into two treatment groups by randomization. Then analyze the data using Fisher's test from Chapter 8 or a population-based procedure that will be presented in a later chapter.

12.5.3 Theory as the Source of p_0

Zener cards were popular in the early twentieth century for investigating whether a person possesses extra sensory perception (ESP). Each Zener card had one of five shapes—circle, cross, waves, square or star—printed on it. There were various protocols for using the Zener cards. The protocol I will talk about is available on a website:

`http://www.e-tarocchi.com/esptest/index.php`

I request that you take a break from this fascinating exposition and test yourself for ESP. Click on the site above. When you arrive at the site click on the *Test Me* box and take the 25 item exam.

Let's match the above to the ideas of this chapter and section and focus on a point in time *before* you take the ESP exam. You plan to observe 25 dichotomous trials. Each trial will be a success if you correctly identify the shape chosen by the computer and a failure if you don't. I will assume

that your trials are Bernoulli trials and I will denote your probability of success by p. I don't want to prejudge your psychic skills; thus, I do not know the value of p, Of all the possible values of p, however, there is definitely one possibility that is of special interest to me. Can you determine what it is? (By the way, if you can correctly determine my special value of interest, this is **not** an indication of ESP!) My special value of interest is $p_0 = 1/5 = 0.20$ because *if you are guessing* then the probability you guess correctly is one-in-five. In other words, my choice of p_0 follows from my **theory** that you are guessing.

Thus, I select the null hypothesis $p = 0.20$. Although we could debate the choice of alternative I will use $p > 0.20$. From Equation 12.7 in Result 12.2, if you score x correct, the exact P-value is

$$P(X \geq x), \text{ where } X \sim \text{Bin}(25, 0.20).$$

So, what is your P-value? Well, since I have not figured out how to make these notes interactive, I cannot respond to your answer. Thus, I will tell you how I did. I scored $x = 6$ correct responses in $n = 25$. I went to the website

`http://stattrek.com/Tables/Binomial.aspx`.

and found that the exact P-value for my data is:

$$P(X \geq 6) = 0.3833.$$

12.5.4 Contracts or Law as the Source of p_0

Company B manufactures tens of thousands of widgets each month. Any particular widget can either work properly or be defective. Because defectives are rare, we label a defective widget a success. By contract or by law Company B is required to manufacture no more than 1% defective widgets.

Suppose we have the ability to examine 500 widgets in order to investigate whether the contract/law is being obeyed. How should we do this? Before we collect data, let's assume that we will be observing $n = 500$ Bernoulli trials with unknown probability of success p. I don't know what p equals, but I am particularly interested in the value 0.01, because 0.01 is the threshold between the manufacturing process being fine and being in violation of the contract/law. I take my null hypothesis to be $p = 0.01$, following the philosophy that it seems fair to begin my process by assuming the company is in compliance with the law/contract. In problems like this one, the alternative is always taken to be $p > p_0$ because, frankly, there is not much interest in learning that $p < p_0$—unless we are trying to decide whether to give Company B an award for good corporate citizenship! Note that my stated goal above was to investigate whether the company is obeying the law/contract. I don't want to accuse the company of misbehaving unless I have strong evidence to that effect.

From Equation 12.7 in Result 12.2, if the sample of 500 widgets yields x defectives, the exact P-value is

$$P(X \geq x), \text{ where } X \sim \text{Bin}(500, 0.01).$$

Below are some possibilities for the exact P-value.

x :	5	6	7	8	9	10
$P(X \geq x)$:	0.5603	0.3840	0.2371	0.1323	0.0671	0.0311

I will now tie this example to the idea of a critical region and the concept of power, introduced in Chapters 8 and 9.

Recall that a critical region is a rule that specifies all values of the test statistic that will lead to rejecting the null hypothesis. The critical regions $(X \geq 10)$ is the rule we get if we follow classical directive to reject the null hypothesis if, and only if, the P-value is 0.05 or smaller. If one uses this critical region, we see that the probability of making a Type 1 error is:

$$\alpha = P(\text{Reject } H_0 | H_0 \text{ is true}) = P(X \geq 10 | p = p_0 = 0.01) = 0.0311.$$

Now that we have determine the critical region, we can investigate the power of the test. With the help of the binomial website, I obtained:

$$P(X \geq 10 | p = 0.015) = 0.2223; P(X \geq 10 | p = 0.02) = 0.5433;$$

$$P(X \geq 10 | p = 0.03) = 0.9330; \text{ and } P(X \geq 10 | p = 0.04) = 0.9956.$$

Let me briefly interpret these four powers.

If, in fact, $p = 0.015$—i.e., a defective rate 50% larger than allowed by law/contract—there is only about 2 chances in 9 (0.2223) that the test will detect it. If $p = 0.02$, then the chance of correctly rejecting the null climbs to a bit more than 54%. If $p = 0.03$, the probability of detecting such a large violation of the law/contract is extremely high, 93.30%. Finally, if $p = 0.04$, it is almost certain—a 99.56% chance—that the test will detect the violation of the contract/law.

12.6 Summary

A researcher plans to observe n Bernoulli trials and doesn't know the value of p. The researcher wants to use the data that will be obtained to **infer** the value of p. Late in this chapter we explore the use of a familiar method—a statistical test of hypotheses—as an inference procedure for p. Most of this chapter, however, is focused on introducing you to a new method of inference, **estimation**.

As with our earlier work involving probabilities, it is important to distinguish the time **before** the data are collected from the time **after** the data are collected. Before the data are collected, there is a random variable X, the total number of successes that **will be** obtained in the n Bernoulli trials. After the data are collected, the researcher will know the observed value, x, of X.

The notion of a point estimate/estimator is a natural starting point for inference. After the data are collected, x is used to calculate the value of $\hat{p} = x/n$. This single number is called the **point estimate** of p; the name is somewhat suggestive: the word *point* reminds us that we are dealing with *one number* and the word *estimate* is, well, how statisticians refer to this activity. The point estimate $\hat{p}$ is called **correct** if, and only if, $\hat{p} = p$.

Having a correct point estimate is a good thing, but because the researcher does not *know* the value of p, he/she will not know whether the point estimate is correct. To avoid having this all become too abstract, it is convenient for me to reintroduce our supernatural friend **Nature**, first

introduced in Chapter 8. In this current chapter Nature knows the value of p. Thus, Nature—but not the researcher—will know whether a particular point estimate is correct.

Let's travel back in time to before the data are collected. The researcher announces, "After I collect data I will calculate the point estimate $\hat{p}$." In symbols, the researcher is interested in the random variable $\hat{P} = X/n$, which we call the **point estimator** of p. Note the distinction, the point estimator is a random variable $\hat{P}$ that will take on observed value $\hat{p}$, the point estimate.

Thus, before the data are collected, Nature—but, again, not the researcher—can calculate the probability that the *point estimator will be correct*. Taking the role of Nature, we looked at one specific possibility ($n = 200$ and $p = 0.600$) and found that this probability is very small. We could have looked at many more examples and, except for quite uninteresting situations, the probability that a point estimator will be correct is very small. (By uninteresting I mean situations for which n is very small. For example, if $n = 2$ and p happens to equal exactly 0.5, then there is a 50% probability that $\hat{p} = p$.) The lesson is quite clear: we need something more sophisticated than point estimation.

Thus, I introduced you to the notion of an interval estimate/estimator. The first type of interval estimate/estimator—the so-called fixed width interval—is intuitively appealing—but, as you will see in Practice Problem 1 of this chapter, is unsatisfactory in terms of the probability that it is correct.

Next, you learned about the approximate 95% confidence interval estimate of p. This interval is really quite amazing. Before collecting data it can be said that for any value of p, the probability that the 95% confidence interval that **will be obtained** is correct is approximately 95%. The lone flaw—and it is serious—is that for this approximation to be good, both np and nq must equal or exceed 25. I give an example with $n = 100$ and $p = 0.02$—hence, $np = 2$ is much smaller than the magic threshold of 25—and show that the probability that the 95% confidence interval will be correct is only 86.73%.

I introduce you to a misnamed *exact* method developed by Clopper and Pearson in 1934 with the property that for all values of p **and** n the probability that the Clopper and Pearson 95% confidence interval will be correct is 95% or larger.

In this chapter, you learned how to extend both the approximate and exact confidence interval estimates to levels other than 95%. Also, you learned how to obtain the Clopper and Pearson upper confidence bound for p, which is very useful if you believe that p is close to zero.

Section 12.4 explores why—when the approximate method performs well—most researchers prefer it to the Clopper and Pearson conservative intervals. In particular, one can see how the sample size n, the confidence level and the value of $\hat{p}$ influence the half-width of the approximate confidence interval. In contrast, the Clopper and Pearson intervals come from a *black box* and, hence, we cannot see useful patterns in their answers.

Finally, Section 12.5 provides a brief—mostly critical—introduction to a test of hypotheses for the value of p. This problem is not very useful in science, but I want you to be aware of its existence, if only for intellectual completeness. In addition, this test allows us to compute its power quite easily, which is a nice feature.

12.7 Practice Problems

1. Diana plans to observe $n = 100$ Bernoulli trials. She decides to estimate p with a fixed-width interval estimate with half width equal to 0.06. Thus, her interval estimate of p will be $[\hat{p} - 0.06, \hat{p} + 0.06]$.

   Diana wonders, "What is the probability that my interval estimator will be correct?" She understands that the probability might depend on the value of p. Thus, she decides to use the website binomial calculator:

   `http://stattrek.com/Tables/Binomial.aspx`.

   to find the missing entries in the following table.

Actual value of p:	0.03	0.06	0.20
The **event** the interval is correct:	$(0 \leq X \leq 9)$	$(0 \leq X \leq 12)$	$(14 \leq X \leq 26)$
P(The interval is correct$\|p$):			
Actual value of p:	0.30	0.40	0.50
The **event** the interval is correct:	$(24 \leq X \leq 36)$	$(34 \leq X \leq 46)$	$(44 \leq X \leq 56)$
P(The interval is correct$\|p$):			

   Find Diana's six missing probabilities for her and comment.

2. During his NBA career in regular season games, Michael Jordan attempted 1,778 three point shots and made a total of 581.

   Assume that these shots are the result of observing 1,778 Bernoulli trials.

   (a) Calculate the approximate 95% confidence interval for p.

   (b) Calculate the exact 95% confidence interval for p.

3. Example 12.1 introduced you to my friend Bert's data from playing mahjong solitaire online. In my discussion of these data, I promised that we would revisit them in a Practice Problem. I am now keeping that promise.

   This is a different kind of practice problem because none of the things I ask you to do involve Chapter 12.

   (a) Calculate the mean and standard deviation of the null distribution of R. Explain why there is no need to specify an alternative or compute a P-value.

   (b) I performed a 10,000 rep simulation experiment to obtain an approximate sampling distribution for V, the length of the longest run of successes given that the total number of successes equals 29. Recall that for Bert's data, $V = 3$. My experiment yielded:

   $$\text{The relative frequency of } (V \geq 3) = 0.8799.$$

   Comment on this result.

(c) I performed a 10,000 rep simulation experiment to obtain an approximate sampling distribution for W, the length of the longest run of failures given that the total number of failures equals 71. Recall that for Bert's data, $W = 14$. My experiment yielded:

$$\text{The relative frequency of } (W \geq 14) = 0.1656.$$

Bert remarked that he became discouraged while he was experiencing a long run of failures. (He actually had two runs of failures of length 14 during his 100 games.) It was his feeling that being discouraged led to him concentrating less and, thus, perhaps, playing worse. Comment on the simulation result and Bert's feeling.

(d) We can create the following 2×2 table from the information given.

	Outcome			Row Prop.	
Half:	Win	Lose	Total	Win	Lose
First	16	34	50	0.32	0.68
Second	13	37	50	0.26	0.74
Total	29	71	100		

This table looks like the tables we studied in Chapter 8. Bert's data, however, are not from a CRD; games were not assigned by randomization to a half. The first 50 games necessarily were assigned to the first half. As you will learn in Chapter 15, there is a population-based justification for performing Fisher's test for these data. Thus, use the Fisher's test website:

`http://www.langsrud.com/fisher.htm`

to obtain the three Fisher's test P-values for these data.

4. During his NBA career, Shaquille O'Neal attempted a total of 22 three-point shots, making one. Assuming that these shots are 22 observations of Bernoulli trials:

   (a) Calculate the 95% two-sided confidence interval for p.

   (b) Calculate the 95% one-sided upper confidence bound for p.

5. Manute Bol, at 7 feet, 7 inches, is tied (with Gheorghe Muresan) for being the tallest man to ever play in the NBA. Not surprisingly, Bol's specialty was blocking opponents' shots. He was, however, a horrible offensive player. So horrible that he hurt his team because the man guarding him could safely ignore him. In 1988–89, Golden State's innovative coach, Don Nelson, decided to make Bol a three-point shot threat. If nothing else, given the NBA's rules, a defensive player would need to stand near Bol, who would position himself in a corner of the court, just beyond the three-point line. During the 1988–89 season, Bol attempted 91 three points shots, making 20 of them.

   Assume that his attempts are 91 observations of a sequence of Bernoulli trials. Calculate the approximate and exact 95% confidence intervals for Bol's p and comment.

6. Refer to the investigation of ESP using Zener cards presented in Section 12.5.3. Recall that this led to the null hypothesis that $p = 0.20$. I am interested in the alternative $p > 0.20$. Suppose that we decide to test Shawn Spencer, famed psychic police consultant in Santa Barbara, California. (Well, at least in the USA network world.)

   We decide that the study of Shawn will involve $n = 1{,}000$ trials.

   (a) Use the website

   `http://stattrek.com/Tables/Binomial.aspx`.

   to obtain the exact P-value if Shawn scores: $x = 221$ correct; $x = 222$ correct; $x = 240$ correct.

   (b) What is the value of α for the critical region $X \geq 222$?

   (c) Using the critical region in part (b), calculate the power of the test if, indeed, Shawn's p equals 0.23.

   (d) Explain the practical importance of having $p = 0.23$ when guessing Zener cards. (Yes, this is a trick question.)

7. Years ago, I received a gift of two round-cornered dice. Have you ever seen round-cornered dice? Regular dice have pointy squared-corners. When I cast regular dice they bounce and then settle. By contrast, round-cornered dice spin a great deal before coming to a rest. I played with my round-cornered dice a great deal and noticed that both of them seemed to give way too many 6's; I did not record any data, but I had a very strong feeling about it. Thus, with my past experience suggesting that 6 seemed to be special, I decided to perform a formal study.

   I took my white round-cornered die and cast it 1,000 times—call me Mr. Excitement! I will analyze the data as follows. I will assume that the casts are Bernoulli trials, with the outcome '6' deemed a success and any other outcome a failure. I will test the null hypothesis that $p = 1/6$ versus the alternative that $p > 1/6$ because, given my past experience, I felt that $p < 1/6$ is inconceivable. If you don't like my alternative, read on please. My 1,000 casts yielded a total of 244 successes. From Equation 12.7, the P-value for the alternative $>$ is

$$P(X \geq 244 | n = 1000, p = 1/6) = 2.91 \times 10^{-10},$$

   with the help of the website binomial calculator. This is an incredibly small P-value! (Even if you use the alternative $\neq$ and double this probability, it is still incredibly small.) The null hypothesis is untenable.

   By the way, the approximate 99.73% (nearly certain) confidence interval estimate of p is:

$$0.244 \pm 3\sqrt{\frac{0.244(0.756)}{1000}} = 0.244 \pm 0.041 = [0.203, 0.285].$$

   The lower bound of this interval is much larger than $1/6 = 0.167$.

8. I want to give you a bit more information about dumb versus smart sampling. The result of this problem, however, is not important unless you frequently:

   - Select a smart random sample with $n/N > 0.05$.

   In other words, if you choose not to read this problem, no worries.

   The half-width, h, of the approximate confidence interval estimate of p is

   $$h = z^* \sqrt{\frac{\hat{p}\hat{q}}{n}}.$$

   This formula arises from a dumb random sample—which includes Bernoulli trials as a special case—or as an approximation if one has a smart random sample. It turns out that a simple modification of the half-width will handle the situation in which one has a smart random sample **and** does not want to pretend it is a dumb random sample. In this new situation, the half-width, denoted by h_s (s is for smart) is:

   $$h_s = z^* \sqrt{\frac{\hat{p}\hat{q}}{n}} \sqrt{\frac{N-n}{N-1}} = h \times \text{fpc},$$

   where I have implicitly defined fpc—which stands for the **finite population correction**—to equal

   $$\sqrt{\frac{N-n}{N-1}}.$$

   Note the following features of the fpc:

   - If $n = 1$, then fpc $= 1$. This makes sense because if $n = 1$ there is no distinction between smart and dumb sampling. (Of course, if $n = 1$, you would not use the approximate confidence interval formula.)
   - For $n > 1$, fpc < 1; thus, the fpc correction term always leads to a narrower confidence interval; why not use it all the time? Suppose that $N =$ 10,000 and $n = 500$, making $n/N = 0.05$, our threshold value. In this case, fpc equals

     $$\sqrt{9500/9999} = 0.974.$$

     Thus, if you use the fpc, the half-width of the approximate confidence interval will decrease by 2.6%.

     Often times, of course, we don't know the exact value of N, so the fpc cannot be used.

12.8 Solutions to Practice Problems

1. The completed table is below.

Actual value of p:	0.03	0.06	0.20
The **event** the interval is correct:	$(0 \le X \le 9)$	$(0 \le X \le 12)$	$(14 \le X \le 26)$
P(The interval is correct$\|p$):	0.9991	0.9931	0.8973
Actual value of p:	0.30	0.40	0.50
The **event** the interval is correct:	$(24 \le X \le 36)$	$(34 \le X \le 46)$	$(44 \le X \le 56)$
P(The interval is correct$\|p$):	0.8446	0.8157	0.8066

   For example, for $p = 0.20$,

$$P(X \le 26) = 0.9442 \text{ and } P(X \le 13) = 0.0469,$$

$$\text{giving } P(14 \le X \le 26) = 0.9442 - 0.0469 = 0.8973.$$

   My comment: The answers in this table are unsatisfactory. Without knowing the value of p, the probability of a correct interval could be nearly one for $p = 0.03$ or $p = 0.06$ or barely four-in-five, for $p = 0.50$.

2. (a) We compute $\hat{p} = 581/1{,}778 = 0.327$. Thus, $\hat{q} = 0.673$ and the approximate 95% confidence interval is:

$$0.327 \pm 1.96\sqrt{\frac{0.327(0.673)}{1{,}778}} = 0.327 \pm 0.022 = [0.305, 0.349].$$

   (b) The exact website gives $[0.305, 0.349]$, the same answer as the approximate interval.

3. (a) First, from Equation 11.4, we get

$$c = 2x(n - x) = 2(29)(71) = 4118.$$

   From Equation 11.5, we get

$$\mu = 1 + \frac{c}{n} = 1 + (4118/100) = 42.18.$$

   From Equation 11.6, we get

$$\sigma = \sqrt{\frac{c(c-n)}{n^2(n-1)}} = \sqrt{\frac{4118(4018)}{(100)^2(99)}} = 4.088.$$

   The observed number of runs, 41, is almost equal to μ. The Normal curve approximate P-value:

- For $>$ will be larger than 0.5000;
- For $<$ will be quite close to 0.5000 (its actual value: 0.4339); and
- For $\neq$ will be quite close to 1 (its actual value: 0.8678).

Thus, regardless of the choice of alternative, there is only weak evidence in support of it.

(b) With the huge approximate P-value for the test statistic V, there is little reason to doubt the assumption of Bernoulli trials.

(c) The approximate P-value for the test statistic W is small, but not convincing. Perhaps there is some validity to Bert's conjecture that repeated failures adversely affect his ability.

(d) After rounding, the P-values are:

- 0.8109 for the alternative $<$;
- 0.3299 for the alternative $>$; and
- 0.6598 for the alternative $\neq$.

There is very little evidence that Bert's ability changed from the first to the second half of his study.

4. (a) Using the exact confidence interval website:

`http://statpages.org/confint.html`

I obtain $[0.0012, 0.2284]$.

(b) I use the above website, but remember that I need to reset the confidence levels. I put 5 in the **Upper** box and 0 in the **Lower** box and then click on **Compute**. I scroll back up the page and click on **Compute**. The answer is $[0, 0.1981]$.

5. For the approximate confidence interval, I first compute $\hat{p} = 20/91 = 0.220$, which gives $\hat{q} = 0.780$. The interval is

$$0.220 \pm 1.96\sqrt{\frac{0.22(0.78)}{91}} = 0.220 \pm 0.085 = [0.135, 0.305].$$

For the exact confidence interval, the website gives me $[0.140, 0.319]$. The exact interval is a bit wider and is shifted to the right of the approximate interval. The approximate interval seems to be pretty good—because it is similar to the exact—even though $x = 21$ falls far short of my guideline of 35.

6. (a) At the website, we enter $p_0 = 0.20$ as the **Probability of success on a single trial** because we want to compute probabilities on the assumption the null hypothesis is correct. I enter 1,000 for the **Number of trials**. I then enter the various x's values in the question and obtain the following results:

$$P(X \geq 221) = 0.0539; P(X \geq 222) = 0.0459; \text{ and } P(X \geq 240) = 0.0011.$$

(b) Recall that

$$\alpha = P(\text{Rejecting } H_0 | H_0 \text{ is correct});$$

for the critical rule I have given you, this probability is

$$P(X \geq 222 | p = 0.20) = 0.0459 \text{ from part (a).}$$

We can also see from part (a) that if one's goal is to have $\alpha = 0.05$, then this goal cannot be met, but $\alpha = 0.0459$ is the ***Price is Right*** value of α: it comes closest to the target without exceeding it.

(c) I first note that the $p = 0.23$ I ask you to consider is, indeed, part of the alternative hypothesis, $p > 0.20$. Thus,

$$P(X \geq 222 | p = 0.23) = 0.7372, \text{ roughly, 3 out of 4}$$

actually is an example of power. The *power business* is different from the *confidence interval business*; a power of 75% is considered to be quite respectable.

(d) I know of no career in which one gets paid for predicting Zener cards. (Notify me if you know of one.) If Zener-card-based ESP can transfer to gambling or the stock market—a big if—then a person with $p = 0.23$ might be able to make a living.

12.9 Homework Problems

1. In the 1984 Wisconsin Driver Survey, subjects were asked the question:

   > For statistical purposes, would you tell us how often, if at all, you drink alcoholic beverages?

   Each subject was required to choose one of the following categories as the response:

   - Several times a week;
   - Several times a month;
   - Less often; and
   - Not at all.

   Of the n = 2,466 respondents, 330 selected *Several times a week*. If we make the WTP assumption (Definition 10.3) then we may view these data as the result of selecting a random sample from the population of all licensed drivers in Wisconsin. Because n is a very small fraction of N (which, as I opined earlier in these notes, must have been at least one million), we may view these data as the observations from 2,466 Bernoulli trials in which the response *Several times a week* is deemed a success.

   (a) Use these data to obtain the approximate 95% confidence interval estimate of p.

   (b) *In your opinion* what proportion of people would answer this question accurately? (I say accurately instead of honestly because a person's self-perception might not be accurate.) Do you think that giving an accurate answer is related to the response; e.g., are true non-drinkers more or less accurate than those who truly drink several times per week?

   (c) In addition to the 2,466 persons who responded to this question, 166 persons chose not to respond. *Does this extra information* change your interpretation of your answer in part (a)? In other words, do you think that the failure to respond is related to the self-perceived frequency of drinking?

2. Don observes n_1 Bernoulli trials. Later, Tom observes n_2 trials from the same process that Don observed. In other words, Don and Tom are interested in the same p, they have different sets of data and their data sets are independent of each other.

   Don uses his data to construct the approximate 90% confidence interval for p. Tom uses his data to construct both the approximate 95% and approximate 98% confidence intervals for p. The three confidence intervals are:

   $$[0.349, 0.451], [0.363, 0.477] \text{ and } [0.357, 0.443].$$

   (a) Match each interval to its researcher—Don or Tom—and its confidence level.

   (b) Calculate Don's 99% confidence interval for p.

(c) This part is tricky. Find the 95% confidence interval for p for Don's and Tom's combined data. Hint: First, determine the values of n_1 and n_2. There will be round-off error; thus, you may use the fact that the 1's digit for both n's is 0.

3. Recall that my friend Bert enjoys playing mahjong solitaire online, as described in Example 12.1. Bert also plays a second version of mahjong solitaire online, which is much more difficult than the version explored in Example 12.1. For this second version, Bert played 250 games and achieved only 34 victories.

   Assuming that Bert's data are 250 observations of a sequence of Bernoulli trials, calculate the exact and approximate 98% confidence interval for p.

   Here is a side note for those of you who are fascinated with Bernoulli trials or mahjong solitaire online or the lengths of longest runs. For the first version Bert played, the length of his longest run of failures—while not convincing—was a bit long for Bernoulli trials. For this second version, the length of Bert's longest run of failures was $w = 25$, which—according to him—was very frustrating. I performed a 10,000 rep simulation study to investigate this issue and found that 5,799 of my simulated arrangements yielded a value of W that was greater than or equal to 25. Thus, Bert's observed value of W does little to diminish my belief in Bernoulli trials.

   With such a small proportion of successes, looking at either the runs test or the value of V is not likely to be fruitful. In particular, Bert's observed value of V was 3. In a 10,000 rep simulation study, 4,227 arrangements yielded $V \geq 3$; thus, $V = 3$ is not remarkable. Finally, Bert's data had 59 runs, which is almost equal to the mean number of runs under the assumption of Bernoulli trials, 59.752.

4. Refer to Practice Problem number 6, a study of Shawn Spencer's power of ESP. I decided to test his partner, Burton 'Gus' Guster too. There was time to test Gus with only 500 cards. Again, I will assume Bernoulli trials. The null hypothesis is $p = 0.20$ and the alternative is $p > 0.20$.

   (a) Use the website

   `http://stattrek.com/Tables/Binomial.aspx`.

   to obtain the exact P-value if Gus scores: $x = 115$ correct; $x = 116$ correct; $x = 125$ correct.

   (b) What is the value of α for the critical region $X \geq 116$?

   (c) Using the critical region in part (b), calculate the power of the test if, indeed, Gus's p equals 0.23.

   (d) Compare your answer to part (c) to the power for the study of Shawn. Comment.

Chapter 13

The Poisson Distribution

Jeanne Antoinette Poisson (1721–1764), Marquise de Pompadour, was a member of the French court and was the official chief mistress of Louis XV from 1745 until her death. The pompadour hairstyle was named for her. In addition, poisson is French for fish. The Poisson distribution, however, is named for Simeon-Denis Poisson (1781–1840), a French mathematician, geometer and physicist.

13.1 Specification of the Poisson Distribution

In this chapter we will study a family of probability distributions for a countably infinite sample space, each member of which is called a **Poisson distribution**. Recall that a binomial distribution is characterized by the values of two parameters: n and p. A Poisson distribution is simpler in that it has only one parameter, which we denote by θ, pronounced *theta*. (Many books and websites use λ, pronounced lambda, instead of θ. We save λ for a related purpose.) The parameter θ must be positive: $\theta > 0$. Below is the formula for computing probabilities for the Poisson.

$$P(X = x) = \frac{e^{-\theta}\theta^x}{x!}, \text{ for } x = 0, 1, 2, 3, \ldots. \tag{13.1}$$

In this equation, e is the famous number from calculus,

$$e = \lim_{n\to\infty}(1 + 1/n)^n = 2.71828\ldots.$$

You might recall, from the study of infinite series in calculus, that

$$\sum_{x=0}^{\infty} b^x/x! = e^b,$$

for any real number b. Thus,

$$\sum_{x=0}^{\infty} P(X = x) = e^{-\theta}\sum_{x=0}^{\infty}\theta^x/x! = e^{-\theta}e^{\theta} = 1.$$

Table 13.1: A comparison of three probability distributions.

	Distribution of X is:		
	Poisson(1)	Bin$(1000, 0.001)$	Bin$(500, 0.002)$
Mean :	1	1	1
Variance :	1	0.999	0.998
x	$P(X = x)$	$P(X = x)$	$P(X = x)$
0	0.3679	0.3677	0.3675
1	0.3679	0.3681	0.3682
2	0.1839	0.1840	0.1841
3	0.0613	0.0613	0.0613
4	0.0153	0.0153	0.0153
5	0.0031	0.0030	0.0030
6	0.0005	0.0005	0.0005
≥ 7	0.0001	0.0001	0.0001
Total	1.0000	1.0000	1.0000

Thus, we see that Formula 13.1 is a mathematically valid way to assign probabilities to the nonnegative integers; i.e., all probabilities are nonnegative—indeed, they are positive—and they sum to one.

The mean of the Poisson is its parameter θ; i.e., $\mu = \theta$. This can be proven using calculus and a similar argument shows that the variance of a Poisson is also equal to θ; i.e., $\sigma^2 = \theta$ and $\sigma = \sqrt{\theta}$.

When I write $X \sim$ Poisson(θ) I mean that X is a random variable with its probability distribution given by the Poisson distribution with parameter value θ.

I ask you for patience. I am going to delay my explanation of why the Poisson distribution is important in science.

As we will see, the Poisson distribution is closely tied to the binomial. For example, let's spend a few minutes looking at the three probability distributions presented in Table 13.1.

There is a wealth of useful information in this table. In particular,

1. If you were distressed that a Poisson random variable has an infinite number of possible values—namely, every nonnegative integer—agonize no longer! We see from the table that for $\theta = 1$, 99.99% of the Poisson probability is assigned to the event $(X \leq 6)$.

2. If you read down the three columns of probabilities, you will see that the entries are nearly identical. Certainly, any one column of probabilities provides good approximations to the entries in any other column. Thus, in some situations, a Poisson distribution can be used as an approximation to a binomial distribution.

3. What do we need for the Poisson to be a good approximation to a binomial? First, we need to have the means of the distributions match; i.e., we need to use the Poisson with $\theta = np$, as I did in Table 13.1. The variance of a binomial npq is necessarily smaller than the mean

np because $q < 1$. Thus, the variance of a binomial *cannot be made to match* the variance of the Poisson:

$$\text{Variance of binomial } = npq < np = \theta = \text{ variance of Poisson.}$$

If, however, p is very close to 0, then q is very close to one and the variances *almost match* as illustrated in Table 13.1.

I will summarize the above observations in the following result.

Result 13.1 (The Poisson approximation to the binomial.) *The Bin(n, p) distribution can be well-approximated by the Poisson(θ) distribution if the following conditions are met:*

1. *The distributions have the same mean; i.e., $\theta = np$;*

2. *The value of n is large and p is close to zero. In particular, the variance of the binomial npq should be very close to the variance of the Poisson, $\theta = np$.*

As a practical matter, we mostly use this result if $n >$ 1,000 because we can easily obtain exact binomial probabilities from a website for $n \leq$ 1,000. Also, if $np \geq 25$, our general guideline from Chapter 11 states that we may use a Normal curve to obtain a good approximation to the binomial. Thus, again as a practical matter, we mostly use this result if $\theta = np \leq 25$, allowing us some indecision as to which approximation to use at $np = 25$, Normal or Poisson.

Poisson probabilities can be computed by hand with a scientific calculator. Alternatively, the following website can be used:

`http://stattrek.com/Tables/Poisson.aspx`.

I will give an example to illustrate the use of this site.

Let $X \sim$ Poisson(θ). The website calculates five probabilities for you:

$$P(X = x); P(X < x); P(X \leq x); P(X > x); \text{ and } P(X \geq x).$$

You must give as input your value of θ and a value of x. Suppose that I have $X \sim$ Poisson(10) and I am interested in $P(X = 8)$. I go to the site and enter 8 in the box *Poisson random variable*, and I enter 10 in the box *Average rate of success.* I click on the *Calculate* box and the site gives me the following answers:

$$P(X = 8) = 0.1126; P(X < 8) = 0.2202; P(X \leq 8) = 0.3328; P(X > 8) = 0.6672;$$

$$\text{and } P(X \geq 8) = 0.7798.$$

As with our binomial calculator, there is a great deal of redundancy in these five answers.

13.1.1 The Normal Approximation to the Poisson

Please look at the Poisson(1) probabilities in Table 13.1. We see that $P(X = 0) = P(X = 1)$ and as x increases beyond 1, $P(X = x)$ decreases. Thus, without actually drawing the probability histogram of the Poisson(1) we know that it is strongly skewed to the right; indeed, it has no left tail! For $\theta < 1$ the probability histogram is even more skewed than it is for our tabled $\theta = 1$. As the value of θ increases the amount of skewness in the probability histogram decreases, but the Poisson is never perfectly symmetric.

In this course, I advocate the general guideline that if $\theta \geq 25$, then the Poisson's probability histogram is approximately symmetric and bell-shaped. (One can quibble about my choice of 25 and I wouldn't argue about it much.) This last statement suggests that we might use a Normal curve to compute approximate probabilities for the Poisson, provided θ is large.

For example, suppose that $X \sim$ Poisson(25) and I want to calculate $P(X \geq 30)$. We will use a modification of the method we learned for the binomial.

First, we note that $\mu = 25$ and $\sigma = \sqrt{25} = 5$. Thus, our approximating curve will be the Normal curve with these values for its mean and standard deviation. Using the continuity correction, we replace $P(X \geq 30)$ with $P(X \geq 29.5)$. Next, going to the Normal curve website, we find that the area above (to the right of) 29.5 is 0.1841. From the Poisson website, I find that the exact probability is 0.1821.

13.2 Inference for a Poisson distribution

If θ is unknown then its point estimator is X, with point estimate equal to x, the observed value of X. We have two options for obtaining a confidence interval estimate of θ: an approximate interval based on using a Normal curve approximation and an exact (conservative) confidence interval using the Poisson equivalent of the work of Clopper and Pearson.

It is possible to perform a test of hypotheses on the value of θ. The test is not widely useful in science; thus, I won't present it.

13.2.1 Approximate Confidence Interval for θ

I will very briefly sketch the rational behind the Normal curve approximation. The main ideas are pretty much exactly the ideas we had for the binomial in Chapter 12. We standardize our point estimator X to obtain

$$Z = \frac{X - \theta}{\sqrt{\theta}}.$$

Next, we replace the unknown parameter in the denominator by its point estimator, yielding

$$Z' = \frac{X - \theta}{\sqrt{X}}.$$

Slutsky's theorem applies; for θ sufficiently large, probabilities for Z' can be well-approximated by using the N(0,1) curve. With the same algebra we used in Chapter 12, we obtain the following

approximate confidence interval estimate of θ:

$$x \pm z^* \sqrt{x}, \tag{13.2}$$

where the value of z^* is determined by the choice of confidence level *in exactly the same way as it was for the binomial*. Thus, you can find the z^* you need in Table 12.1 on page 296.

I have investigated the performance of Formula 13.2 and I have concluded that the approximation is good for any $\theta \geq 40$; i.e., for any $\theta \geq 40$ the actual probability that this formula will give a correct confidence interval is close to the target reflected by the choice of z^*. As always, one can quibble with my choice of 40 as the magic threshold. It is larger than my choice, 25, for using a Normal curve to approximate Poisson probabilities in part because the confidence interval also relies on Slutsky's approximation.

In practice, of course, we estimate θ because we don't know its value. Thus, if you are concerned with having a guideline based on the value of θ, an alternative guideline is to use the approximate confidence interval if $x \geq 50$.

13.2.2 The 'Exact' (Conservative) Confidence Interval for θ

Suppose that we plan to observe a random variable X and we are willing to assume that $X \sim$ Poisson(θ). We want to use the observed value of X to obtain a confidence interval for θ, but the condition for using the approximate method of the previous subsection is not met. For example, suppose that we observe $X = 10$; what should we do?

In Chapter 12, when you learned how to use the website:

`http://statpages.org/confint.html`

you probably noticed that the website also can be used for Poisson distribution. Click on this website now and scroll down to the section **Poisson Confidence Intervals**. You will see that there is one box for data entry, called **Observed Events**; this is where you place the observed value of X. Note that the default value is 10, which, coincidentally, is the value I asked you to use! Click on the *Compute* box and the site gives you the exact—which, as in Chapter 12, really means conservative—two-sided 95% confidence interval for θ:

$$[4.7954, 18.3904].$$

If, instead, you want the two-sided 98% confidence interval for θ, then you proceed exactly as you did in Chapter 12. Scroll down to **Setting Confidence Levels**, type 98 in **Confidence Level** and click on *Compute*. Scroll back up to **Poisson Confidence Intervals** and make sure that 10 is still in the **Observed Events** box. Click on the *Compute* box and the site gives the answer:

$$[4.1302, 20.1447].$$

Suppose that I want the one-sided 90% upper confidence bound for θ, still with $x = 10$. Scroll down to **Setting Confidence Levels**, enter 10 in the **Upper Tail**, enter 0 in the **Lower Tail** and

click on *Compute*. Scroll back up to **Poisson Confidence Intervals** and make sure that 10 is still in the **Observed Events** box. Click on the *Compute* box and the site gives the answer:

$$[0.4749, 15.4066].$$

This answer is a bit strange; the lower bound in the interval should be 0, but it's not. I played around with this website a bit and here is what I learned. If $x \leq 2$ then the site gives 0 as the (correct) lower bound for the one-sided interval. If, however, $x \geq 3$, it gives a positive lower bound, which seems to be incorrect. This is not incorrect for two reasons:

1. We are free to replace the non-zero lower bound with 0 if we want; by making the interval wider, the probability of a correct interval becomes a bit larger.

2. Without examining either the programmer's code or performing a huge analysis—which I have neither the time nor interest to do—I can't be sure, but I believe that having a non-zero lower bound is part of the conservative nature of the site's intervals. Here is what I mean. If θ actually equaled the lower bound I have above for $x = 10$, which is 0.4749, then the probability of 10 or more successes is 10^{-10} (you can find this on our website for computing Poisson probabilities). Thus, if $x = 10$, values of θ smaller than 0.4749 are pretty much impossible anyways.

The next example shows why this material provides insight into some of our work in Chapter 12.

Example 13.1 (Don K. and high hopes) *Don K. was a teammate on my high school basketball team. Don wasn't very tall, but he was very quick and had a very strong throwing arm. He started his senior year as first or second player off the bench, but as the year progressed his playing time diminished. A highlight of his year was when he sank a half-court shot at the end of a quarter in a blow-out 93-40 victory. After his amazing shot, Don would spend most of his practice free time attempting very long shots. I don't remember him making many such shots, but everyone on the team noted how our coach, Mr. Pasternak— whom we affectionately dubbed* Boris *either because of his resemblance to the actor Boris Karloff or because* Doctor Zhivago *was* **the** *movie of 1965— was doing a slow boil from frustration. Finally, one day at practice, Coach could contain himself no longer and berated Don at length for not practicing a more useful basketball skill. Eight minutes later during a scrimmage as the time clock was running down to zero, Don grabbed a defensive rebound, pivoted and threw the ball 70 (?) feet, resulting in a perfect basket—swish through the net. Don ran around the court yelling, "See, Boris, I have been practicing a useful shot," while the rest of us collapsed in laughter.*

Perhaps because of my friend Don's experience, I have always been interested in situations in which successes are rare. Thus, let's look at some examples. I used the site

`http://statpages.org/confint.html`

to obtain the exact (conservative) 95% upper bound for p in each of the situations below.

- A total of $x = 0$ successes are obtained in $n = 10$ Bernoulli trials; the exact (conservative) 95% upper bound for p is: $p \leq 0.2589$.

- A total of $x = 0$ successes are obtained in $n = 100$ Bernoulli trials; the exact (conservative) 95% upper bound for p is: $p \leq 0.0295$.
- A total of $x = 0$ successes are obtained in $n = 1{,}000$ Bernoulli trials; the exact (conservative) 95% upper bound for p is: $p \leq 0.0030$.

As I have mentioned a number of times in these notes, the weakness of exact answers is that they are a black box; we can't see a pattern in the answers. There is a pattern in the above answers, as I will now demonstrate. (Indeed, you might see the pattern above, but you won't know *why* until you read on.)

Let's suppose now that our random variable X has a Poisson distribution and we observe $x = 0$. Using the same website, I can obtain an upper 95% confidence bound for θ; it is $\theta \leq 2.9957$, which, when I am feeling especially daring, I round to $\theta \leq 3.000$. Now we are going to use the fact that, under certain conditions, we can use the Poisson to approximate the binomial. Ignoring the conditions for a moment, recall that the key part of the approximation is to set θ for the Poisson equal to np from the binomial. Thus—and this is the key point—an exact confidence interval for θ is an approximate confidence interval for np. Thus, the upper bound $\theta \leq 3.000$ becomes $np \leq 3.000$ which becomes the following result.

Result 13.2 (Approximate 95% Confidence Upper Bound for p When $x = 0$.) *If $n \geq 100$,*

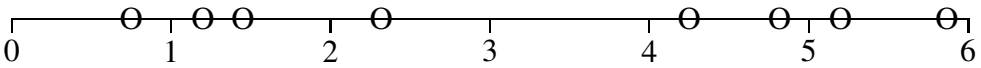
$$p \leq 3/n, \tag{13.3}$$

is a good approximation to the exact 95% confidence upper bound for p when $x = 0$. This result is sometimes referred to as the rule of 3.

13.3 The Poisson Process

The binomial distribution is appropriate for counting successes in n i.i.d. trials. For p small and n large, the binomial can be well approximated by the Poisson. Thus, it is not too surprising to learn that the Poisson distribution is also a model for counting successes.

Consider a process evolving in time in which at *random times* successes occur. What does this possibly mean? Perhaps the following picture will help.



In this picture, observation begins at time $t = 0$ and the passage of time is denoted by moving to the right on the number line. At various times successes will occur, with each success denoted by the letter 'O' placed on the number line. Here are some examples of such processes.

1. A 'target' is placed near radioactive material and whenever a radioactive particle hits the target we have a success.

2. A road intersection is observed. A success is the occurrence of an accident.

3. A hockey (or soccer) game is watched. A success occurs whenever a goal is scored.

4. On a remote stretch of highway, a success occurs when a vehicle passes.

The idea is that the times of occurrences of successes cannot be predicted with certainty. We would like, however, to be able to calculate probabilities. To do this, we need a mathematical model, much like our mathematical model for Bernoulli trials.

Our model is called the **Poisson Process**. A careful mathematical presentation and derivation is beyond the goals of this course. Here are the basic ideas:

1. **Independence:** The number of successes in disjoint intervals are independent of each other.

   For example, in a Poisson Process, the number of successes in the interval $[0, 3]$ is independent of the number of successes in the interval $[5, 6]$.

2. **Identically distributed:** The probability distribution of the number of successes counted in any time interval depends only on the length of the interval.

   For example, the probability of getting exactly five successes is the same for interval $[0, 2.5]$ as it is for interval $[3.5, 6.0]$.

3. Successes cannot be simultaneous. (This assumption is needed for technical reasons that we won't discuss.)

With these assumptions, it turns out that the probability distribution of the number of successes in *any* interval of time is the Poisson distribution with parameter θ, where $\theta = \lambda \times w$, where $w > 0$ is the length of the interval and $\lambda > 0$ is a feature of the process, often called its **rate**.

I have presented the Poisson Process as occurring in one dimension—time. It also can be applied if the one dimension is, say, distance. For example, a researcher could be walking along a path and occasionally finds successes. Also, the Poisson Process can be extended to two or three dimensions. For example, in two dimensions a researcher could be searching a field for a certain plant or animal that is deemed a success. In three dimensions a researcher could be searching a volume of air, water or dirt looking for something of interest.

The modification needed for two or three dimensions is quite simple: the Poisson Process still has a rate, again called λ, and now the number of successes in an area or volume has a Poisson distribution with θ equal to the rate multiplied by the area or volume, whichever is appropriate. Also, of course, to be a Poisson Process in two or three dimensions requires the assumptions of independence and identically distributed to be met.

13.4 Independent Poisson Random Variables

Earlier we learned that if $X_1, X_2, \ldots, X_n$ are i.i.d. dichotomous outcomes (success or failure), then we can calculate probabilities for the sum of these guys X:

$$X = X_1 + X_2 + \ldots X_n.$$

Probabilities for X are given by the binomial distribution. There is a similar result for the Poisson, but the conditions are actually weaker. The interested reader can think about how the following result is implied by the Poisson Process.

Result 13.3 (The sum of independent Poisson random variables.) *Suppose that for $i = 1, 2, 3, \ldots, n$, the random variable $X_i \sim$ Poisson(θ_i) and that the sequence of X_i's are independent. Define $\theta_+ = \sum_{i=1}^{n} \theta_i$. Then $X \sim$ Poisson(θ_+).*

Because of this result we will often (as I have done above), but not always, pretend that we have *one* Poisson random variable, even if, in reality, we have a sum of independent Poisson random variables. I will illustrate what I mean with an estimation example.

Suppose that Cathy observes 10 i.i.d. Poisson random variables, each with parameter θ. She summarizes the ten values she obtains by computing their total, X, remembering that $X \sim$ Poisson(10θ). Cathy can then calculate a confidence interval for 10θ and convert it to a confidence interval for θ.

For example, suppose that Cathy observes a total of 92 when she sums her 10 values. Because 92 is sufficiently large (it exceeds 50), I will use the formula for the approximate two-sided 95% confidence interval for 10θ. It is:

$$92 \pm 1.96\sqrt{92} = 92 \pm 18.800 = [73.200, 110.800].$$

The interpretation of this interval is, of course:

$$73.200 \leq 10\theta \leq 110.800.$$

If we divide through by 10, we get

$$7.3200 \leq \theta \leq 11.0800.$$

Thus, the two-sided approximate 95% confidence interval for θ is $[7.320, 11.080]$. By the way, the exact confidence interval for 10θ is $[74.165, 112.83]$. This is typically what happens; the exact confidence interval for a Poisson is shifted to the right of the approximate confidence interval because the Poisson distribution is skewed to the right.

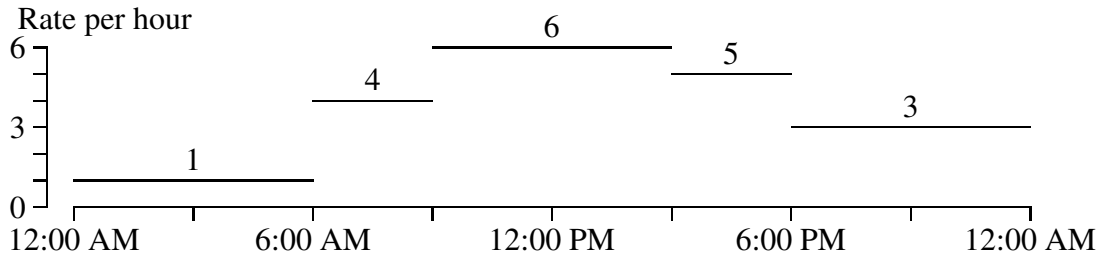
13.4.1 A Comment on the Assumption of a Poisson Process

Recall my four examples of possible Poisson Processes given on page 327. My first example, radioactive decay, was, by far, the most popular example in textbooks on probability theory, circa 1970, when I was an undergraduate student. Literally, radioactive decay involves a source of radioactive material comprised of a huge number of atoms, each of which has a very small probability of decaying in a short time period. Because atoms don't talk to each other, "Hey, Adam, I am about to decay, will you join me?" it seems extremely reasonable to believe we have a huge number of Bernoulli trials with a very small value of p. Hence, assuming a Poisson Process is simply restating the idea that the Poisson distribution approximates the binomial. All models have an implicit

expiration date; for example, if I am still shooting free throws at age 80, I definitely won't have the same p I had at age 17. For radioactive decay, if the length of observation approaches the half-life of the element then the rate will definitely decrease because—by definition—half the atoms have *decayed* at the half life. For example, uranium-232 has a half-life of 69 years and carbon-14, which is used to date fossils, has a half-life of 5,730 years.

I hope that you will agree that radioactive decay is a pretty solid example of a Poisson Process. My second and fourth examples—both involving traffic—appear, however, to be on shaky ground. Let's examine the fourth example, in which a success is the passage of a car on a remote stretch of highway. When I think of a remote highway, it is hard for me to imagine that the rate of traffic at, say, 3:00 AM is the same as it is at 3:00 PM. Thus, you might think that the assumption of a Poisson Process is reasonable only for a very limited period of time, say, 9:00 AM to 4:00 PM. You would be correct, except for what I am now going to tell you, which is the point of this subsection.

I want to make this argument very concrete. To that end, suppose that I am Nature and I know that the rate is as given in the following picture.



Let's make sure that this picture is clear. From 12:00 AM (midnight) to 6:00 AM a car passing the spot follows a Poisson Process with an average of one car per hour. From 6:00 AM to 9:00 AM the rate of the Poisson Process quadruples to four cars per hour; and so on.

If we watch the road continuously, then we do not have a Poisson Process over the 24 hours of a day because the rate is not constant. If I look at the process for *certain limited periods of time*, then I will have a Poisson Process; for example, if I observe the process over the six hour time period of 9:00 AM to 3:00 PM, I am observing a Poisson Process with rate equal to six cars per hour.

Now let's imagine, however, that we **do not** observe the process continuously at all. Instead, every day at the same time, say midnight, we are told how many cars passed the spot in the day just completed. Call this observed count x with corresponding random variable X. I will now demonstrate that X has a Poisson distribution.

We can write X as the sum of five random variables:

$$X = X_1 + X_2 + X_3 + X_4 + X_5,$$

where

- X_1 is the number of cars that pass the spot between midnight and 6:00 AM.
- X_2 is the number of cars that pass the spot between 6:00 AM and 9:00 AM.

Table 13.2: The number of homicides, by year, in Baltimore, Maryland.

Year:	2003	2004	2005	2006	2007
Number of homicide deaths:	270	276	269	276	282

- And so on, for X_3, X_4 X_5, throughout the day.

From the above picture, being Nature I know that:

- $X_1 \sim$ Poisson($6 \times 1 = 6$); $X_2 \sim$ Poisson($4 \times 3 = 12$); $X_3 \sim$ Poisson($6 \times 6 = 36$); $X_4 \sim$ Poisson($5 \times 3 = 15$); and $X_5 \sim$ Poisson($3 \times 6 = 18$).
- Also, the random variables X_1, $X_2 \ldots X_5$ are statistically independent.
- From Result 13.3, we know that X has a Poisson distribution with parameter

$$\theta_+ = 6 + 12 + 36 + 15 + 18 = 87.$$

I might even abuse language a bit and say that the number of cars passing the spot is a Poisson Process with a rate of 87 cars per day. I shouldn't say this of course, but sometimes we get a bit lazy in probability and statistics!

Of course, I am not Nature, so I would never know the exact rate. The following example with real data is illustrative of the above method.

Example 13.2 (Homicides in Baltimore.) *I recently discovered data on homicides, by year, in Baltimore, Maryland. The data are presented in Table 13.2.*

I am going to assume that the number of homicides per year is a Poisson Process with unknown rate of λ homicides per year. I will revisit this example in Chapter 14. With my assumption, I have observed the process for five units of time—five years—and counted a total of

$$270 + 276 + 269 + 276 + 282 = \text{ 1,373 successes.}$$

(Remember that whatever we are counting, no matter how tragic it might be, is called a success. Hence, a homicide death is a success.) We view 1,373 as the observed value of a random variable X with Poisson(θ) distribution. Because my observed value of X is much larger than 50, I feel comfortable using the approximate confidence interval for θ, given in Formula 13.2. For 95% confidence, we get

$$1373 \pm 1.96\sqrt{1373} = 1373 \pm 72.6 = [1300.4, 1445.6].$$

Because the process was observed for five time units, we have $\theta = 5\lambda$. Thus, the above confidence interval for θ becomes

$$1300.4 \leq 5\lambda \leq 1445.6;$$

after dividing through by 5, we get

$$260.08 \leq \lambda \leq 289.12.$$

Thus, $[260.08, 289.12]$ is my approximate 95% confidence interval for the rate of homicides per year in Baltimore during the years 2003–2007.

13.5 Summary

The Poisson is a probability distribution—see Equation 13.1—concentrated on the nonnegative integers. The Poisson distribution has a single parameter, θ, which can be any positive number. The mean and variance of a Poisson distribution both equal θ and the standard deviation equals $\sqrt{\theta}$.

Poisson probabilities can be calculated with the help of the website:

`http://stattrek.com/Tables/Poisson.aspx`.

If $\theta \geq 25$, then the Normal curve with $\mu = \theta$ and $\sigma = \sqrt{\theta}$ will give good approximations to the Poisson(θ) distribution.

The first use for the Poisson distribution is as an approximation to the binomial distribution. In particular, suppose we have a Bin(n, p) distribution, with n large, p small and npq approximately equal to np; i.e., q is very close to one. If we set $\theta = np$, then the Poisson distribution is a good approximation to the Binomial distribution.

If $X \sim$ Poisson(θ), then X is the point estimator of θ. The standardized version of the point estimator X is

$$Z = \frac{X - \theta}{\sqrt{\theta}}.$$

As implied above, if $\theta \geq 25$, then the N(0,1) curve provides good approximate probabilities for Z. Combining the above with Slutsky's theorem, we obtain the following approximate confidence interval for θ:

$$x \pm z^* \sqrt{x},$$

where the value of z^* depends on the choice of confidence level and is given in Table 12.1 on page 296. My advice is that this interval performs as advertised provided $x \geq 50$. For smaller values of x, see the next paragraph.

There is an exact—actually conservative—confidence interval for θ, available on the website:

`http://statpages.org/confint.html`

The Poisson distribution also arises from a mathematical model for successes occurring randomly in time. In particular, the first two of the three assumptions of a Poisson Process are similar to the assumptions of Bernoulli trials. If we have a Poisson Process then the number of successes in any time interval of length w has a Poisson distribution with parameter $\theta = w\lambda$, where $\lambda > 0$ is a parameter of the process, called its rate. (If $w = 1$, then $\theta = \lambda$. Thus, the mean number of successes in one unit of time is λ; hence, the name rate.)

When I talk about a Poisson Process in general, I will speak of it evolving in time. It could, alternatively, evolve in distance. Moreover, a Poisson Process can be used for counting successes in two or three dimensions.

The Poisson distribution has the following very useful property. If the random variables X_1, X_2, $\ldots$, X_n, are independent with $X_i \sim$ Poisson(θ_i)—i.e., the X_i's need not be identically distributed—then the new random variable

$$X = X_1 + X_2 + \ldots X_n = \sum X_i,$$

has a Poisson distribution with parameter

$$\theta_+ = \theta_1 + \theta_2 + \ldots \theta_n = \sum \theta_i.$$

In words, the sum of independent Poisson random variables has a Poisson distribution; and the parameter for the sum is the sum of the parameters. This property of Poisson distributions can be very useful; I illustrate its use with data on the annual number of homicide deaths in Baltimore, Maryland.

13.6 Practice Problems

1. Suppose that $X \sim$ Poisson(20). Use the website

   `http://stattrek.com/Tables/Poisson.aspx`

   to calculate the following probabilities.

   (a) $P(X = 20)$.
   (b) $P(X \leq 20)$.
   (c) $P(X > 20)$.
   (d) $P(16 \leq X \leq 24)$.

2. Suppose that $X \sim$ Bin(2000,0.003). I want to know $P(X \leq 4)$. Help me by calculating an approximate probability for this event.

3. Wayne Gretzky is perhaps the greatest hockey player ever. We have the following data from his NHL (National Hockey League) career.

   - During the 1981–82 season he played 80 games and scored 92 goals.
   - During the 1982–83 season he played 80 games and scored 71 goals.
   - During the 1983–84 season he played 74 games and scored 87 goals.

   Assume that Gretzky's goal scoring followed a Poisson Process with a rate of λ goals per game. Use the three seasons of data given above to obtain an approximate 98% confidence interval for λ.

4. Let $X \sim$ Poisson(θ). Given $X = 1$, find the exact 95% upper confidence bound for θ. Apply your finding to create *the rule of 4.75 when* $X = 1$.

13.7 Solutions to Practice Problems

1. For parts (a)–(c), go to the website and enter 20 for both x and the **Average rate of success**. You will obtain:

   (a) $P(X = 20) = 0.0888$.

   (b) $P(X \leq 20) = 0.5591$.

   (c) $P(X > 20) = 0.4409$.

   (d) There are several ways to get the answer. I suggest:

   $$P(16 \leq X \leq 24) = P(X \leq 24) - P(X \leq 15).$$

   I enter the website twice and obtain:

   $$P(16 \leq X \leq 24) = 0.8432 - 0.1565 = 0.6867.$$

2. Our binomial calculator website does not work for $n >$ 1,000; hence, I want an approximate answer. For the binomial, the mean is $np = 2000(0.003) = 6$. This is much smaller than 25, so I will not use the Normal curve approximation. In addition, the binomial variance is $npq = 6(0.997) = 5.982$ which is only a bit smaller than the mean. Thus, I will use the Poisson approximation. I go to the website

   `http://stattrek.com/Tables/Poisson.aspx`

   and enter 4 for x and $\theta = np = 6$ for **Average rate of success**. The website gives me 0.2851 as its approximation of $P(X \leq 4)$.

   By the way, Minitab is able to calculate the exact probability; it is 0.2847. Thus, the Poisson approximation is very good.

3. Combining the data, we find that Gretzky scored 250 goals in 234 games. We view $x = 250$ as the observed value of a random variable X which has a Poisson distribution with parameter θ. Also, $\theta = 234\lambda$. For 98% confidence, we see from Table 12.1 that $z^* = 2.326$. Thus, the approximate 98% confidence interval for θ is

   $$250 \pm 2.326\sqrt{250} = 250 \pm 36.78 = [213.22, 286.78].$$

   Literally, we are asserting that

   $$213.22 \leq \theta \leq 286.78 \text{ or } 213.22 \leq 234\lambda \leq 286.78.$$

   Dividing through by 234, we get

   $$213.22/234 \leq \lambda \leq 286.78/234 \text{ or } 0.911 \leq \lambda \leq 1.226.$$

4. Go to the website

   `http://statpages.org/confint.html`.

   Scroll down to **Setting Confidence Levels**. Enter 5 in the **Upper** box, 0 in the **Lower** box and click on **Compute**. The site now knows that we want the 95% upper confidence bound.

   Scroll up to **Poisson Confidence Intervals**, enter 1 in the **Observed Events** box and click on **Compute**. The site gives us $[0, 4.7439]$ as the upper 95% confidence bound for θ.

   If $X \sim \text{Bin}(n, p)$ with n large and the observed value of X is 1, then 4.7439, rounded rather clumsily to 4.75, is the approximate 95% upper confidence bound for np. Thus, for n large and $X = 1$,

   $$4.75/n \text{ is the approximate } 95\% \text{ upper confidence bound for } p.$$

   As a partial check, I scrolled up to **Binomial Confidence Intervals**, entered 1 for x, entered 100 for n, and clicked on **Compute**. The site gave me 0.0466 as the exact 95% upper confidence bound for p, which is reasonably approximated by $4.75/n = 4.75/100 = 0.0475$.

Table 13.3: Traffic accident data in Madison, Wisconsin.

Year:	2005	2006	2007	2008	2009	Total
Average weekday arterial volume	26,271	25,754	25,760	24,416	24,222	126,423
Total crashes	4,577	4,605	4,779	4,578	4,753	23,292
Bike crashes	97	95	118	95	115	520
Pedestrian crashes	84	87	80	76	77	404
Fatal crashes	9	12	13	6	14	54

13.8 Homework Problems

1. Suppose that $X \sim$ Poisson(10). Use the website

   `http://stattrek.com/Tables/Poisson.aspx`

   to calculate the following probabilities.

   (a) $P(X = 8)$.

   (b) $P(X \leq 6)$.

   (c) $P(X \leq 15)$.

   (d) $P(7 \leq X \leq 15)$.

2. Suppose that $X \sim$ Bin(5000,0.0001). According to Minitab, $P(X \leq 2) = 0.9856$. Find the Poisson approximation to this probability. Compare your approximate answer with the exact answer and comment.

3. Let $X \sim$ Poisson(θ). Given $X = 2$, find the exact 95% upper confidence bound for θ. Apply your finding to create *the rule of 6.30 when* $X = 2$.

4. The data in Table 13.3 appeared in the Wisconsin State Journal on July 13, 2010, for accidents involving autos in Madison, Wisconsin.

   In parts (a)–(c), assume that the number of crashes of interest follows a Poisson Process with unknown rate λ per year. Use the data in the *Total* column to obtain the approximate 95% confidence interval estimate of λ.

   (a) Bike crashes.

   (b) Pedestrian crashes.

   (c) Fatal crashes. Also obtain the exact confidence interval.

Chapter 14

Rules for Means and Variances; Prediction

14.1 Rules for Means and Variances

The result in this section is very technical and algebraic. And dry. But it is useful for understanding many of prediction results we obtain in this course, beginning later in this chapter.

We have independent random variables W_1 and W_2. Note that, typically, they are not identically distributed. The result below is an extremely special case of a much more general result. It will suffice, however, for our needs; thus, I see no reason to subject you to the pain of viewing the general result.

We need some notation:

- Let μ_1 [μ_2] denote the mean of W_1 [W_2].
- Let Var(W_1) [Var(W_2)] denote the variance of W_1 [W_2].
- Let b denote any number. Define

$$W = W_1 - bW_2.$$

Result 14.1 (The mean and variance of W.) *For the notation given above,*

- *The mean of W is*

$$\mu_W = \mu_1 - b\mu_2 \tag{14.1}$$

- *The variance of W is*

$$\textit{Var}(W) = \textit{Var}(W_1) + b^2\, \textit{Var}(W_2). \tag{14.2}$$

In our two applications of this result in this chapter, the number b will be taken to equal μ_1/μ_2; thus,

$$\mu_W = \mu_1 - (\mu_1/\mu_2)\mu_2 = \mu_1 - \mu_1 = 0,$$

for our applications.

14.2 Predicting for Bernoulli Trials

Predictions are tough, especially about the future—Yogi Berra.

We plan to observe m Bernoulli trials and want to predict the total number of successes that will be obtained. Let Y denote the random variable and y the observed value of the total number of successes in the future m trials. Similar to estimation, we will learn about point and interval predictions of the value of Y.

14.2.1 When p is Known

Because prediction is a new idea in this course, I want to present a gentle introduction to it. Suppose that you have a favorite pair of dice, one colored blue and the other white. Let's focus on the blue die. And let's say your favorite number is 6—perhaps you play *Risk* a great deal; in *Risk*, 6 is the best outcome by far when one casts a die. Ghengis Khan playing *Risk* would roll a lot of 6's.

You plan to cast the die 600 times and you want to predict the number of 6's that you will obtain. You believe that the die is balanced; i.e., that the six possible outcomes are equally likely to occur.

OK. Quick. Don't think of any of the wonderful things you have learned in this course. Give me your point (single number) prediction of how many 6's you will obtain. I conjecture that your answer is 100. (I asked this question several times over the years to a live lecture and always—save once—received the answer 100 from the student who volunteered to answer. One year a guy said 72 and got a big laugh. I failed him because it is my job to make the jokes, such as they are. No, I didn't really fail him, but I was more than a bit annoyed that he got a larger laugh than I did with my much cleverer anecdotes.)

My academic grandfather (my advisor's advisor, who happened to be male) is Herb Robbins, a very brilliant and witty man. Herb was once asked what mathematical statisticians do, and he replied, "They find out what non-statisticians do and prove it's optimal."

Thus, I am going to argue that your answer of 100 is the best answer to the die question I posed above. In order to show that something is best mathematically, we find a way to measure *good* and whichever candidate answer has the largest amount of good is best. This is the approach for the glass-half-full people. More often, one finds a way to measure *bad* and whichever candidate answer has the smallest amount of bad is best.

We want to predict, in advance, the value that Y will yield. We denote the point prediction by the single number $\hat{y}$. We adopt the criterion that we want the probability of being correct to be as large as possible. (Thus, we define being correct as good and seek to maximize the probability that we will get a good result.)

Result 14.2 (The best point prediction of Y.) *Calculate the mean of Y, which is mp.*

- **If mp is an integer**, *then it is uniquely the most probable value of Y and our point prediction is $\hat{y} = mp$.*

- **If mp is not an integer**, *then the most probable value of Y is one of the integers immediately on either side of mp. Check them both; whichever is more probable is the point prediction. If they are equally probably, I choose the even integer.*

Below are some examples of this result.

- For my die example, $m = 600$ and $p = 1/6$, giving $mp = 600(1/6) = 100$. This is an integer; thus, 100 is the point prediction of Y. With the help of the website calculator (details not given), I find that $P(Y = 100) = 0.0437$. For comparison, $P(Y = 99) = 0.0436$ and $P(Y = 101) = 0.0432$. Thus, if 99 is your life-long favorite number, it is difficult for me to criticize using it as your point prediction. In the long-run, you will have one fewer correct point prediction for every 10,000 times you cast the blue die 600 times. That's a lot of die casting!

- Suppose that $m = 200$ and $p = 0.50$. Then, $mp = 200(0.5) = 100$ is an integer; thus, 100 is the point prediction of Y. With the help of the website calculator, I find that $P(Y = 100) = 0.0563$.

- Suppose that $m = 300$ and $p = 0.30$. Then, $mp = 300(0.3) = 90$ is an integer; thus, 90 is the point prediction of Y. With the help of the website calculator, I find that $P(Y = 90) = 0.0502$.

- Suppose that $m = 20$ and $p = 0.42$. Then, $mp = 20(0.42) = 8.4$ is not an integer. The most likely value of Y is either 8 or 9. With the help of the website calculator, I find that $P(Y = 8) = 0.1768$ and $P(Y = 9) = 0.1707$. Thus, $\hat{y} = 8$.

- Suppose that $m = 75$ and $p = 0.50$. Then, $mp = 75(0.50) = 37.5$ is not an integer. The most likely value of Y is either 37 or 38. With the help of the website calculator, I find that $P(Y = 37) = 0.0912$ and $P(Y = 38) = 0.0912$. Because these probabilities are identical, I choose the even integer; thus, $\hat{y} = 38$.

- Suppose that $m = 100$ and $p = 0.615$. Then, $mp = 100(0.615) = 61.5$ is not an integer. The most likely value of Y is either 61 or 62. With the help of the website calculator, I find that $P(Y = 61) = 0.0811$ and $P(Y = 62) = 0.0815$. Thus, $\hat{y} = 62$.

In each of the above examples we saw that the probability that a point prediction is correct is very small. As a result, scientists usually prefer a prediction interval. It is possible to create a one-sided prediction interval, but we will consider only two-sided prediction intervals.

We have two choices: using a Normal curve approximation or finding an exact interval. Even if you choose the exact interval, it is useful to begin with the Normal curve approximation.

Result 14.3 (Predicting when p is known.) *Let Y denote the total number of successes in m future Bernoulli trials. The Normal curve approximate prediction interval for Y when p is known is:*

$$mp \pm z^* \sqrt{mpq} \tag{14.3}$$

where the value of z^* is determined by the desired probability of getting a correct prediction interval. The value of z^* is given in Table 12.1 on page 296. It is the same number that is used for the two-sided confidence interval for p.

I won't indicate a proof of this result, other than to say it is the same derivation we used in Chapter 12 to find the approximate confidence interval for p, except that in the current situation we don't need Slutsky's theorem because the value of p is known.

For our die example with $m = 600$, I want to have a prediction interval for which the probability it will be correct equals approximately 98%. From Table 12.1, I find that $z^* = 2.326$. The prediction interval is:

$$600(1/6) \pm 2.326\sqrt{600(1/6)(5/6)} = 100 \pm 21.23 = [78.77, 121.23].$$

Let me make a couple of remarks concerning this answer. First, it makes no sense to predict that I will obtain a fractional number of 6's; thus, I round-off my endpoints to obtain the closed interval $[79, 121]$. Second, I remember that this answer is *not really an interval of numbers*; more accurately, I predict that Y will be one of the numbers in the set $79, 80, 81, \ldots, 121$. This more accurate statement is way too tedious for me; thus, I will abuse language and say that $[79, 121]$ is my approximate 98% prediction interval for Y.

You might be thinking: Hey, Bob, why did you use the Normal curve approximation; why not use exact binomial probabilities? Good question. If I am really serious about this prediction problem, I take my approximate answer, $[79, 121]$, as my *starting point*. I go to the binomial calculator website and find that for $m = 600$ and $p = 1/6$:

$$P(Y \leq 121) = 0.9894 \text{ and } P(Y \leq 78) = 0.0078.$$

Thus,

$$P(79 \leq Y \leq 121) = 0.9894 - 0.0078 = 0.9816,$$

which is very close to my target of 98% probability. Thus, I am really happy with the prediction interval $[79, 121]$.

14.2.2 When p is Unknown

We now consider the situation in which p is unknown. We will begin with point prediction. The first difficulty is that we cannot achieve our criterion's goal: we cannot find the most probable value of Y. The most probable value, as we saw above, is at or near mp, but we don't know what p is. Thus, we won't concern ourselves with point prediction.

Because p is unknown, we cannot use Formula 14.3 as a prediction interval. In order to proceed, we need to add another ingredient to our procedure. We assume that we have past data from the process that will generate the m future trials. We denote the past data as consisting of n trials which yielded x successes, giving $\hat{p} = x/n$ as our point estimate of the unknown p and, as always, $\hat{q} = 1 - \hat{p}$. We will now use the Result 14.1 to derive a prediction interval for Y.

It will be convenient to begin by defining a symbol r, the ratio of the future number of trials to the past number of trials:

$$r = m/n. \tag{14.4}$$

Note that if r is close to zero, then we are using a relatively large amount of past data to predict a relatively small number of future trials. Conversely, if r is large, then we are using a relatively small amount of past data to predict a relatively large number of future trials.

The algebra between this point and Result 14.4 is pretty intense. Unless you find that working through messy algebra improves your understanding, feel free to jump (skip, hop, dash) ahead to Result 14.4.

In Result 14.1, let $W_1 = Y$, $W_2 = X$ and $b = r$. Thus,

$$W = Y - rX.$$

The mean of W is:

$$\mu_W = \mu_Y - r\mu_X = mpq - (m/n)npq = mpq - mpq = 0$$

and the variance of W is:

$$\text{Var}(W) = \text{Var}(Y) + r^2 \text{Var}(X) = mpq + r(m/n)npq = mpq(1 + r).$$

Let me make a couple of quick comments about this formula for the variance of W. As you will see below, the larger the variance, the wider the prediction interval. We see that the variance is proportional to m, the number of trials being predicted. This makes sense; more trials means more uncertainty. The variance is also proportional to $(1 + r)$. To see why this makes sense, refer to my comments above immediately after Equation 14.4.

We can standardize W:

$$Z = \frac{W - \mu_W}{\sqrt{\text{Var}(W)}} = \frac{W - 0}{\sqrt{mpq}\sqrt{1 + r}} = \frac{Y - rX}{\sqrt{mpq}\sqrt{1 + r}}.$$

It can be shown that if m and n are both large and p is not too close to either 0 or 1, then probabilities for Z can be well approximated by the N(0,1) curve. Slutsky's work can also be applied here; the result is:

$$Z' = \frac{Y - rX}{\sqrt{rX[1 - (X/n)]}\sqrt{1 + r}}.$$

It can be shown that if m and n are both large and p is not too close to either 0 or 1, then probabilities for Z' can be well approximated by the N(0,1) curve. As a result, using the same algebra we had in Chapter 12 (the names have changed) we can expand Z' and get the following prediction interval for Y.

Result 14.4 (Predicting when p is unknown.) *The formula below is the approximate prediction interval for Y, the total number of successes in m future Bernoulli trials, when p is unknown. In this formula, r is given in Equation 14.4; x is the observed number of successes in the past data*

of n trials; $\hat{q} = (n - x)/n$ is the proportion of failures in the past data; and z^* is determined by the desired probability of the interval being correct. The relationship between z^* and the desired probability is given in Table 12.1 on page 296.

$$rx \pm z^* \sqrt{rx\hat{q}}\sqrt{1+r} = rx \pm z^* \sqrt{rx(1+r)\hat{q}}. \tag{14.5}$$

I will illustrate the use of this formula with real data from basketball.

On page 274, I introduced you to the data Katie Voigt collected and shared with me. I will use some of Katie's data to illustrate the current method.

I will put myself in time on day 3 of Katie's study *before* she collected the day's data. I want to use the combined data from Katie's first two days of shooting to predict the number of successes that she will achieve on her $m = 100$ trials on day 3.

Let me explicitly review the necessary assumptions. Katie's 300 shots on days 1–3 are Bernoulli trials with an unknown value of p. Note, in particular, that I assume that her future trials are governed by the same process that generated her past data. Now, let's get to Katie's data.

On day 1, Katie obtained 56 successes and on day 2 she obtained 59 successes. Combining these days, the past data consist of $n = 200$ trials with $x = 56 + 59 = 115$ successes. I will use these data to obtain the approximate 95% prediction interval for Y, the number of successes she will obtain on day 3. I will use Formula 14.5 to obtain my answer. Thus, I need to identify the values of the various symbols in the formula.

$$r = m/n = 100/200 = 0.5; x = 115; z^* = 1.96; \text{ and } \hat{q} = (200 - 115)/200 = 0.425.$$

Next, I substitute these values into Formula 14.5 and obtain:

$$0.5(115) \pm 1.96\sqrt{0.5(115)(0.425)}\sqrt{1+0.5} = 57.5 \pm 1.96\sqrt{24.4375}\sqrt{1.5} =$$

$$57.5 \pm 11.86 = [45.64, 69.36],$$

which I will round to $[46, 69]$.

This is a very wide interval. (Why do I say this? Well, in my opinion, a basketball player will believe that making 46 out of 100 is significantly different than making 69 out of 100.) A great thing about prediction (not shared by estimation or testing) is that we find out whether our answer is correct. It is no longer the case that **only Nature knows!** In particular, when Katie attempted the $m = 100$ future shots, she obtained $y = 66$ successes. The prediction interval is correct because it includes $y = 66$.

You should note that there is no exact solution to this problem. Even a simulation experiment is a challenge. I will discuss how a simulation experiment is performed for a limited situation motivated by Katie's data.

In order to do a simulation study we must specify three numbers: m, n and p. (Even though the researcher does not know p, Nature, the great simulator, must know it.) And, as we shall see, it is a two-stage simulation.

To be specific, I will simulate something similar to Katie's problem. I will take $m = 100$, $n = 200$ and $p = 0.60$. (Of course, I don't know Katie's p, but 0.60 seems to be not ridiculous.) A single rep of the simulation experiment is as follows.

1. Simulate the value of $X \sim$ Bin(200,0.60), our simulated past data for Katie.

2. Use our value of X from step 1 to estimate p and then compute the 95% prediction interval for Y, remembering to use the interval for p unknown.

3. Simulate the value of $Y \sim$ Bin(100,0.60) and see whether or not the interval from step 2 is correct.

I performed this simulation with 10,000 reps and obtained 9,493 (94.93%) correct prediction intervals! This is very close to the nominal (advertised) rate of 95% correct.

Before we leave this section, I will use Katie's data to obtain one more prediction interval.

I will put myself in time on day 9 of Katie's study *before* she collected the day's data. I want to use the combined data from Katie's first eight days of shooting ($n = 800$) to predict the number of successes that she will achieve on her $m = 200$ trials on days 9 and 10 combined ($m = 200$).

Let me explicitly review the necessary assumptions. Katie's 1,000 shots on days 1–10 are Bernoulli trials with an unknown value of p. Note, in particular, that I assume that her future trials are governed by the same process that generated her past data. Now, let's get to Katie's data.

On days 1–8, Katie obtained a total 505 successes. Thus, the past data consist of $n = 800$ trials with $x = 505$ successes. I will use these data to obtain the approximate 95% prediction interval for Y, the number of successes she will obtain on days 9 and 10 combined. I will use Formula 14.5 to obtain my answer. Thus, I need to identify the values of the various symbols in the formula.

$$r = m/n = 200/800 = 0.25; x = 505; z^* = 1.96; \text{ and } \hat{q} = (800 - 505)/800 = 0.369.$$

Next, I substitute these values into Formula 14.5 and obtain:

$$0.25(505) \pm 1.96\sqrt{0.25(505)(0.369)}\sqrt{1 + 0.25} = 126.25 \pm 1.96\sqrt{46.5862}\sqrt{1.25} =$$

$$126.25 \pm 14.96 = [111.29, 141.21],$$

which I will round to $[111, 141]$.

On days 9 and 10 combined, Katie achieved $y = 130$ successes. The prediction interval is correct because it includes $y = 130$.

14.3 Predicting for a Poisson Process

Compared to my above work on the binomial distribution, this section will be quite brief. I will consider only one of several possible problems, namely the following. I plan to observe a Poisson Process with unknown rate λ for t_2 units of time. I have past data on the same process that gave x successes in an observational period of t_1 units of time. After I derive the prediction interval, Formula 14.7, I will illustrate the method with the Baltimore homicide data introduced in Chapter 13 in Table 13.2 on page 331. In particular, I will put myself in time at the end of 2004, which means that my past data are for $t_1 = 2$ years and I will want to predict the total number of homicides for the year 2005, giving me a future observation of $t_2 = 1$ year.

Returning to the general problem, the observed number of successes in the past data, x, is the observed value of a random variable X which has distribution Poisson(λt_1). I want to predict the value of Y which has distribution Poisson(λt_2).

For the binomial problem with p unknown, recall that the ratio r, defined in Equation 14.4, played a key role in our prediction interval. We need a similar ratio for the current Poisson problem:

$$r' = t_2/t_1. \tag{14.6}$$

Note that, similar to the binomial problem, this is the ratio of the length of future observation of the process to the length of past observation of the process. I put a *prime symbol* on the notation for the current ratio to avoid confusion with the binomial problem's ratio.

The algebra between this point and Result 14.5 is pretty intense. Unless you find that working through messy algebra improves your understanding, feel free to jump ahead to Result 14.5.

In Result 14.1, let $W_1 = Y$, $W_2 = X$ and $b = r'$. Thus,

$$W = Y - r'X.$$

The mean of W is:

$$\mu_W = \mu_Y - r'\mu_X = \lambda t_2 - (t_2/t_1)\lambda t_1 = \lambda t_2 - \lambda t_2 = 0.$$

and the variance of W is:

$$\text{Var}(W) = \text{Var}(Y) + (r')^2\,\text{Var}(X) = \lambda t_2 + r'(t_2/t_1)\lambda t_1 = \lambda t_2(1 + r').$$

We can standardize W:

$$Z = \frac{W - \mu_W}{\sqrt{\text{Var}(W)}} = \frac{W - 0}{\sqrt{\lambda t_2(1 + r')}} = \frac{Y - r'X}{\sqrt{\lambda t_2(1 + r')}}.$$

It can be shown that if t_1 and t_2 are both large and λ is not too close to 0, then probabilities for Z can be well approximated by the N(0,1) curve. Slutsky's work can also be applied here. The idea is that in the denominator we replace the unknown λ by X/t_1. This seems sensible because the mean of X is λt_1. The result is:

$$Z' = \frac{Y - r'X}{\sqrt{r'X(1 + r')}}.$$

It can be shown that if t_1 and t_2 are both large and λ is not too close to 0, then probabilities for Z' can be well approximated by the N(0,1) curve. As a result, using the same algebra we had in Chapter 12 (the names have changed) we can expand Z' and get the following prediction interval for Y.

Result 14.5 (Predicting for a Poisson Process.) *A Poisson Process will be observed for t_2 units of time; let Y denote the number of successes that will occur. The process has been observed previously for t_1 units of time, during which x successes were counted. In the formula below, r' is given by Equation 14.6; and z^* is determined by the desired probability of the interval being correct.*

The relationship between z^* and the desired probability is given in Table 12.1 on page 296. The formula below is the approximate prediction interval for Y when the rate of the Poisson Process is unknown.

$$r'x \pm z^*\sqrt{r'x(1+r')} \tag{14.7}$$

I will illustrate the use of this formula with the Baltimore homicide data, presented in Table 13.2 on page 331. The past data consist of the $t_1 = 2$ years, 2003 and 2004. The total number of homicides, x, in those two years is $270 + 276 = 546$. The future period of interest consists of $t_2 = 1$ year, 2005. Thus, $r' = 1/2 = 0.5$. I want to have an interval for which the probability it will be correct is approximately 95%; thus, my choice for z^* is 1.96. Substituting these values into Formula 14.7, I get:

$$0.5(546) \pm 1.96\sqrt{0.5(546)(1+0.5)} = 273 \pm 39.66 = [233.34, 312.66],$$

which I round to $[233, 313]$. In words, at the beginning of 2005, using the 2003 and 2004 data, my 95% prediction interval for the number of homicides in Baltimore in 2005 is $[233, 313]$. This is a very wide interval! I imagine that if—after homicide totals of 270 and 276 in 2003 and 2004—the number declined to 233 or increased to 313 in 2005, many people would conclude that *something must have changed*. Our interval shows that such a large change is within the bounds of Poisson variation.

The actual number of homicides in 2005 turned out to be $y = 269$; thus, the prediction interval is correct.

As with Bernoulli trials with p unknown, there is no exact formula for a prediction interval. The only way to evaluate the quality of the Normal curve approximation contained in our prediction interval is to perform a simulation experiment. A simulation study is a challenge. For simplicity, I will limit my presentation of how a simulation experiment is performed to the context of the above Baltimore homicide example.

In order to do a simulation experiment we must specify one number, the rate of the Poisson Process. For the Baltimore process, I will specify that $\lambda = 270$ in my computer simulation. Each rep of the computer simulation will consist of the following three steps:

1. Simulate the value of $X \sim$ Poisson($2\lambda = 540$).
2. Use our value of X from step 1 to compute the 95% prediction interval for Y.
3. Simulate the value of $Y \sim$ Poisson($\lambda = 270$) and see whether or not the interval from step 2 is correct.

I performed this simulation with 10,000 reps and obtained 210 intervals that were too large and 253 intervals that were too small. Thus,

$$10{,}000 - (210 + 253) = \text{ 9,537—or 95.37\%—of the intervals were correct.}$$

This is a very good agreement between simulation results and the nominal probability of being correct, 95%.

14.4 Summary

This chapter introduces the notion of prediction, in the context of Bernoulli trials or a Poisson Process.

The first situation is that we plan to observe m future Bernoulli trials and we want to predict the total number of successes, Y, that will be obtained. There are two situations of interest: p known and p unknown.

For the case in which p is known, we first consider the point prediction $\hat{y}$. We adopt the criterion that we want to maximize the probability that the point prediction will be correct; i.e., we want to choose $\hat{y}$ to maximize $P(Y = \hat{y})$. The result is:

- If the mean of Y, mp, is an integer, then $\hat{y} = mp$ is the unique maximizer of the probability of obtaining a correct point prediction.
- If the mean of Y, mp, is not an integer, then calculate the probability of Y equaling each of the two integers nearest to mp. Whichever of these two integers has the larger probability of occurring is the point prediction. If they have the same probability of occurring, then I arbitrarily decide that the even integer will be the point prediction.

We looked at several examples and found that in every case the probability that the point prediction will be correct is quite small. Thus, the somewhat tedious steps outlined above for finding the point prediction are arguably somewhat (no pun intended) *pointless*. As a result, we turn our attention to finding a prediction interval for the value of Y.

The first decision is to select the desired probability that the prediction interval will be correct. The popular choices—80%, 90%, 95%, 98% and 99%—are familiar from our work on confidence interval estimation. The approximate prediction interval for Y is given in Formula 14.3, reproduced below:

$$mp \pm z^* \sqrt{mpq}.$$

In this formula, the value of z^* depends on the desired probability and is the same number we used for approximate confidence intervals.

There are three comments to remember about this approximate prediction interval. First, because we are predicting the number of successes—which must be an integer—the endpoints of the interval should be rounded-off to their nearest integers. For example, if the formula yields $[152.27, 191.33]$ we should round this to $[152, 191]$. Second, although for convenience I will always refer to the answer as an **interval**, it really isn't. For example, if we predict that Y will be in the interval $[152, 191]$, we actually are predicting that Y will take on one of the values:

$$152, 153, 154, \ldots, 191.$$

Third and finally, if $m \leq$ 1,000, once we have a prediction interval, we should use the binomial calculator website to obtain the exact probability that it will be correct. (Recall that for $m >$ 1,000, the calculator website does not yield exact binomial probabilities.)

Now we turn to the scientifically more interesting problem of finding a prediction interval for Y when p is unknown. This new problem cannot be solved unless we add a new ingredient to our

set-up. We assume that we have observed n past Bernoulli trials from the same process that will generate the m future Bernoulli trials; in particular, the p for the past is the same as the p for the future and the future is statistically independent of the past. The approximate prediction interval for Y is given in Formula 14.5, reproduced below:

$$rx \pm z^* \sqrt{rx\hat{q}}\sqrt{1+r} = rx \pm z^* \sqrt{rx(1+r)\hat{q}}.$$

In this formula, $r = m/n$, x is the total number of successes in the n past trials; z^* is the same as it was for p known; and $\hat{q} = (n-x)/n$ is the proportion of failures in the n past trials.

Lastly, we learned how to obtain a prediction interval in the context of a Poisson Process. Assume that there is a Poisson Process with unknown rate λ. The process has been previously observed for t_1 units of time, yielding a total of x successes. The process will be observed for t_2 units of time in the future and will yield Y successes. The approximate prediction interval for Y is given in Formula 14.7, reproduced below

$$r'x \pm z^* \sqrt{r'x(1+r')}.$$

In this formula, $r' = t_2/t_1$; and z^* is the same as it was for both of our earlier prediction intervals.

Finally, I want to comment on Formula 14.5. In particular, I want to explain why I present it in two different ways:

$$rx \pm z^* \sqrt{rx\hat{q}}\sqrt{1+r} \text{ and } rx \pm z^* \sqrt{rx(1+r)\hat{q}}.$$

I will refer to these as the *outside* and *inside* versions, respectively, because the former has the term $(1+r)$ within its own radical sign *outside* the first radical sign; obviously, *inside* has the term $(1+r)$ inside its only radical sign. What possible reason do I have for doing this?

The term rx is equal to

$$(m/n)x = m(x/n) = m\hat{p}.$$

Thus, the outside version is equivalent to

$$m\hat{p} \pm z^* \sqrt{m\hat{p}\hat{q}}\sqrt{(1+r)}.$$

This expression makes clear the relationship between the situations when p is unknown, Formula 14.5, and p is known, Formula 14.3. Namely, in the p *is known formula* we replace the unknown p and q by their estimates from the previous data. To make this work, we need to include the *correction term* $\sqrt{(1+r)}$. Somewhat whimsically, I like to think of this correction term as representing the *cost of ignorance*.

The inside version gives insight of the connection between the binomial and Poisson formulas for prediction. Recalling that I use r [r'] to denote the ratio of future to past for the binomial [Poisson], the prediction formulas for binomial and Poisson are identical except that the binomial formula contains the term $\hat{q}$ under its radical sign.

The derivations of Formulas 14.5 and 14.7 involve some intense algebra and the use of the fairly advanced Result 14.1. I include these *things* in these *Course Notes* for completeness; feel free to ignore them if you don't find excessive algebra to be helpful.

14.5 Practice Problems

1. The entry

   `http://en.wikipedia.org/wiki/Mendelian_inheritance`

   in Wikipedia provides a colorful illustration of the results of a dihybrid cross involving brown or white cats with short or long tails. According to Mendelian inheritance, the probability that a cat created by such a cross will be brown with a short tail is $9/16$. For the sake of this question, let's all agree that Mendelian inheritance is correct.

   I plan to observe $m = 320$ future cats created by such a cross and I want to obtain a 98% prediction interval for the total number of these cats that will have brown fur and a short tail.

2. Refer to the previous problem. According to Mendelian inheritance, the probability that a cat created by such a cross will have a long tail—regardless of fur color—is 0.25.

   I plan to observe $m = 400$ future cats created by such a cross and I want to obtain a 99% prediction interval for the total number of these cats that will have a long tail.

3. During Michael Jordan's first season with the Chicago Bulls, 1984–85, he made 630 out of 746 free throw attempts. In his last season with the Chicago Bulls, 1997–98, he made 565 out of 721 free throw attempts. Use the data from his first season to predict the number of free throws he would make during his last season. Use the 99% prediction level.

4. The purpose of this example is to show you the folly of using a small amount of data to predict a large amount of the future. On day 1, Katie obtained 56 successes in 100 trials. Use these data to obtain the 98% prediction interval for the $m = 900$ future trials on days 2–10 combined.

   After day 10, it was determined that Katie had made 579 of her shots on days 2–10 combined. Comment on your prediction interval.

5. Data for hockey player Wayne Gretzky were presented on page 333 in a Practice Problem. Use his data from the 1981–82 season—92 goals in 80 games— to predict his number of goals—71 in 80 games—in the 1982–83 season. Assume that we have a Poisson Process with unknown rate of λ goals per game. Calculate the 95% prediction interval.

6. Refer to the previous problem. In the two seasons 1981–83 combined, Gretzky scored a total of $92 + 71 = 163$ goals. given that Gretzky would play 74 games during the 1983–84 season, use the data from 1981–83 to obtain the 95% prediction interval for the number of goals he would score in 1983–84.

   Given that Gretzky scored 87 goals in 1983–84, comment on your prediction interval.

14.6 Solutions to Practice Problems

1. This is a standard problem with $p = 9/16$ known and $m = 320$. The desired probability, 98%, gives $z^* = 2.326$. Using Formula 14.3, we get

$$320(9/16) \pm 2.326\sqrt{320(9/16)(7/16)} = 180 \pm 20.64 = [159.36, 200.64],$$

which I round to $[159, 201]$.

I go to the binomial calculator website and obtain

$$P(Y \leq 201) = 0.9926 \text{ and } P(Y \leq 158) = 0.0079.$$

Thus,

$$P(159 \leq Y \leq 201) = 0.9926 - 0.0079 = 0.9847.$$

This is a bit larger than the target of 98%. Let's see if we can do better. The idea is to make the prediction interval narrower, but not much narrower. Too much narrower and the probability of being correct would fall below the target of 98%.

First, I try the interval $[160, 201]$; i.e., I increase the lower bound by one and leave the upper bound alone. I find that its probability of including Y is 0.9819 (details not given; check if you need practice with the binomial calculator).

Next, I try $[159, 200]$ and find that its probability of including Y is 0.9820.

Finally, I try $[160, 200]$; i.e., I take the original approximate interval, increase the lower bound by one and decrease the upper bound by one. I find that its probability of including Y is 0.9792. This is a bit smaller than the target probability, but it's actually the closest of the three probabilities.

Any one of these modifications of the approximate interval is fine; I (very slightly) prefer the interval $[160, 200]$ because of its symmetry around the point prediction 180.

Note: Usually at this time I tell you the value of Y and we can see whether the prediction interval is correct. Sorry, but acquiring 320 cats is too much for this dedicated author.

2. This is a standard problem with $p = 0.25$ known and $m = 400$. The desired probability, 99%, gives $z^* = 2.576$. Using Formula 14.3, we get

$$400(0.25) \pm 2.576\sqrt{400(0.25)(0.75)} = 100 \pm 22.31 = [77.69, 122.31],$$

which I round to $[78, 122]$.

I go to the binomial calculator website and obtain

$$P(Y \leq 122) = 0.9946 \text{ and } P(Y \leq 77) = 0.0039.$$

Thus,

$$P(78 \leq Y \leq 122) = 0.9946 - 0.0039 = 0.9907.$$

This is very close to the target of 99%. Thus, I won't bother to see what happens if I change either endpoint of the interval.

3. This is prediction with p unknown. The past data give $x = 630$, $n = 746$ and $\hat{q} = 116/746 = 0.155$. The future number of trials is $m = 721$, giving $r = 721/746 = 0.9665$. (Note: A weakness with this method is that at the beginning of the 1997–98 season, nobody knew the eventual value of m. One way around this was to restate the problem as "Of his first $m = 500$ free throws during the season, predict his number of successes." It is *possible* that the number m is somehow related to how well he was shooting, but I will ignore this possibility.)

   I will substitute these values into the prediction formula, Formula 14.5, noting that 99% gives $z^* = 2.576$:

$$rx \pm z^* \sqrt{rx\hat{q}}\sqrt{1+r} = 608.90 \pm 2.576\sqrt{608.90(0.155)}\sqrt{1+0.9665} =$$

$$608.90 \pm 35.09 = [573.81, 643.99],$$

   which I round to $[574, 644]$. This prediction interval is too large, because $y = 565$. Perhaps using data from 13 years earlier was a bad choice for the past data.

   In 1997–98, Jordan had his lowest free throw success percentage of his career; even worse than his two (misguided) later years in Washington.

4. This is prediction with p unknown. The past data give $x = 56$, $n = 100$ and $\hat{q} = 44/100 = 0.44$. The future number of trials is $m = 900$, giving $r = 900/100 = 9$.

   I will substitute these values into the prediction formula, Formula 14.5, noting that 98% gives $z^* = 2.326$:

$$rx \pm z^* \sqrt{rx\hat{q}}\sqrt{1+r} = 504 \pm 2.326\sqrt{504(0.44)}\sqrt{1+9} = 504 \pm 109.54 = [394.46, 613.54],$$

   which I round to $[394, 614]$. This extremely wide prediction interval is correct, because $y = 579$.

5. We have $t_1 = t_2 = 80$ games; thus $r' = 80/80 = 1$. The 95% level gives us $z^* = 1.96$. The observed value of X is 92. We substitute this information into Formula 14.7 and obtain:

$$r'x \pm z^* \sqrt{r'x(1+r')} = 92 \pm 1.96\sqrt{92(1+1)} = 92 \pm 26.6 = [65, 119],$$

   after rounding. This very wide interval is correct because it includes $y = 71$. I opine that many hockey fans would interpret the change from 92 to 71 goals as significant, but it falls within the range of Poisson variation.

6. We have $t_1 = 80 + 80 = 160$ games and $t_2 = 74$ games; thus $r' = 74/160 = 0.4625$. The 95% level gives us $z^* = 1.96$. The observed value of X is 163, giving us $r'x = 75.39$. We substitute this information into Formula 14.7 and obtain:

$$r'x \pm z^* \sqrt{r'x(1+r')} = 75.39 \pm 1.96\sqrt{75.39(1+0.4625)} = 75.39 \pm 20.6 = [55, 96],$$

   after rounding. This very wide interval is correct because it includes $y = 87$.

14.7 Homework Problems

1. Refer to Practice Problems 1 and 2. Let A be the event that the cat has white fur and let B be the event that the cat has a long tail.

   (a) Define a success to be that event (A or B) occurs. Remember, or means and/or. According to Mendelian inheritance, what is the probability of a success?

   (b) For $m = 160$ cats created by such a cross, calculate the 95% prediction interval for the number of successes.

   (c) Use the binomial calculator website to determine the exact probability that your interval in (b) will be correct.

2. On days 1–19, Katie made 1,227 out of 1,900 attempted shots. Use these 1,900 observations to calculate the 99% prediction interval for the number of shots, out of $m = 100$, that she would make on day 20.

   Given that Katie made $y = 71$ shots on day 20, comment on your prediction interval.

3. Refer to the data on traffic accidents in Table 13.3 on page 336. Use the combined data from years 2005–2008, 405 crashes with bikes, to obtain the 95% prediction interval for the number of bike crashes in 2009.

   Given that the number of bike crashes in 2009 was $y = 115$, comment on your prediction interval.

Chapter 15

Comparing Two Binomial Populations

In this chapter and the next, our data are presented in a 2×2 contingency table. As you will learn, however, not all 2×2 contingency tables are analyzed the same way. I begin with an introductory section.

15.1 The Ubiquitous 2×2 Table

I have always liked the word *ubiquitous*; and who can argue taste? According to dictionary.com the definition of ubiquitous is:

> Existing or being everywhere, especially at the same time; omnipresent.

Table 15.1 is a partial recreation of Table 8.5 on page 170 in Chapter 8 of these notes.

Let me explain why I refer to this table as *ubiquitous*. You will learn of many different scientific scenarios that yield data of the form presented in Table 15.1. Depending on the scenario, you will learn the different appropriate ways to summarize and analyze the data in this table.

I say that Table 15.1 is only a partial recreation of Chapter 8's table because of the following changes:

1. In Chapter 8, the rows were treatments 1 and 2; in Table 15.1, they are simply called rows 1 and 2. In some scenarios, these rows will represent treatments and in some scenarios they won't.

2. In Chapter 8, the columns were the two possible responses, success and failure; in Table 15.1, they are simply called columns 1 and 2. In some scenarios, these columns will represent **the** response and in some scenarios they won't.

3. The table in Chapter 8 also included row proportions that served two purposes: they described/summarized the data; and they were the basis (via the computation of $x = \hat{p}_1 - \hat{p}_2$) for finding the observed value of the test statistic, X, for Fisher's test. Again, in some scenarios in this and the next chapter we will compute row proportions, and in some scenarios we won't.

Table 15.1: The general notation for the ubiquitous 2×2 contingency table of data.

	Column		
Row	1	2	Total
1	a	b	n_1
2	c	d	n_2
Total	m_1	m_2	n

Here are the main features to note about the ubiquitous 2×2 contingency table of data.

- The values a, b, c and d are called the cell counts. They are necessarily nonnegative integers.
- The values n_1 and n_2 are the row totals of the cell counts.
- The values m_1 and m_2 are the column totals of the cell counts.
- The value n is the sum of the four cell counts; alternatively, it is the sum of the row [column] totals.

Thus, there are nine counts in the 2×2 contingency table, all of which are determined by the four cell counts.

15.2 Comparing Two Populations; the Four Types of Studies

The first appearance of the ubiquitous 2×2 contingency table was in Chapter 8 for a CRD with a dichotomous response. Recall that Fisher's Test is used to evaluate the Skeptic's Argument. As stated at the beginning of this Part II of these notes, a limitation of the Skeptic's Argument is that it is concerned *only* with the units under study. In this section, you will learn how to extend the results of Chapter 8 to populations. In addition, we will extend results to observational studies that, as you may recall from Chapter 1, do not involve randomization.

In Chapter 8 the units can be trials or subjects. The listing below summaries the studies of Chapter 8.

- **Units are subjects:** The infidelity study; the prisoner study; and the artificial Headache Study-2.
- **Units are trials:** The golf putting study.

The idea of a population depends on the type of unit. In particular,

- When units are subjects, we have a finite population. The members of the finite population comprise all potential subjects of interest to the researcher.
- When units are trials, we assume that they are Bernoulli Trials.

The number **four** in the title of this section is obtained by multiplying 2 by 2. When we compare two populations both populations can be Bernoulli trials or both can be finite populations. In addition, as we shall discuss soon, a study can be **observational** or **experimental**. Combining these two dichotomies, we get four types of study; for example an observational study on finite populations.

It turns out that the mathematical formulas are identical for the four types of studies, but the *interpretation* of our analysis depends on the type of study. In addition, the assumptions are somewhat different in the four settings.

We begin with an **observational study on two finite populations**. This study was published in 1988; see [1].

Example 15.1 (The Dating study.) *The first finite population is undergraduate men at at the University of Wisconsin–Madison and the second population is undergraduate men at Texas A&M University. Each man's response is his answer to the following question:*

> *If a woman is interested in dating you, do you generally prefer for her: to* **ask** *you out; to* **hint** *that she wants to go out with you; or to* **wait** *for you to act.*

After data were collected, it was found that only two or three (I can't remember for sure) of the 207 subjects selected wait *for the response. As a result, the researchers decided that* **ask** *is a success and either of the other responses is a failure. The purpose of the study is to compare the proportion of successes at Wisconsin with the proportion of successes at Texas A&M.*

These two populations obviously fit our definition of finite populations. Why is it called observational? The dichotomy of observational/experimental refers to the *control* available to the researcher. Suppose that Matt is a member of one of these populations. As a researcher, I have control over whether I have Matt in my study, but I do *not* have control over the population to which he belongs. Consistent with our usage in Chapter 1, the variable that determines a subject's population, is called the **study factor**. In the current example, the study factor is school attended and it has two **levels**: Wisconsin and Texas A&M. This is an observational factor, sometimes called, for obvious reasons, a classification factor, because each subject is classified according to his school.

Table 15.2 presents the data from this *Dating Study*. Please note the following *decisions* that I made in creating this table.

1. Similar to our tables in Chapter 8, the columns are for the response and the rows are for the levels of the study factor; i.e., the populations. Note that because the researcher did not assign men to university by randomization, we **do not** refer to the rows as treatments.

2. As in Chapter 8, I find the row proportions to be of interest. In particular, we see that 56% of the Wisconsin men sampled are successes compared to only 31% of the Texas men.

Next, we have an **experimental study on two finite populations**. Below is my slight alteration of an actual study of Crohn's disease that was published in 1988; see [2]. I have taken the original sample sizes, 37 and 34, and made them both 40. My values of the $\hat{p}$'s differ from the original ones by less than 0.005. Why did I make these changes?

Table 15.2: Data from the Dating study.

	Counts			Row Proportions		
	Prefer:			Prefer:		
Population	Ask	Other	Total	Ask	Other	Total
Wisconsin	60	47	107	0.56	0.44	1.00
Texas A&M	31	69	100	0.31	0.69	1.00
Total	91	116	207			

Table 15.3: Data from the modified study of Crohn's disease.

	Counts			Row Proportions		
	Response:			Response:		
Population	S	F	Total	S	F	Total
Drug C	24	16	40	0.600	0.400	1.000
Placebo	13	27	40	0.325	0.675	1.000
Total	37	43	80			

1. I can't believe that a 25 year-old study, albeit apparently a very good study, is the *final word* in the treatment of Crohn's disease. Thus, I don't see much *upside* in preserving the exact data.

2. The computations become much friendlier—I am not a big fan of dividing by 37—by changing both sample sizes to 40. Perhaps a minor point, but why pass on a chance to reduce the tedium of hand computations?

3. Most importantly, I want to use this example to make some general comments on how the WTP assumption relates to medical studies. I don't have perfect knowledge of how the actual study was performed and I don't want to be unfairly critical of a particular study; instead, I will be fair in my criticism of my version of the study.

Example 15.2 (A study of Crohn's Disease.) *Medical researchers were searching for an improved treatment for persons with* **Crohn's disease***. They wanted to compare a new therapy, called* **drug C***, to an inert drug, called a* **placebo***. Eighty persons suffering from Crohn's disease were available for the study. Using randomization, the researcher assigned 40 persons to each treatment. After three months of treatment, each person was examined to determine whether his/her condition had improved (a success) or not (a failure).*

The data from this study of Crohn's disease is presented in Table 15.3. There is a very important distinction between the study of Crohn's disease and the Dating study. Below we are going to talk about comparing the *drug C population* to the *placebo population.* But, as we shall see, and perhaps is already obvious, there is, in reality, neither a *drug C population* nor a *placebo*

population. Certainly not in the physical sense of there being a University of Wisconsin and a Texas A&M University.

Indeed, as I formulate a *population approach* to this medical study, the only population I can imagine is one **superpopulation** of all persons, say in the United States, who have Crohn's disease. This superpopulation gives rise to two imaginary populations: first, imagine that everybody in the superpopulation is given drug C and, second, imagine that everybody in the superpopulation is given the placebo.

To summarize the differences between observational and experimental studies:

1. For observational studies, there exists two distinct finite populations. For experimental studies, there exists two **treatments** of interest and one superpopulation of subjects. The two populations are generated by imagining what would happen if each member of the superpopulation was assigned each treatment. (It is valid to use the term treatment because, as you will see, an experimental study always includes randomization.)

2. Here is a very important consequence of the above: For an observational study, the two populations have distinct members, but for an experimental study, the two populations consist of the same members.

Let me illustrate this second comment. For the Dating study, the two populations are comprised of different men—Bubba, Bobby Lee, Tex, and so on, for one population; and Matt, Eric, Brian, and so on, for the other population. For the Crohn's study, both populations consist of the same persons, namely the persons in the superpopulation.

15.3 Assumptions and Results

We begin with an **observational study on finite populations.** Assume that we have a random sample of subjects from each population and that the samples are independent of each other. For our Dating study, independence means that the method of selecting subjects from Texas was totally unrelated to the method used in Wisconsin. *Totally unrelated* is, of course, rather vague, but bear with me for now.

The sample sizes are n_1 from the first population and n_2 from the second population. We define the random variable X [Y] to be the total number of successes in the sample from the first [second] population. Given our assumptions, $X \sim \text{Bin}(n_1, p_1)$ and $Y \sim \text{Bin}(n_2, p_2)$, where p_i is the proportion of successes in population i, $i = 1, 2$.

Always remember that you can study the populations separately using the methods of Chapter 12. The purpose of this chapter is to *compare* the populations, or, more precisely, to compare the two p's. We will consider both estimation and testing.

For estimation, our goal is to estimate $p_1 - p_2$. Define point estimators $\hat{P}_1 = X/n_1$ and $\hat{P}_2 = Y/n_2$. The point estimator of $p_1 - p_2$ is

$$W = \hat{P}_1 - \hat{P}_2 = X/n_1 - Y/n_2.$$

A slight modification of Result 14.1 (I won't give the details) yields the mean and variance of W:

$$\mu_W = \mu_X/n_1 - \mu_Y/n_2 = n_1p_1/n_1 - n_2p_2/n_2 = p_1 - p_2, \text{ and}$$

$$\text{Var}(W) = \text{ Var}(X)/n_1^2 + \text{ Var}(Y)/n_2^2 = p_1q_1/n_1 + p_2q_2/n_2.$$

Thus, it is easy to standardize W:

$$Z = \frac{(\hat{P}_1 - \hat{P}_2) - (p_1 - p_2)}{\sqrt{(p_1q_1)/n_1 + (p_2q_2)/n_2}}. \tag{15.1}$$

It can be shown that if both n_1 and n_2 are large and neither p_i is too close to either 0 or 1, then probabilities for Z can be well approximated by using the N(0,1) curve. Slutsky's theorem also applies here. Define

$$Z' = \frac{(\hat{P}_1 - \hat{P}_2) - (p_1 - p_2)}{\sqrt{(\hat{P}_1\hat{Q}_1)/n_1 + (\hat{P}_2\hat{Q}_2)/n_2}}, \tag{15.2}$$

where $\hat{Q}_i = 1 - \hat{P}_i$, for $i = 1, 2$. Subject to the same conditions we had for Z, probabilities for Z' can be well approximated by using the N(0,1) curve. Thus, using the same algebra we had in Chapter 12, Formula 15.2 can be expanded to give the following two-sided confidence interval estimate of $(p_1 - p_2)$:

$$(\hat{p}_1 - \hat{p}_2) \pm z^*\sqrt{(\hat{p}_1\hat{q}_1)/n_1 + (\hat{p}_2\hat{q}_2)/n_2}. \tag{15.3}$$

I will use this formula to obtain the 95% confidence interval estimate of $(p_1 - p_2)$ for the Dating study. First, 95% confidence gives $z^* = 1.96$. Using the summaries in Table 15.2 we get the following:

$$(0.56 - 0.31) \pm 1.96\sqrt{(0.56)(0.44)/107 + (0.31)(0.69)/100} =$$

$$0.25 \pm 1.96(0.0666) = 0.25 \pm 0.13 = [0.12, 0.38].$$

Let me briefly discuss the interpretation of this interval.

The first thing to note is that the confidence interval **does not include zero**; all of the numbers in the interval are positive. Thus, we have the qualitative conclusion that $(p_1 - p_2) > 0$, which we can also write as $p_1 > p_2$. Next, we turn to the question: *How much larger?*. The endpoints of the interval tell me that p_1 is at least 12 percentage points and at most 38 percentage points larger than p_2. I remember, of course, that my confidence level is 95%; thus, approximately 5% of such intervals, in the long run, will be incorrect.

I choose to make no assessment of the *practical importance* of p_1 being between 12 and 38 percentage points larger than p_2; although I enjoy the Dating study, to me it is essentially frivolous. (Professional sports, in my opinion, are essentially frivolous too. Not everything in life needs to be serious!)

For a test of hypotheses, the null hypothesis is H_0: $p_1 = p_2$. There are three choices for the alternative:

$$H_1\text{: } p_1 > p_2;\ H_1\text{: } p_1 < p_2;\ \text{or } H_1\text{: } p_1 \neq p_2.$$

Recall that we need to know how to compute probabilities given that the null hypothesis is true. First, note that if the null hypothesis is true, then $(p_1 - p_2) = 0$. Making this substitution into Z in Equation 15.1, we get:

$$Z = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{(p_1 q_1)/n_1 + (p_2 q_2)/n_2}}. \tag{15.4}$$

We again have the problem of unknown parameters in the denominator. For estimation we used Slutsky's results and handled the two unknown p's separately. But for testing, we proceed a bit differently.

On the assumption that the null hypothesis is true, $p_1 = p_2$; let's denote this common value by p (nobody truly loves to have subscripts when they aren't needed!). It seems obvious that we should **combine** the random variables X and Y to obtain our point estimator of p. In particular, define

$$\hat{P} = (X + Y)/(n_1 + n_2) \text{ and } \hat{Q} = 1 - \hat{P}.$$

We replace the unknown p_i's [q_i's] in Equation 15.4 with $\hat{P}$ [$\hat{Q}$] and get the following test statistic.

$$Z = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{(\hat{P}\hat{Q})(1/n_1 + 1/n_2)}}. \tag{15.5}$$

Assuming that n_1 and n_2 are both large and that p is not too close to either 0 or 1, probabilities for Z in this last equation can be well approximated with the N(0,1) curve.

The observed value of the test statistic Z is:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{(\hat{p}\hat{q})(1/n_1 + 1/n_2)}}, \tag{15.6}$$

The rules for finding the approximate P-value are given below.

- For H_1: $p_1 > p_2$: The approximate P-value equals the area under the N(0,1) curve to the right of z.
- For H_1: $p_1 < p_2$: The approximate P-value equals the area under the N(0,1) curve to the left of z.
- For H_1: $p_1 \neq p_2$: The approximate P-value equals twice the area under the N(0,1) curve to the right of $|z|$.

I will demonstrate these rules with the Dating study.

Personally, I would have chosen the alternative $p_1 > p_2$, but I will calculate the approximate P-value for all three possibilities. I begin by calculating $\hat{p} = (60 + 31)/(107 + 100) = 0.44$, giving $\hat{q} = 0.56$. The observed value of the test statistic is

$$z = \frac{0.25}{\sqrt{(0.44)(0.56)(1/107 + 1/100)}} = \frac{0.25}{0.0690} = 3.62.$$

Using a Normal curve website calculator, the P-value for $p_1 > p_2$ is 0.00015; for $p_1 < p_2$ it is 0.99985; and for $p_1 \neq p_2$ it is $2(0.00015) = .00030$.

The Fisher's test site can be used to obtain the exact P-value, using the same method you learned in Chapter 8. Using the website I get the following exact P-values: 0.00022, 0.99993 and 0.00043. The Normal curve approximation is pretty good.

(Technical Note: Feel free to skip this paragraph. There is some disagreement among statisticians as to whether the numbers from the Fisher's website should be called the **exact** P-values. Note, as above, I will call them exact because it is a convenient distinction from using the N(0,1) curve. The P-values are exact only if one decides to condition of the values of m_1 and m_2 in Table 15.1. Given the Skeptic's Argument in Part I of these notes, these values **are fixed**, but for the independent binomial sampling of this section, they are not. Statisticians debate the importance of the information lost if one conditions on the observed values of m_1 and m_2. It is not a big deal for this course; I just want to be intellectually honest with you.)

We will turn now to the study of Crohn's disease, our example of an **experimental study on finite populations**. Because the two populations do not actually exist in the physical world, we modify our sampling a bit.

- Decide on the numbers n_1 and n_2, where n_i is the number of subjects who will be given treatment i. Calculate $n = n_1 + n_2$, the total number of subjects who will be in the study.
- Select a smart random sample of n subjects from the superpopulation; i.e., we need to obtain n distinct subjects.
- Divide the n subjects selected for study into two treatment groups by **randomization**. Assign n_1 subjects to the first treatment and n_2 subjects to the second treatment.

If we now turn to the two imaginary populations, we see that our samples are not quite independent. The reason is quite simple. A member of the superpopulation, call him Ralph, *cannot* be given both treatments. Thus, if, for example, Ralph is given the first treatment he cannot be given the second treatment. Thus, knowledge that Ralph is in the sample from the first population tells us that he is not in the sample from the second population; i.e., the samples depend on each other. But if the superpopulation has a large number of members compared to n, which is usually the case in practice, then the dependence between samples is very weak and can be safely ignored, which is what we will do.

Ignoring the slight dependence between samples, we can use the same estimation and testing methods that we used for the Dating study. The details are now given for the study of Crohn's disease. The data I use below can be found in Table 15.3.

Here is the 95% confidence interval estimate of $p_1 - p_2$:

$$(0.600-0.325)\pm 1.96\sqrt{(0.600)(0.400)/40 + (0.325)(0.675)/40} = 0.275\pm 0.210 = [0.065, 0.485].$$

As with the confidence interval for the Dating study, let me make a few comments. This interval does not include zero. I conclude that p_1 is at least 6.5 and at most 48.5 percentage points larger than p_2. While the researchers were no doubt pleased to conclude that drug C is superior to the

placebo, this interval is too wide to be of practical value. Like it or not, modern medicine must pay attention to *cost-benefit analyses* and the benefit at 6.5 percentage points is much less than the benefit at 48.5 percentage points. Ideally, more data should be collected, which will result in a narrower confidence interval.

For the test of hypotheses, I choose the first alternative, $p_1 > p_2$. Using the Fisher's test website, the exact P-value is 0.0122. To use the Normal curve approximation, first we need $\hat{p} = 37/80 = 0.4625$. Plugging this into Equation 15.6, we get

$$Z = \frac{0.275}{\sqrt{(0.4625)(0.5375)[1/40 + 1/40]}} = \frac{0.275}{0.11149} = 2.467.$$

From the Normal curve calculator, the approximate P-value is 0.0068. This is not a very good approximation of the exact value, 0.0122, because n_1 and n_2 are both pretty small. There is a continuity correction that improves the approximation, but it's a bit of a mess; given that we have the Fisher's test website, I won't bother with the continuity correction.

15.3.1 'Blind' Studies and the Placebo Effect

Well, if I ever become **tsar** of the research world I will eliminate the use of the word **blind** to describe studies! Until that time, however, I need to tell you about it. (By the way, I would replace *blind* by *ignorant* which, as you will see, is a much more accurate description.)

Look at the data from our study of Crohn's disease, in Table 15.3 on page 356. Note the value $\hat{p}_2 = 0.325$, which is just a bit smaller than one-third. In words, of the 40 persons given the inert drug—the placebo—nearly one-third improved! This is an example of what is called **the placebo effect**. Note that I will not attempt to give a precise definition of the placebo effect. For more information, see the internet or consult one of your other professors.

It is natural to wonder:

> Why did the condition improve for 32.5% of the subjects on the placebo?

Two possible explanations come to mind immediately:

1. Like many diseases, if one suffers from Crohn's disease, then one will have good days and bad days. Thus, for some of the 13 persons on the placebo who improved, the improvement might have been due to routine variation in symptoms.

2. For some of the patients, being in a medical study and receiving attention might provide a psychological boost, resulting in an improvement.

Regarding the second item above, if you don't know about studies being **blind**, you might think,

> Hey, if the physician tells me that I am receiving a placebo, how is this going to help my outlook? Indeed, I might be a little annoyed about it!

Our study of Crohn's disease was **double blind**.

The first blindness was that the 80 subjects were blind to—ignorant of—the treatment they received. Each subject signed some kind of *informed consent* document that included a statement that the subject *might* receive drug C or *might* receive a placebo. This is actually a very good feature of the study. Because of this blindness, it is reasonable to view the difference, $\hat{p}_1 - \hat{p}_2 = 0.275$, as a measure of the biochemical effectiveness of drug C. Without blindness, this difference measures biochemical effectiveness **plus** the psychological impact of knowing one is receiving an active drug rather than an inactive one. Without blindness, how much of the 27.5 percentage point improvement is due to biochemistry? Who knows? If not much, then how can one honestly recommend drug C?

Of course, it is easy to say, "I will have my subjects be blind to treatment;" it might not be so easy to achieve. For example, if drug C has a nasty side-effect, subjects might infer their treatment depending on the presence or absence of the side-effect.

I said that our study of Crohn's disease was double blind. The second blindness involved the evaluator(s) who determined each subject's response. An assessment of *improvement*, either through examining biological material or asking the subject a series of questions, can be quite subjective. It is better if the person making the assessment is ignorant—again, a much better word than blind—of the subject's treatment.

Not all medical studies *can* be blind. For example, in a study of breast cancer, the treatments were mastectomy versus lumpectomy. Obviously, the patient will know which treatment she receives! If blindness is possible, however, the conventional wisdom is that it's a good idea.

(This reminds me of an example of what passes for humor among statisticians: A triple blind study is one in which the subjects are blind to treatment; the evaluators are blind to treatment; and the statistician analyzing the data is blind to treatment!)

In medical studies, on humans especially, there are always ethical issues. In the early days of research for a treatment for HIV infection, for example, many people felt that it was highly unethical to give a placebo to a person who is near death. I am not equipped to help you navigate such treacherous waters, but I do want to make a comment about our study of Crohn's disease. Looking at the data on the 40 persons who received drug C, we see that 60% were improved. Sixty percent sounds less impressive, however, when you are reminded that 32.5% improved on the placebo. Thus, a placebo helps us to gauge how wonderful the new therapy actually is.

There is a final feature of the actual study of Crohn's disease, a feature that makes me admire the researchers even more. The data in Table 15.3 must have made the researchers happy. Their proposal—treat Crohn's disease with drug C—was shown to be correct, based on the criterion of statistical significance. It would have been easy after collecting their data to write-up their results and submit them for publication.

Instead, the researchers chose to collect data again, three months after the treatment period had ended. These *follow-up* data (modified, again, by me for reasons given earlier) are presented in Table 15.4. The follow-up data are statistically significant—the exact P-value for $>$ is 0.0203, details not shown—but the benefits of both drug C and the placebo diminished over time. I am not a physician, but this suggests that drug C should perhaps be viewed as a maintenance treatment rather than a one-time treatment.

Table 15.4: Modified data from the study of Crohn's disease at follow-up.

	Observed Frequencies			Row Proportions		
	Response:			Response:		
Population	S	F	Total	S	F	Total
Drug C	15	25	40	0.375	0.625	1.000
Placebo	6	34	40	0.150	0.850	1.000
Total	21	59	80			

15.3.2 Assumptions, Revisited

In this brief subsection I will examine the WTP assumption for both the Dating study and our study of Crohn's disease.

From all I have read, heard and observed, the University of Wisconsin–Madison has a strong Department of Psychology. Thus, please do not interpret what follows as a criticism.

I used the Dating study in my in-person classes for almost 20 years. My students seem to have felt that it is a *fun study*. We discussed how the data were collected at Madison; what follows is based on what my students told me over the years.

Psychology 201, *Introduction to Psychology*, is a very popular undergraduate course. Students tended to take **my** course during the junior or senior year. Every semester, a strong majority of my students had finished or were currently enrolled in Psychology 201; a strong majority of those students had taken it during their first year on campus. Obviously, I had no access to students who took Psychology 201 during, for example, their last semester on campus unless, of course, they also took my class during their last semester on campus. Despite the caveat implied above, it is most likely accurate to say that a majority of students in Psychology 201 take it during their first year on campus.

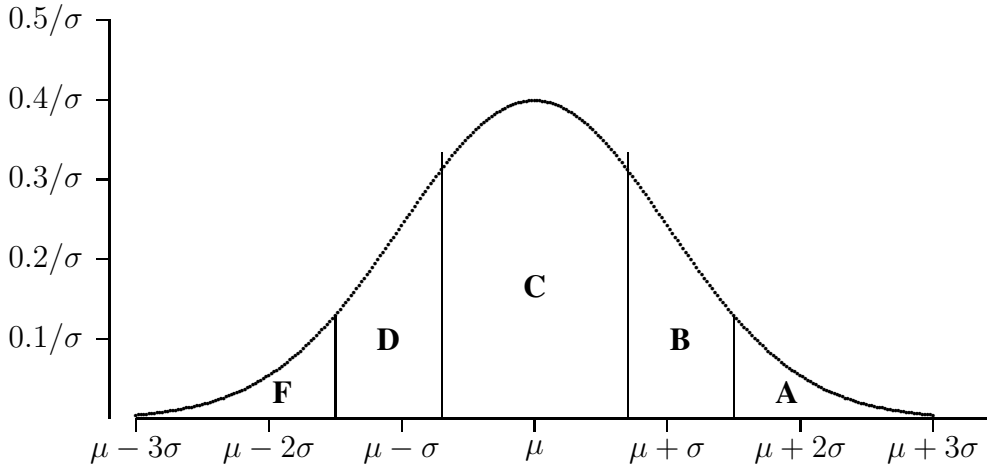
On the first day of lecture in Psychology 201 the instructor draws a Normal curve on the board—Normal curves are **big** in psychology—similar to the one I present to you in Figure 15.1. As I have done in this figure, the instructor divides the curve into five regions, with a grade assigned to each region. The instructor tells the class that this curve represents the distribution of final grades in the class: equal numbers of A's and F's; equal numbers of B's and D's; C the most common grade; and A and F tied for rarest grade. This presentation, I have been told, really gets the students' attention! Having been thus psychologically prepared, the students are then told how to avoid this curve: sign-up for extra credit. After class they race into the hallway outside the lecture hall to enroll as a subject in one or more research studies. Including, once upon a time, the Dating study.

Again, let me emphasize that I am not being critical of this method of recruiting subjects; chemists need beakers, microscopes and chemicals; psychologists need subjects. In addition, I suspect that the experience of being a subject is an invaluable part of one's education.

The point of the above is the following: First year men are almost certainly overrepresented in the group of 107 Wisconsin men in the Dating study. Does this matter? Should it affect our willingness to make the WTP assumption?

I polled my students over the years, and while there was some disagreement, a solid majority

Figure 15.1: The promised (threatened?) distribution of grades in Psychology 201.



believed that first year men are much more likely than other men to be successes. Something about the excitement and newness of being on campus. I suspect that you have much better direct knowledge of young men on campus; thus, I will end my speculations. Except to say that if my students are correct, then the 56% success rate among Wisconsin men in Table 15.2 is likely to be an overestimate of the true p_1 for all Wisconsin men.

I have no idea how the men from Texas A & M were selected for the Dating study. Perhaps at Texas A & M, the first year of study is devoted to farming and/or military activities—after all, its original name is the Agricultural and Mechanical College of Texas with an education that consisted of agricultural and military techniques. (Texas A & M is a fine modern university with a very strong Statistics Department.)

I will return to these issues surrounding the Dating study later in this chapter when we consider **Simpson's Paradox**.

Let's turn our attention now to our study of Crohn's disease. Researchers in such studies don't even try to obtain a random sample of subjects for study. Here are some reasons why:

1. There does not exist a listing of all persons in the United States with Crohn's disease. If there was such a list. it would be inaccurate soon after it is created.

2. If there was a listing of the population, it would be easy to select 80 persons at random, but these subjects would be scattered across the country and we could not **force** them to go through the inconvenience of being in our study.

3. Our study of Crohn's disease was performed to determine whether or not drug C has any benefit. Thus, we would not want to spend a huge amount of money and effort to obtain a sample that is nearly random, only to find that the drug is ineffective.

Now I am going to tell you a really nice feature of our study of Crohn's disease, a feature that indeed is present in all of our experimental studies.

If you decide to make the WTP assumption, then you can conclude that drug C is superior to the placebo, overall, for the entire population of persons suffering from Crohn's disease. You can even use the confidence interval to proclaim *how much better* drug C is than the placebo.

If, however, you decide **not** to make the WTP assumption, then you can revert to a Part I analysis of the Skeptic's Argument. The P-value for the population-based inference is the same as the P-value for the randomization-based inference. This is the nice feature: for an experimental study we always have a valid Part I analysis; if we are willing to make the WTP assumption, our analysis becomes richer and more useful, at the risk of being less valid if, indeed, our belief in WTP is seriously misguided.

15.4 Bernoulli Trials

In the previous section, I presented two of the promised four types of studies, namely the studies—observational or experimental—on finite populations. In this section I will tell you about observational and experimental studies on Bernoulli trials.

The distinction between observational and experimental studies with Bernoulli trials confuses many people. It's a good indicator of whether somebody thinks like a mathematician or a scientist.

Mathematicians think as follows: Given that we assume we have two sequences of Bernoulli trials, both p_1 and p_2 are constant and there is no memory. Thus, one can prove (and they are correct on this) that there is no reason to randomize.

Scientists think as follows: In a real problem we can never be certain that we have Bernoulli trials and frequently we have serious doubts about it. As a result, if we *can* randomize, it adds another *level of validity* to our findings, similar to what I stated at the end of the last section when discussing our study of Crohn's disease.

The good news is that Formula 15.3 on page 358, our earlier confidence interval estimate of $p_1 - p_2$, is valid in this section too. In addition, the Fisher's test website gives exact P-values. Alternatively, you may use the N(0,1) curve to obtain approximate P-values, using the rules given on page 359 and the observed value of the test statistic, z, given in Equation 15.6 on page 359.

I will begin with an example of an **experimental study on Bernoulli trials**.

Example 15.3 (Clyde's study of 3-point shooting.) *For his project for my class, former star college basketball player, Arnold (Clyde) Gaines studied his ability to shoot baskets. He decided to perform a balanced CRD with a total of $n = 100$ shots. His treatments were the locations of the attempted shot, both behind the three-point line. Treatment 1 was shooting from the front of the basket and treatment 2 was shooting from the left corner of the court.*

Clyde's data are in Table 15.5. First, I note, descriptively, that Clyde's performance was almost identical from the two locations. Using Fisher's test, the exact P-values are: 0.5000 for $>$; 0.6577 for $<$; and 1 for $\neq$. The confidence interval, however, provides an interesting insight.

The 95% confidence interval estimate of $p_1 - p_2$ is

$$(0.42 - 0.40) \pm 1.96\sqrt{(0.42)(0.58)/50 + (0.40)(0.60)/50} = 0.02 \pm 0.19 = [-0.17, 0.21].$$

Table 15.5: Data from Clyde's Three-point basket study.

	Observed Frequencies			Row Proportions		
Population	Basket	Miss	Total	Basket	Miss	Total
Front	21	29	50	0.42	0.58	1.00
Left Corner	20	30	50	0.40	0.60	1.00
Total	41	59	100			

Table 15.6: Data from the High-temperature forecast study.

	Counts			Row Proportions		
Population	S	F	Total	S	F	Total
Spring	46	43	89	0.517	0.483	1.00
Summer	50	39	89	0.562	0.438	1.00
Total	96	82	178			

Let's examine this interval briefly. Unlike our intervals for the Dating and Crohn's data, this interval includes zero. It also includes both negative and positive numbers. I label such an interval **inconclusive** because it states that p_1 might be **less than, equal to or greater than** p_2. Next, I look at the endpoints. The upper bound, 0.21, tells me that p_1 might be as much as 21 percentage points larger than p_2. The lower bound, -0.17, tells me that p_1 might be as much as 17 percentage points smaller than p_2.

This interval is so wide that, in my opinion, it has no practical value. Here is what I mean. Based on my *expertise* on basketball, I would be amazed if the two p's for a highly skilled basketball player differed by more than 0.15. (Feel free to disagree with me.) As a result, this confidence interval tells me **less** than what I 'knew' before the data were collected. Clearly, to study this problem one needs many more than $n = 100$ shots. (We would reach the same conclusion by performing a power analysis, but I won't bother you with the details.)

Next, we have an example of an **observational study on Bernoulli trials**. Having lived in Michigan or Wisconsin my entire life, I had noted that weather seems to be less predictable in spring than in summer. In 1988, I collected some data to investigate this issue.

Example 15.4 (High temperature prediction in Madison, Wisconsin.) *Every day, the morning Madison newspaper would print a predicted high temperature for that day and the actual high for the previous day. Using these data over time, one could evaluate how well the predictions performed. I arbitrarily decided that a predicted high that came within two degrees of the actual high was a success and all other predictions were failures. Thus, for example, if the predicted high was 60 degrees; then an actual high between 58 and 62 degrees, inclusive, would be a success; any other actual high would be a failure.*

Table 15.6 presents the data that I collected. The summer predictions were better (descriptively),

but not by much. I found this surprising. My choice of alternative is $<$, which has exact P-value equal to 0.3260. The other exact P-values are 0.7739 for $>$; and 0.6520 for $\neq$.

The 95% confidence interval estimate of $p_1 - p_2$ is

$$(0.517 - 0.562) \pm 1.96\sqrt{(0.517)(0.483)/89 + (0.562)(0.438)/89} =$$

$$-0.045 \pm 0.146 = [-0.191, 0.101].$$

I am not an expert at weather forecasting; thus, I cannot really judge whether this confidence interval is useful scientifically, but I doubt that it is because it is so wide.

15.5 Simpson's Paradox

The most important difference between an observational and experimental study is in how we interpret our findings. Let us compare and contrast the Dating study and the study of Crohn's disease.

In both studies we concluded that the populations had different p's. Of course these conclusions could be wrong, but let's ignore that issue for now. We have concluded that the populations are different, so it is natural to wonder **why are they different?**

In the Dating study we don't know why. Let me be clear. We have concluded that Wisconsin men and Texas men have very different attitudes, but we don't know why. Is it because the groups differ on:

Academic major? Ethnicity? Religion? Liberalism/Conservatism? Wealth?

We don't know why. Indeed, perhaps the important differences are in the **women** mentioned in the question. Perhaps being asked out by a Texas woman is very different from being asked out by a Wisconsin woman.

The above comments (not all of which are silly!) are examples of what is true for any observational study. We can conclude that the two populations are different, but we don't know why.

Let us contrast the above with the situation for our study of Crohn's disease. In this study, the two populations consist of exactly the same subjects! Thus, the only possible explanation for the difference between populations is that drug C is better than the placebo. (This is a good time to remember that our conclusion that the populations differ could be wrong.)

Simpson's Paradox (no, not named for Homer, Marge, Bart, Maggie, Lisa or even O.J.) provides another, more concrete, way to look at this same issue.

Years ago, I worked as an expert witness in several cases of workplace discrimination. As a result of this work, I was invited to make a brief presentation at a continuing education workshop for State of Wisconsin administrative judges. (In Wisconsin, the norm was to have workplace discrimination cases settled administratively rather than by a jury of citizens.) Below I am going to show you what I presented in my 10 minutes. Note that my analysis below is *totally descriptive, not inferential*. For the current discussion, I don't really care whether the data come from a population and, if they do, whether or not the WTP assumption seems reasonable. I simply want to show you possible *hidden patterns* in data.

Table 15.7: Hypothetical data from an observational study. The study factor is the sex of the employee. The response is whether the employee was released (lost job) or not.

	Released?			
Sex	Yes	No	Total	$\hat{p}$
Female	60	40	100	0.60
Male	40	60	100	0.40
Total	100	100	200	

Table 15.8: **Case 1**: Data from Table 15.7 with a background factor—job type—added.

	Job A					**Job B**			
	Released?					Released?			
Sex	Yes	No	Total	$\hat{p}$	Sex	Yes	No	Total	$\hat{p}$
Female	30	20	50	0.60	Female	30	20	50	0.60
Male	20	30	50	0.40	Male	20	30	50	0.40
Total	50	50	100			50	50	100	

These are totally and extremely hypothetical data. A company with 200 employees decides it must reduce its work force by one-half. Table 15.7 reveals the relationship between the sex of the worker and the employment outcome. The table shows that the proportion of women who were released was 20 percentage points larger than the proportion of men who were released. This is an observational study—the researcher did not assign, say, Sally to be a woman by randomization. This means that we do not know *why* there is a difference. In particular, it would be presumptuous to say it is because of *discrimination*. (Aside: There is a legal definition of discrimination and I found that there is one thing lawyers really hate: When statisticians think they know the law.)

The idea we will pursue is: What else do we know about these employees? In particular, do we know anything other than their sex? Let us assume that we know their job type and that, for simplicity, there are only two job types, denoted by A and B. We might decide to incorporate the job type into our description of the data. I will show you four possibilities for what could occur. As will be obvious, this is not an exhaustive listing of possibilities.

My first possibility is shown in Table 15.8; it shows that bringing job type into the analysis might have no effect whatsoever. The proportion in each sex/job type combination matches exactly what we had in Table 15.7.

Henceforth, we will refer to our original table as the **collapsed table** and tables such as the two in Table 15.8 as the **component tables**.

Before we move on to Case 2, let me say something about the numbers in Table 15.8. Yes, I *made up* these numbers, but I was **not** free to make them any numbers I might want; I had a major constraint. Note, for example, that 30 females were released from Job A and 30 females were released from Job B, giving us a total of 60 females released, which matches exactly the

Table 15.9: **Case 2**: Hypothetical observational data with a background factor—job type—added.

	Job A					**Job B**			
	Released?					Released?			
Sex	Yes	No	Total	$\hat{p}$	Sex	Yes	No	Total	$\hat{p}$
Female	30	10	40	0.75	Female	30	30	60	0.50
Male	30	30	60	0.50	Male	10	30	40	0.25
Total	60	40	100			40	60	100	

Table 15.10: **Case 3**: Hypothetical observational data with a background factor—job type—added.

	Job A					**Job B**			
	Released?					Released?			
Sex	Yes	No	Total	$\hat{p}$	Sex	Yes	No	Total	$\hat{p}$
Female	60	15	75	0.80	Female	0	25	25	0.00
Male	40	10	50	0.80	Male	0	50	50	0.00
Total	100	25	125			0	75	75	

number of females released by the company, as reported in Table 15.7. In fact, for **every position** in the ubiquitous 2×2 contingency table, Table 15.1, the sum of the numbers in the component tables **must equal** the number in the collapsed table, such as our $30 + 30 = 60$ for the numbers in the 'a' position. I summarize this fact by saying that the collapsed and component tables must be **consistent**. I hope that the terminology makes this easy to remember: if you combine the component tables, you get the collapsed table; if you break the collapsed table down—remember, you don't lose any data—the result is the component tables.

My next possibility is in Table 15.9. In this Case 2 we find that job type does matter and it matters in the sense that women are doing even worse than they are doing in the collapsed table—the difference, $\hat{p}_1 - \hat{p}_2$, equals 0.25 in both job types, compared to 0.20 in the collapsed table. Our next possibility, Case 3 in Table 15.10, shows that if we incorporate job type into the description, the difference between the experiences of the sexes could disappear.

Finally, Case 4 in Table 15.11 shows that if we incorporate job type into the description, the result can be quite remarkable. In the collapsed table, $\hat{p}_1 > \hat{p}_2$, but this relationship is reversed, to $\hat{p}_1 < \hat{p}_2$, in **all (both)** component tables. This reversal is called **Simpson's Paradox**.

Let's take a moment and catch our breaths. I have shown you four possible examples (Cases 1–4) of what could happen if we incorporate a background factor into an analysis. My creation of the component tables is sometimes referred to as **adjusting for a background factor** or **accounting for a background factor**. It is instructive, at this time, to look at Cases 1 and 4 in detail.

We have a response (released or not), study factor (sex) and background factor (job type). In the collapsed table we found an association between response and study factor and in Case 1 the association remained unchanged when we took into account the background factor, job type. To

Table 15.11: **Case 4: Simpson's Paradox**: Hypothetical observational data with a background factor—job type—added.

	Job A					**Job B**			
	Released?					Released?			
Sex	Yes	No	Total	$\hat{p}$	Sex	Yes	No	Total	$\hat{p}$
Female	56	24	80	0.70	Female	4	16	20	0.20
Male	16	4	20	0.80	Male	24	56	80	0.30
Total	72	28	100			28	72	100	

Table 15.12: Relationships of background factor (job type) with study factor (sex) and background factor (job type) with response (released or not) in **Case 1**.

	Job					Job			
Sex	A	B	Total	$\hat{p}$	Released?	A	B	Total	$\hat{p}$
Female	50	50	100	0.50	Yes	50	50	100	0.50
Male	50	50	100	0.50	No	50	50	100	0.50
Total	100	100	200		Total	100	100	200	

see why this is so, examine Table 15.12. We see that the background factor has no association (the row $\hat{p}$'s are identical) with either the study factor or the response. As a result, incorporating it into the analysis has no effect.

By contrast, in Case 4, putting the background factor into the analysis had a huge impact. We can see why in Table 15.13. In this case, the background factor is strongly associated with both the study factor (sex) and the response (outcome); in particular, women are disproportionately in Job A and persons in Job A are disproportionately released. It can be shown that *something like Cases 2–4* can occur only if the background factor is associated with **both the study factor and the response**. Here is where randomization becomes relevant. If subjects are assigned to study factor level by randomization, then there should be either no or only a weak association between study factor and background factor. With randomization, it would be extremely unlikely to obtain

Table 15.13: Relationships of background factor (job type) with study factor (sex) and background factor (job type) with response (released or not) in **Case 4**.

	Job					Job			
Sex	A	B	Total	$\hat{p}$	Released?	A	B	Total	$\hat{p}$
Female	80	20	100	0.80	Yes	72	28	100	0.72
Male	20	80	100	0.20	No	28	72	100	0.28
Total	100	100	200			100	100	200	

an association between sex (study factor) and job (background factor) as strong as (or stronger than) the one in Table 15.13.

Thus, with randomization, the message in the collapsed table will be pretty much the same as the message in any of the component tables; well, unless randomization yields a particularly bizarre—and unlikely—assignment.

Let's look at Simpson's Paradox and the Dating study. First, I need a candidate for my background factor. Given the opinions of my students, I will use the man's year in school. All of the computations are easier if we have only two levels for the background factor; thus, I will specify two levels. It is natural for you to wonder:

> Hey, Bob. Why don't you look at the data? Perhaps three or four levels would be better.

This is a fair point, except that **I don't have any data to look at**. Years ago, I phoned one of the researchers who had conducted the Dating study—a very talented person I had helped in 1976 when she was doing research for her M.S. degree—and, after explaining what I wanted to do, asked if she could provide me with some background data. She told me, "We already looked at that." She did kindly give me permission to include her study in a textbook I wrote. While writing the text, it was my experience that nasty researchers would not let me use their data; kindly researchers let me use their published data; no researchers provided me with additional data. Deservedly or not—my guess is deservedly—there seems to be a fear among researchers that statisticians like to criticize published work by finding other ways to analyze it. (Indeed, I am guilty of this charge, as shown in the example I present on free throws in the next subsection.)

In any event, I will look at the Dating study because it is a real study with which you are familiar, but all aspects of the component tables will be hypothetical. In addition to having hypothetical component tables, I will modify the data from the Dating study to make the arithmetic much more palatable. The modified data are presented in Table 15.14. The modified data has three changes:

1. I changed n_1 from 107 to 100.
2. I changed $\hat{p}_1$ from 0.56 to 0.55.
3. I changed $\hat{p}_2$ from 0.31 to 0.30.

I hope that you will agree that these modifications **do not** distort the basic message in the actual Dating study data, Table 15.2 on page 356.

My background factor is years on campus, with levels: *less than one* (i.e., a first year student) and *one or more*. My hypothetical component tables are given in Table 15.15. In this table, we see that if we adjust for years on campus, the two schools are **exactly the same**! Note that I am not claiming that the two schools are the same; I manufactured the data in Table 15.15. I am saying that it *could be the case* that the two schools are the same.

The moral is as follows. Whenever you see a pattern in an observational study, be aware that the following **is possible**: If one adjusts for a background factor, the pattern could become stronger; it could be unchanged; it could become weaker while preserving its direction; it could disappear; or it could be reversed. The last of these possibilities is referred to as Simpson's Paradox.

Table 15.14: Modified data from the Dating study.

	Observed Frequencies Prefer Women to:			Row Proportions Prefer Women to:		
Population	Ask	Other	Total	Ask	Other	Total
Wisconsin	55	45	100	0.55	0.45	1.00
Texas A&M	30	70	100	0.30	0.70	1.00
Total	85	115	200			

Table 15.15: Hypothetical component tables for the modified Dating study data presented in Table 15.14.

Less than 1 year on campus

	Response			
School	Ask	Other	Total	$\hat{p}$
Wisconsin	49	21	70	0.70
Texas A&M	14	6	20	0.70
Total	63	27	90	

At least 1 year on campus

	Response			
School	Ask	Other	Total	$\hat{p}$
Wisconsin	6	24	30	0.20
Texas A&M	16	64	80	0.20
Total	22	88	110	

15.5.1 Simpson's Paradox and Basketball

When I gave my presentation to the administrative judges, they seemed quite interested in the four possible cases I gave them for what could occur when one adjusts/accounts for a background factor. I told them, as I will remind you, that my Cases 1–4 **do not provide an exhaustive list of possibilities**, they are simply four examples of the possibilities.

An audience member asked me, "Which case occurred in the court case on which you were working?" I reminded him that the data are totally hypothetical.

Another audience member asked, "OK, which of the four cases is the most likely to occur in practice?" This is a trickier question. If you are really bad (nicer than saying stupid) at selecting background factors, then you will continually pick background factors that are **not associated with the response**. As a result, as I stated above, controlling for the background factor will result in Case 1 or something very similar to it. If you pick background factors that are associated with the response, then it all depends on whether the background factor is associated with the study factor. If it is, then something interesting **might** happen when you create the component tables. I will note, however, that I only occasionally see reports of Simpson's Paradox occurring. Does this mean it is rare or that people rarely look for it? In this subsection I will tell you a story of my finding Simpson's Paradox in some data on basketball free throws.

My description below is a dramatization of some published research; see [3]. It will suffice, I hope, for the goals of this course.

Researchers surveyed a collection of basketball fans, asking the question:

> A basketball player in a game is rewarded two free throw attempts. Do you agree or

disagree with the following statement?

> The player's chance of making his second free throw is greater if he made his first free throw than if he missed his first free throw.

The researchers reported that a large majority of the fans agreed with the statement. The researchers then claimed to have data that showed the fans were wrong and went so far as to say that, "The fans see patterns where none exists." My initial reaction was that the researchers' statement was rather rude. I do not call someone delusional just because we disagree! You can probably imagine my happiness when I demonstrated—well, at least in my opinion, you can judge for yourself—that the fans might have been accurately reporting—albeit misinterpreting—a pattern they had seen.

The researchers' method was, to me, quite curious. They reported data on nine players and because their analysis of the nine players showed no pattern, they concluded that the fans could not possibly have seen a pattern! As if there are only nine basketball players in the world! (The researchers in question actually have had very illustrious careers—much more than I have had, for example—and I don't entirely fault their work. Indeed, the problem is that their preliminary analysis should not have been published as it was; this was definitely a case of the peer review system failing—hardly an unusual occurrence!)

The good news is that I am not going to show you data on nine players; the data from two of the players will suffice.

Table 15.16 presents the data we will analyze. There is a great deal of information in this table and I will, therefore, spend some time explaining it. The data come from two NBA seasons combined, 1980–81 and 1981–82. Larry Bird and Rick Robey were two NBA players. Let's look at the first set of tables in Table 15.16, which presents data from Larry Bird. On 338 occasions in NBA games, Bird attempted a pair of free throws. In the next chapter, we will see how to analyze Bird's data as **paired data**, but it can also be fit into our model of independent Bernoulli trials. In particular, we will view the outcome of the second shot as the response and the outcome of the first shot as the determinant of the population. In particular, my model is that if he makes the first free throw, then the second free throw comes from the first Bernoulli trial process, with success probability equal to p_1. If he misses the first free throw, then the second free throw comes from the second Bernoulli trial process, with success probability equal to p_2. This is clearly the model the researchers had in mind; the statement they gave to the fans, in my symbols, is the statement that $p_1 > p_2$.

We see that the data for both Larry Bird and Rick Robey contraindicates the majority fan belief; i.e., $\hat{p}_1 < \hat{p}_2$ for both men. We could perform Fisher's test for either man's data, but the smallest of the six P-values (three alternatives for each man) is 0.4020—details not given, trust me on this or use the website to verify my claim. Thus, for these two men, there is only weak evidence of the first shot's outcome affecting the second shot. The researchers' conclusion—and it's valid—is that the data on either Bird or Robey, as well as their seven other players, contradicts the majority fan opinion in their survey.

If one focuses on row proportions and P-values for Fisher's tests, one might overlook the most striking feature in the data: Bird was much better than Robey at shooting free throws! I had a clever idea—no false modesty for me—and it led to a publication. I realized that we could "Do

Table 15.16: Observed frequencies and row proportions for pairs of free throws shot by Larry Bird and Rick Robey, and the collapsed table.

Larry Bird

Observed Frequencies				Row Proportions			
First Shot:	Second Shot: Hit	Miss	Total	First Shot:	Second Shot: Hit	Miss	Total
Hit	251	34	285	Hit	0.881	0.119	1.000
Miss	48	5	53	Miss	0.906	0.094	1.000
Total	299	39	338				

Rick Robey

Observed Frequencies				Row Proportions			
First Shot:	Second Shot: Hit	Miss	Total	First Shot:	Second Shot: Hit	Miss	Total
Hit	54	37	91	Hit	0.593	0.407	1.000
Miss	49	31	80	Miss	0.612	0.388	1.000
Total	103	68	171				

Collapsed Table

Observed Frequencies				Row Proportions			
First Shot:	Second Shot: Hit	Miss	Total	First Shot:	Second Shot: Hit	Miss	Total
Hit	305	71	376	Hit	0.811	0.189	1.000
Miss	97	36	133	Miss	0.729	0.271	1.000
Total	402	107	509				

Simpson's Paradox in reverse." In particular, I decided to view Bird's and Robey's tables as my component tables and I threw them together to create the collapsed table, which is the bottom table in Table 15.16.

In the collapsed table, $\hat{p}_1 = 0.811$ is much larger than $\hat{p}_2 = 0.729$. By contrast, in both component tables, $\hat{p}_1$ is about 2 percentage points smaller than $\hat{p}_2$. For these data, Simpson's Paradox is occurring! Let's take a few minutes to see *why* it is occurring. In the collapsed table, we don't know who is shooting free throws. Given that the first shot is a hit, the probability that Bird is shooting increases, while if the first shot is a miss, the probability that Robey is shooting increases. (I ask you to trust these intuitive statements; they can be made rigorous, but I don't want that diversion.) Thus, the pattern in the collapsed table is due to the difference in ability between Bird and Robey; it is not because one shot influences the next.

In my paper, I asked the following question:

> Which of the following scenarios seems more reasonable?
>
> 1. A basketball fan has in his/her brain an individual 2×2 table for **every one** of the thousands of players that he/she has ever viewed.
> 2. A basketball fan has in his/her brain one 2×2 table, collapsed over all the players that he/she has ever viewed.

I opined that only the second scenario is reasonable. My conclusions were:

- Researchers **should not assume** that the data they analyze are the data that are available to everyone else.
- For the current study, instead of scolding people for seeing patterns that don't exist, it is better to teach them that a pattern one finds in a collapsed table is not necessarily the pattern one would find in component tables.

I hope that if you become a creator of research results, you will always remember this first conclusion; I believe that it is very important.

A side note on something that turned out to be humorous, but easily could have turned out to be tragic. (Admittedly, only if you view having a paper unfairly rejected as tragic.) The referee for my paper clearly was annoyed that I had found a flaw in the researchers' conclusion because he/she said that my paper should be rejected unless and until I presented a theory on the operation of the brain that would establish that one table is easier to remember than thousands of tables. Some associate editors, no doubt, are lazy and rarely contradict their referees; I was fortunate to have an associate editor who—with knowledge of the identity of the referee—wrote to me, "I don't think you could ever make this referee happy. Your paper is interesting and I am going to publish it."

15.6 Summary

In this chapter and the next we study population-based inference for studies that yield data in the form of a 2×2 contingency table. In this chapter we focus on the so-called independent random samples type of data.

There are four types of studies considered in this chapter:

1. Observational studies to compare two finite populations;
2. Experimental studies to compare two finite populations;
3. Observational studies to compare two sequences of Bernoulli trials; and
4. Experimental studies to compare two sequences of Bernoulli trials.

First, we consider the assumptions behind the data. In the first type listed above, we assume that we have independent random samples from the two finite populations. For an experimental study on finite populations, we begin with one superpopulation and two treatments. With these ingredients, we define the first [second] finite population to be the result of giving treatment 1 [2] to every member of the superpopulation. We select a random sample of n members from the superpopulation to be the subjects in the study. The n selected members are divided into two groups by the process of randomization. The members in the first [second] group are assigned to treatment 1 [2] and, hence, they become the sample from the first [second] population. The resultant two samples are not quite independent, but if the size of the superpopulation is much larger than n, the dependence can be safely ignored.

For an observational study on Bernoulli trials, we simply observe n_1 trials from the first *process* and n_2 trials from the second *process*. As an example in these notes, the first [second] process is temperature forecasting in the spring [summer]. For an experimental study on Bernoulli trials, we begin with a plan to observe n trials, each of which can be assigned to either process of interest. The assignment is made by randomization. As an example in these notes, a basketball player's trials consisted of 100 shots. Shots were assigned to treatments (different processes, representing different locations) by randomization.

Whatever type of study we have, we define p_1 to be the proportion [probability] of success in the first population [Bernoulli trial process] and we define p_2 to be the proportion [probability] of success in the second population [Bernoulli trial process].

All four types of studies have the same formula for the confidence interval estimate of $p_1 - p_2$. It is given in Formula 15.3 and is reproduced below:

$$(\hat{p}_1 - \hat{p}_2) \pm z^* \sqrt{(\hat{p}_1\hat{q}_1)/n_1 + (\hat{p}_2\hat{q}_2)/n_2}.$$

For a test of hypotheses the null hypothesis is H_0: $p_1 = p_2$. There are three choices for the alternative:

$$H_1\text{: } p_1 > p_2;\ H_1\text{: } p_1 < p_2;\ \text{or } H_1\text{: } p_1 \neq p_2.$$

We have two methods for finding the P-value of a set of data and a specific alternative.

First, the Fisher's test website can be used to obtain an exact P-value. Second, an approximate P-value can be obtained by using the N(0,1) curve. For this second method, the observed value of the test statistic is presented in Formula 15.6 and is reproduced below:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{(\hat{p}\hat{q})(1/n_1 + 1/n_2)}}$$

In this formula,

$$\hat{p} = (x + y)/(n_1 + n_2),$$

the total number of successes in the two samples divided by the sum of the sample sizes. Also, $\hat{q} = 1 - \hat{p}$. The rules for finding the approximate P-value are given on page 359.

I introduce the placebo effect and the idea of blindness in a study.

With an observational study, it can be instructive to explore the possible impact of a background factor. We view our original data as the collapsed table which can be subdivided into component tables, with one component table for each value of the background factor. The result of creating component tables can be very interesting or a waste of time. If the pattern in the data in the collapsed table is **reversed in every component table**, then we say that Simpson's Paradox is occurring.

Table 15.17: Response to question 1 by sex for the 1986 Wisconsin survey of licensed drivers. See Practice Problem 1 for the wording of question 1 and the definition of a success.

	Counts			Row Proportions			
	Response				Response		
Sex	S	F	Total	Gender:	S	F	Total
Female	409	329	738	Female	0.554	0.446	1.000
Male	349	392	741	Male	0.471	0.529	1.000
Total	758	721	1479				

15.7 Practice Problems

1. We will analyze some data from the 1986 Wisconsin survey of licensed drivers. Question 1 on the survey asked,

   How serious a problem do you think drunk driving is in Wisconsin?

   There were four possible responses: extremely serious; very serious; somewhat serious; and not very serious. I decided to label the first of these responses—extremely serious—a success and any of the others a failure. Each subject also was asked to self-report his/her sex. A total of 1,479 subjects answered both of these questions, plus a third question, see problem 2. A total of 210 subjects failed to answer at least one of these three questions; I will ignore these 210 subjects.

   The data are presented in Table 15.17. I will view these data as independent random samples from two finite populations; the first population being female drivers and the second being male drivers. If this seems suspicious to you, I will discuss this issue in Chapter 16.

   (a) Calculate the 95% confidence interval estimate of $p_1 - p_2$. Comment.

   (b) Find the exact P-value for the alternative $\neq$. Comment.

   (c) Find the approximate P-value for the alternative $\neq$. Comment.

2. Refer to problem 1. Each of the 1,479 subjects also self-reported his/her category of consumption of alcoholic beverages, which we coded to: light drinker; moderate drinker; and heavy drinker. By the way, *light drinker* includes self-reported non-drinkers. I used these three coded responses as my three levels of a background factor. The resultant three component tables are presented in Table 15.18.

   (a) If you fix the population—either female or male—what happens to $\hat{p}$ as you move from one component table to another?

   (b) If you create 2×3 table for which rows are sex and columns are self-reported drinking behavior, what will you find? Note that you don't need to create this table; you should be able to see the pattern from Table 15.18.

Table 15.18: Three component tables for the data in Table 15.17.

Light Drinkers

Observed Frequencies				Row Proportions			
	Response				Response		
Gender:	S	F	Total	Gender:	S	F	Total
Female	234	133	367	Female	0.638	0.362	1.000
Male	151	100	251	Male	0.602	0.398	1.000
Total	385	233	618				

Moderate Drinkers

Observed Frequencies				Row Proportions			
	Response				Response		
Gender:	S	F	Total	Gender:	S	F	Total
Female	133	142	275	Female	0.484	0.516	1.000
Male	115	146	261	Male	0.441	0.559	1.000
Total	248	288	536				

Heavy Drinkers

Observed Frequencies				Row Proportions			
	Response				Response		
Gender:	S	F	Total	Gender:	S	F	Total
Female	42	54	96	Female	0.438	0.562	1.000
Male	83	146	229	Male	0.362	0.638	1.000
Total	125	200	325				

(c) In the collapsed table, $\hat{p}_1 - \hat{p}_2 = 0.083$. Calculate this difference for each of the component tables. Comment.

(d) In each of the component tables, find the exact P-value for the alternative $\neq$; thus, I am asking you to find three P-values. Comment.

(e) Write a brief summary of what you have learned in these first two problems.

3. In Example 12.1, I introduced you to my friend Bert's playing mahjong solitaire online. I will use Bert's data to investigate whether he had Bernoulli trials.

   Let's model his first [second] 50 games as Bernoulli trials with success rate p_1 [p_2]. I want to test the null hypothesis that $p_1 = p_2$; if this hypothesis is true, then he has the same p for all 100 games. Thus, in the context of assuming independence, this is a test of whether Bert has Bernoulli trials.

   Recall that I reported that Bert won 16 of his first 50 games and 13 of his second 50 games. Find the exact P-value for the alternative $\neq$.

4. An observational study yields the following *collapsed table.*

Group	S	F	Total
1	99	231	330
2	86	244	330
Total	185	475	660

   Below are two component tables for these data.

Subgroup A				Subgroup B			
Group	S	F	Total	Group	S	F	Total
1	24	96	120	1	75	135	210
2	c_A		225	2	c_B		105
Total			345	Total			315

   Complete these tables so that Simpson's Paradox is occurring **or** explain why Simpson's Paradox *cannot* occur for these data. You must present computations to justify your answer.

5. An observational study yields the following *collapsed table.*

Group	S	F	Total
1	180	320	500
2	117	183	300
Total	297	503	800

   Below are two component tables for these data.

Subgroup A				Subgroup B			
Group	S	F	Tot	Group	S	F	Tot
1	130	130	260	1	50	190	240
2	c_A		180	2	c_B		120
Total			440	Total			360

Complete these tables so that Simpson's Paradox is occurring **or** explain why Simpson's Paradox *cannot* occur for these data. You must present computations to justify your answer.

15.8 Solutions to Practice Problems

1. (a) We will use Formula 15.3 with $z^* = 1.96$. The values of the two $\hat{p}$'s, the two $\hat{q}$'s, n_1 and n_2 are given in Table 15.17. Substituting these values into Formula 15.3 gives:

$$(0.554 - 0.471) \pm 1.96\sqrt{(0.554)(0.446)/738 + (0.471)(0.529)/741} =$$

$$0.083 \pm 0.051 = [0.032, 0.134].$$

The interval does not include zero; I conclude, first, that p_1 is larger than p_2. More precisely, I conclude that the proportion of female drivers who are successes is between 3.2 and 13.4 percentage points larger than the proportion of male drivers who are successes.

(b) Using the Fisher's test website, I find that the exact P-value for the alternative $\neq$ is 0.0015. The data are highly statistically significant. There is very strong evidence that the female proportion is larger than the male proportion.

(c) For the approximate test, I need to calculate

$$\hat{p} = (409 + 349)/(738 + 741) = 0.513.$$

The observed value of the test statistic is given in Formula 15.6. Substituting into this formula, we obtain:

$$z = \frac{0.083}{\sqrt{0.513(0.487)(1/738 + 1/741)}} = \frac{0.083}{0.0260} = 3.192.$$

From a Normal curve area website calculator, I find that the area under the N(0,1) curve to the right of 3.192 equals 0.0007. Doubling this, we get 0.0014 as the approximate P-value. The approximate and exact P-values are almost identical.

2. (a) For both females and males, as the degree of drinking increases, the proportion of successes declines.

(b) It is clear that men have substantially higher levels of drinking alcohol than women.

(c) By subtraction, the difference is 0.036 for the light drinkers; 0.043 for the moderate drinkers; and 0.076 for the heavy drinkers. The difference is positive in each component table; thus, we are not close to having Simpson's Paradox. The difference, however, is closer to zero in each component table than it is in the collapsed table.

(d) For the light drinkers, the P-value is 0.3982; for the moderate drinkers, the P-value is 0.3409; and for the heavy drinkers, the P-value is 0.2136. None of these P-values approaches statistical significance.

(e) For each category of drinking, women have a higher success rate than men, but none of the data approaches statistical significance. (Aside: The direction of the effect is consistent—women always *better* than men—and there is a way to combine the tests, but we won't cover it in this class. When I did this, I obtained an overall approximate P-value equal to 0.0784, which is much larger than the P-value for the collapsed table.)

In the collapsed table, women have a much higher success rate than men. I would **not** label the collapsed table analysis as **wrong**. If one simply wants to compare men and women, then its analysis is valid. Many people, however, myself included, think it's important to discover that self-reported drinking frequency is linked to attitude, which we can see with the component tables.

3. Go to the Fisher's test website and enter the counts 16, 34, 13 and 37. The resultant two-sided P-value is 0.6598. The evidence in support of a changing value of p is very weak.

4. In the collapsed table the two row totals are equal. Thus, it is easy to see that $\hat{p}_1 > \hat{p}_2$. So, we need a reversal, $\hat{p}_1 < \hat{p}_2$, in both component tables.

   In the A table, this means

   $$c_A/225 > 24/120 \text{ or } c_A > 45 \text{ or } c_A \geq 46.$$

   In the B table, this means

   $$c_B/105 > 75/210 \text{ or } c_B > 37.5 \text{ or } c_B \geq 38.$$

   Consistency requires that $c_A + c_B = 86$. There are three ways to satisfy these three conditions: $c_A = 46$ and $c_B = 40$; $c_A = 47$ and $c_B = 39$; and $c_A = 48$ and $c_B = 38$.

5. In the collapsed table

   $$\hat{p}_1 = 180/500 = 0.36; \hat{p}_2 = 117/300 = 0.39.$$

   Thus, for a reversal we need $\hat{p}_1 > \hat{p}_2$ in both component tables.

   In the A table, this means

   $$c_A/180 < 130/260 \text{ or } c_A < 90 \text{ or } c_A \leq 89.$$

   In the B table, this means

   $$c_B/120 < 50/240 \text{ or } c_B < 25 \text{ or } c_B \leq 24.$$

Consistency requires that $c_A + c_B = 117$. It is impossible to satisfy these three conditions because $c_A \leq 89$ and $c_B \leq 24$ imply that

$$c_A + c_B \leq 89 + 24 = 113 < 117.$$

15.9 Homework Problems

1. Refer to the data in Table 15.18. Consider a data table with the same response, female moderate drinkers in row 1 and male moderate drinkers in row 2.

   (a) Calculate the 90% confidence interval estimate of $p_1 - p_2$.

   (b) Find the exact P-value for the alternative $>$.

   (c) Find the approximate P-value for the alternative $>$.

2. Refer to the data in Table 15.18. Consider a data table with the same response, male moderate drinkers in row 1 and male heavy drinkers in row 2.

   (a) Calculate the 90% confidence interval estimate of $p_1 - p_2$.

   (b) Find the exact P-value for the alternative $>$.

   (c) Find the approximate P-value for the alternative $>$.

3. An observational study yields the following *collapsed table*.

Group	S	F	Total
1	113	287	400
2	102	198	300
Total	215	485	700

   Below are two component tables for these data.

	Subgroup A				Subgroup B		
Gp	S	F	Total	Gp	S	F	Total
1	73	227	300	1	40	60	100
2	c_A		100	2	c_B		200
Total			400	Total			300

   Determine all pairs of values of c_A and c_B so that Simpson's Paradox is occurring **or** explain why Simpson's Paradox *cannot* occur for these data. You must present computations to justify your answer.

4. An observational study yields the following *collapsed table*.

Group	S	F	Total
1	60	40	100
2	c		100
Total			200

Below are two component tables for these data.

Subgroup A				Subgroup B			
Group	S	F	Total	Group	S	F	Total
1	25	25	50	1	35	15	50
2	36	34	70	2	c_B		30
Total	61	59	120	Total			80

Determine all pairs of values of c and c_B so that Simpson's Paradox is occurring **or** explain why Simpson's Paradox *cannot* occur for these data. You must present computations to justify your answer.

Bibliography

[1] Muehlenhard, C. and Miller, E.: 'Traditional and Nontraditional Men's Responses to Women's Dating Initiation," *Behavior Modification*, July, 1988, pp 385–403.

[2] Brynskov, J., et.al., "A Placebo Controlled, Double-Blind, Randomized Trial of Cyclosporine Therapy in Active Chronic Crohn's Disease", *The New England Journal of Medicine*, September 28, 1989, pp 845–850.

[3] Tversky A. and Gilovich T, "The Cold Facts about the Hot Hand in Basketball," *Chance: New Directions in Statistics and Computing*, Winter, 1989, pp. 16-21.

Chapter 16

One Population with Two Dichotomous Responses

This chapter focuses on a new idea. Thus far in these notes, a unit (subject, trial) has yielded one response. In this chapter, we consider situations in which each unit yields **two responses**, both dichotomies. Later in these *Course Notes* we will examine situations in which both responses are numbers and the mixed situation of one response being a number and the other a dichotomy. Multi-category responses can be added to the mix, but—with one exception—we won't have time for that topic.

Sometimes the examples of this chapter will look very much like our examples of Chapter 15. Other times, it will be natural to view our two responses as **paired data**. As a result, you need to be extra careful as you read through this material.

16.1 Populations: Structure, Notation and Results

A population model for two dichotomous responses can arise for a collection of individuals—a finite population—or as a mathematical model for a process that generates two dichotomous responses per trial.

Here are two examples.

1. Consider the population of students at a small college. The two responses are sex with possible values female and male; and the answer to the following question, with possible values yes and no.

   Do you usually wear corrective lenses when you attend lectures?

2. Recall the data on Larry Bird in Chapter 15, presented in Table 15.16 on page 374. We view his shooting a pair of free throws as a trial with two responses: the outcome of the first shot and the outcome of the second shot.

Recall that I treated the Larry Bird data as *Chapter 15 data*; i.e., independent random samples from two Bernoulli trials processes. Later in this chapter we will view his results as paired data. Both

perspectives are valid, but it will require some care for you to be comfortable with such *moving between models*. Also, my example of sex and lenses can be viewed as Chapter 15 data, but I would find it awkward to refer to it as paired data.

We begin with some notation. With two responses per unit, sometimes it would be confusing to speak of successes and failures. Instead, we proceed as follows.

- The first response has possible values A and A^c. Note that A^c is read *A-complement* or *not-A*.
- The second response has possible values B and B^c. Note that B^c is read *B-complement* or *not-B*.

In the above example of a finite population, A could denote female; A^c could denote male; B could denote the answer 'yes;' and B^c could denote the answer 'no.' In the above example of trials, A could denote that the first shot is made; A^c could denote that the first shot is missed; B could denote that the second shot is made; and B^c could denote that the second shot is missed. In fact, with data naturally viewed as paired, such as Larry Bird's shots, it is natural to view A [B] as *a success on the first [second] response* and A^c [B^c] as *a failure on the first [second] response*.

It will be easier if we consider finite populations and trials separately. We will begin with finite populations.

16.1.1 Finite Populations

Table 16.1 presents our notation for population counts for a finite population. Remember that, in practice, only Nature would know these numbers. This notation is fairly simple to remember: all counts are represented by N, with or without subscripts. The symbol N without subscripts represents the total number of members of the population. An N with subscripts counts the number of members of the population with the feature(s) given by the subscripts. For example, N_{AB} is the number of population members with response values A and B; N_{A^c} is the number of population members with value A^c on the first response; i.e., for this, we don't care about the second response.

Note also that these guys sum in the obvious way:

$$N_A = N_{AB} + N_{AB^c}.$$

In words, if you take the number of population members whose response values are A and B; and add to it the number of population members whose response values are A and B^c, then you get the number of population members whose value on the first response is A.

It might help if we have some hypothetical values for the population counts. I put some in Table 16.2.

If we take the table of population counts and divide each entry by N, we get the table of population proportions or probabilities—see the discussion in the next paragraph. I do this in Tables 16.3 and 16.4, for the general notation and our particular hypothetical data.

Now we must face a notational annoyance. Consider the number 0.36 in Table 16.4, derived from our hypothetical population counts for the sex and lenses study. There are two ways to interpret this number. First, it is the *proportion* of the population who have value A (female) on the

Table 16.1: The table of population counts.

	B	B^c	Total
A	N_{AB}	N_{AB^c}	N_A
A^c	N_{A^cB}	$N_{A^cB^c}$	N_{A^c}
Total	N_B	N_{B^c}	N

Table 16.2: Hypothetical population counts for the study of sex and corrective lenses.

	Yes (B)	No (B^c)	Total
Female (A)	360	240	600
Male (A^c)	140	260	400
Total	500	500	1000

Table 16.3: The table of population proportions—lower case p's with subscripts—or probabilities—upper case P's followed by parentheses.

	B	B^c	Total
A	$p_{AB} = P(AB)$	$p_{AB^c} = P(AB^c)$	$p_A = P(A)$
A^c	$p_{A^cB} = P(A^cB)$	$p_{A^cB^c} = P(A^cB^c)$	$p_{A^c} = P(A^c)$
Total	$p_B = P(B)$	$p_{B^c} = P(B^c)$	1

Table 16.4: Hypothetical population proportions or probabilities for the study of sex and corrective lenses.

	Yes (B)	No (B^c)	Total
Female (A)	0.36	0.24	0.60
Male (A^c)	0.14	0.26	0.40
Total	0.50	0.50	1.00

first response **and** value B (yes) on the second response. From this perspective, it is natural to view 0.36 as p_{AB} because we use lower case p's for population proportions—with a subscript, if needed, to clarify which one. But consider our most commonly used chance mechanism when studying a finite population: Select a member of the population at random. For this chance mechanism it is natural to view 0.36 as the *probability* of selecting a person who is female and would answer 'yes.' We use upper case 'P' to denote the word probability. Hence, it is also natural to write $P(AB) = 0.36$.

The point of all this is ...? Well, in this chapter $p_{AB} = P(AB)$ (and $p_A = P(A)$, and so on); the one we use will depend on whether we feel it is more natural to talk about proportions or probabilities.

16.1.2 Conditional Probability

Conditional probability allows us to investigate one of the most basic questions in science: How do we make use of partial information?

Consider again the hypothetical population presented in Tables 16.2 and 16.4. Consider the chance mechanism of selecting one person at random from this population. We see that $P(A) = 0.60$. In words, the probability is 60% that we will select a female. But suppose we are given the partial information that the person selected answered 'yes' to the question. *Given* this information, what is the probability the person selected is a female? We write this symbolically as $P(A|B)$; i.e., the probability that A will occur given that B occurs. How do we compute it?

We reason as follows. Given that B occurs, we know that the selected person is among the 500 in column B of Table 16.2. Of these 500 persons, reading up the column, we see that 360 are female. Thus, by direct reasoning $P(A|B) = 360/500 = 0.72$, which is different than $P(A) = 0.60$. In words, knowledge that the person usually wears corrective lenses in lecture *increases* the probability that the person is female.

We now repeat the above reasoning, but using symbols instead of numbers. Refer to Table 16.1. Given that B occurs, we know that the selected subject is among the N_B subjects in column B. Of these N_B subjects, reading up the column, we see that N_{AB} have property A. Thus, by direct reasoning we obtain the following equation.

$$P(A|B) = N_{AB}/N_B. \tag{16.1}$$

Now, this is a perfectly good equation, relating the conditional probability of A given B to population counts. Most statisticians, however, prefer a modification of this equation. On the right side of the equation divide both the numerator and denominator by N. This, of course, does not change the value of the right side and has the effect of converting counts to probabilities. The result is below, the equation which is usually referred to as the definition of conditional probability.

$$P(A|B) = P(AB)/P(B). \tag{16.2}$$

Now there is nothing uniquely special about our wanting to find $P(A|B)$; we could just as well be interested in, say, $P(B^c|A^c)$. In fact, there are eight possible conditional probabilities of interest; all combinations of the following three dichotomous choices: to use A or A^c; to use B or

Table 16.5: Conditional probabilities of the B's given the A's in the hypothetical study of sex and lenses. For example, $P(B|A) = 0.60$, $P(B^c|A^c) = 0.65$ and $P(B^c) = 0.50$.

	Yes (B)	No (B^c)	Total
Female (A)	0.60	0.40	1.00
Male (A^c)	0.35	0.65	1.00
Unconditional	0.50	0.50	1.00

B^c; and to put the 'A' or the 'B' first. Now, of course, it would be no fun to derive these eight formulas one-by-one; fortunately, if we view Equation 16.2 creatively, we don't need to.

Look at Equation 16.2 again. What it is really saying is,

> If you want the conditional probability of one event given another event, calculate the probability that both events occur divided by the probability of the conditioning event occurring.

With this interpretation, we immediately know how to compute any conditional probability. For example,

$$P(B^c|A^c) = P(A^cB^c)/P(A^c).$$

Note that in the numerator on the right side of this equation, we write our event in 'alphabetical order;' i.e. the 'A' event is written before the 'B' event. We are not *required* to do this, but life is easier if we adopt little conventions like this one: for example, it is easier to see whether different people have obtained the same answer.

I will now return to the study of sex and lenses to show a quick way to obtain all eight conditional probabilities. We can work with either the table of population counts or population proportions; I will use the latter, Table 16.4.

Divide every entry in Table 16.4 by its row total. The results are in Table 16.5. The four numbers in the body of this table are the conditional probabilities of the B's given the A's. For example, $P(B^c|A) = 0.40$. In words, given the selected person is female, the probability is 40% that the person will answer 'no.' More succinctly, 40% of the females would answer 'no.'

Here is a hint to help you remember that Table 16.5 gives the conditional probabilities of the B's given the A's, instead of the A's given the B's. Look at the margins: We see 0.50 as the (marginal, unconditional) probability of both B and B^c. The other marginal totals are both 1.00 and they cannot be the probabilities of A and A^c. Thus, this is a table of probabilities of B's; i.e., probabilities of B's given A's.

Similarly, if you divide every entry in Table 16.4 by its column total you get the table of the conditional probabilities of the A's given the B's. The results are in Table 16.6.

16.1.3 How Many Probabilities are There?

Look at Table 16.3 again. There are nine probabilities in this table: the four cell probabilities, the four marginal probabilities and the overall probability of 1 in the lower right corner. All except

Table 16.6: Conditional probabilities of the A's given the B's in the hypothetical study of sex and lenses. For example, $P(A|B) = 0.72$, $P(A^c|B^c) = 0.52$ and $P(A^c) = 0.40$.

	Yes (B)	No (B^c)	Unconditional
Female (A)	0.72	0.48	0.60
Male (A^c)	0.28	0.52	0.40
Total	1.00	1.00	1.00

the 1 are unknown to a researcher, but this does not mean that there are actually eight unknown probabilities. It turns out that if one chooses wisely (remember the very old knight near the end of *Indiana Jones and the Last Crusade*) then knowledge of three probabilities will suffice to determine all eight probabilities. As we shall see below, there are several possible *sets of three*, although I won't give you an exhaustive list of sets. I will give you four of the possible sets that are of the greatest interest to scientists. The first two of my four sets obviously work; the other two require some care and will be covered in a Practice Problem. I will be referring to the symbols in Table 16.3.

1. **Any three of the cell probabilities will suffice.** For example, if we know the values of $P(AB)$, $P(AB^c)$ and $P(A^cB)$, then we can obtain the remaining five unknown probabilities. For example, in Table 16.4, once we know $P(AB) = 0.36$, $P(AB^c) = 0.24$ and $P(A^cB) = 0.14$, we can obtain the remaining probabilities by addition and subtraction.

2. **A row marginal probability, a column marginal probability and any cell probability will suffice.** For example, if we know the values of $P(A)$, $P(B)$ and $P(AB)$, then we can determine all eight probabilities. For example, in Table 16.4, once we know $P(A) = 0.60$, $P(B) = 0.50$ and $P(AB) = 0.36$, we can obtain the remaining probabilities by subtraction and addition.

3. **A conditional probability for each row plus one of the row marginal probabilities will suffice.** For example, if we know the values of $P(B|A)$, $P(B|A^c)$ and $P(A)$, then we can determine all eight probabilities.

4. **A conditional probability for each column plus one of the column marginal probabilities will suffice.** For example, if we know the values of $P(A|B)$, $P(A|B^c)$ and $P(B)$, then we can determine all eight probabilities.

If you choose your three probabilities *unwisely* you will not be able to determine all probabilities. For one of many possible examples, suppose you know $P(A) = 0.80$, $P(B) = 0.30$ and $P(A^c) = 0.20$. With these three probabilities, you cannot determine the remaining five probabilities. (Try it!) The difficulty lies in the fact that once we know $P(A) = 0.80$, we can deduce that $P(A^c) = 0.20$; i.e., knowing $P(A^c)$ is not *new* information.

16.1.4 Screening Test for a Disease

You might be thinking, "I am not interested in any relationship between sex and lenses." Fair enough. I used that example just to get us going. In this subsection we will consider an extremely important application of the ideas of this chapter, namely the analysis of a screening test for a disease.

For many diseases, early detection can be extremely beneficial, for both the ultimate outcome for the patient and the cost of treatment. Screening tests, however, are often controversial. At my annual physical a few years ago, I learned that the PSA screening test for prostate cancer was no longer routine. More recently, there has been much discussion in the media about new recommendations on mammograms for the detection of breast cancer in women.

Here is the way our model fits. We have a population of people at a particular moment in time and we are interested in one particular disease. Each person either has the disease, denoted by A, or does not have the disease, denoted by A^c. Furthermore, we can imagine giving each person a screening test. For any given person, the screening test can be positive, denoted by B, or negative, denoted by B^c. Thus, for each person in the population there are two dichotomous responses of interest: the actual disease state and the results of the screening test, if it were given.

We can see immediately that the current problem has issues that were not present in our hypothetical study of sex and lenses. First, it might not be easy to learn whether a person has a disease. (If it were easy, inexpensive and painless to determine, nobody would bother with trying to develop a screening test.) Second, we cannot *force* a person to have a screening test. (True, we cannot force a person to tell us his/her sex or whether he/she usually wears corrective lenses, but the issue is much trickier for screening tests that might be painful and might have negative consequences.)

As a result, one must use great care in any attempt to evaluate a real life screening test. What I present below is an idealization of what a researcher or physician will face in practice.

Remember that a positive result on a screening test is interpreted as indicating the disease is present. **But it is very important to remember that screening tests make mistakes!**

In a screening test, the various combinations of response values have special meanings and it is important to note these. In particular,

- Event AB is called a correct positive.
- Event A^cB is called a false positive.
- Event AB^c is called a false negative.
- Event A^cB^c is called a correct negative.

Make sure you understand *why* these labels make sense.

Let's now look at a hypothetical screening test, presented in Table 16.7. Let's summarize the information in this table.

1. Ten percent of the population have the disease.
2. If everyone were given the screening test, 18.5% of the population would test positive.

Table 16.7: Population counts for a hypothetical screening test.

	Screening test result: Positive (B)	Negative (B^c)	Total
Disease Present (A)	95	5	100
Disease Absent (A^c)	90	810	900
Total	185	815	1,000

3. The screening test is correct for 905 persons: 95 correct positives and 810 correct negatives.

4. The screening test is incorrect for 95 persons: 90 false positives and 5 false negatives.

Let's focus on the errors made by the screening test. (I am definitely a *glass half empty* person!)

Consider false negatives; i.e., the combination of A, the disease present, with B^c, the screening test saying the disease is absent. It seems obvious to me that there are three rates, or probabilities, of possible interest. I will illustrate these ideas with our hypothetical screening test.

1. $P(AB^c) = 5/1000 = 0.005$; in words, one-half of one percent of the population will receive a false negative test result.

2. $P(A|B^c) = 5/815 = 0.006$; in words, six-tenths of one percent of those who would test negative will actually have the disease.

3. $P(B^c|A) = 5/100 = 0.05$; in words, five percent of those who have the disease would test negative.

It seems to me that each of these three numbers—0.005, 0.006 and 0.05 in our example—could reasonably be called *a false negative rate*. Biostatisticians call $P(B^c|A)$ **the false negative rate** and, as best I can tell, have not given a name to the other rates.

Does any of this matter? Well, yes, I think it does. First, I would be hard pressed to argue that $P(AB^c)$ is an interesting number, but I believe that both of the conditional probabilities are worthy of attention. I believe it is useful to think of $P(B^c|A)$ as being of most interest to the medical community and $P(A|B^c)$ as being of most interest to regular people. Why do I say this?

First, consider $P(B^c|A)$, the one that is called the false negative rate by the medical community. As a physician, I might think,

> Let's consider all the persons with the disease; what proportion of these *people who need help* will be told that they are fine?

Second, consider $P(A|B^c)$, the one with no name. A person is told that the screening test result is negative. A thoughtful person—some would say *hypochondriac*, but let's not be judgmental—might wonder,

> Do I have the disease; i.e., did the screening test make a mistake?

A definitive answer might be available **only with an autopsy**—thanks, but I'll pass on that! There is some value, however, in considering the number $P(A|B^c)$. Of all the persons who test negative, this number is the proportion that actually have the disease.

As an aside, I hope that you have **not** concluded that I view the medical research community as some nasty organization that gives names only to things of interest to them and refuses to name things of interest to patients. As we will see later in this chapter, data collection might well give us a good estimate of $P(B^c|A)$, but it is often impossible to estimate—without some controversial assumptions—the value of $P(A|B^c)$. My suspicion is that $P(B^c|A)$ is given a name, in part, because it can be estimated without controversy. But, of course, I could be wrong.

My issue of there being two rates of interest becomes really important when we consider false positives. A false positive occurs whenever A^c is matched with B. Again, there are three rates we could calculate, again illustrated with our hypothetical screening test.

1. $P(A^cB) = 90/1000 = 0.09$; in words, nine percent of the population will receive a false positive test result.
2. $P(B|A^c) = 90/900 = 0.10$; in words, ten percent of those for whom the disease is absent would test positive.
3. $P(A^c|B) = 90/185 = 0.487$; in words, 48.7% of those who would test positive are free of the disease.

Look at these last two rates. Biostatisticians call $P(B|A^c)$ **the false positive rate**, but it seems to me that $P(A^c|B)$ deserves a name too! At various times in history, governments have considered or enacted policies in which: everybody gets tested and those who test positive will be quarantined. This last computation shows that, for our hypothetical population, of those quarantined, 48.7% are actually disease free!

16.1.5 Trials and Bayes' Formula

We do not have a table of population counts for trials. We begin with the table of probabilities.

So, how do we obtain the table of probabilities? Well, mostly, we cannot obtain it, but in later sections we will learn how to estimate it from data. But this is a good time to introduce a couple of important ideas.

Recall the definition of conditional probability given in Equation 16.2:

$$P(A|B) = P(AB)/P(B).$$

We can rewrite this as

$$P(AB) = P(B)P(A|B). \tag{16.3}$$

This new equation is called **the multiplication rule for conditional probabilities**. Note what it says, in words:

> The probability that two events both occur is the product of the probability of one event occurring and the conditional probability of the remaining event, given the one we already handled.

With this admittedly awkward statement, we can obtain two *versions* of the multiplication rule for conditional probabilities. For example, if we interchange the roles of A and B in Equation 16.3, we get:

$$P(AB) = P(A)P(B|A). \tag{16.4}$$

We want two ways to calculate $P(AB)$ because sometimes we know the pair $P(A)$ and $P(B|A)$ and other times we know the pair $P(B)$ and $P(A|B)$.

I will now give you a somewhat silly application of these ideas to illustrate how we can *build a table of probabilities*.

Years ago, I lived with a dog named Casey. Periodically, I would exercise on my treadmill for 30 minutes and watch television. Casey would sit by the window watching the yard, with special interest in viewing squirrels. I could not see the yard from the treadmill.

I will view each such 30 minute segment of time as a trial. Two dichotomous responses are obtained—though not by me—during the trial:

- A: One or more squirrels enter the yard; obviously, A^c is that no squirrels enter the yard.
- B: Casey barks at some time during the trial; obviously, B^c is that Casey does not bark during the trial.

From past experience I know (estimate or guess might be more accurate verbs; see the next section) the following numbers:

$$P(A) = 0.30, P(B|A) = 0.80 \text{ and } P(B|A^c) = 0.10.$$

In words, in any given trial, there is a 30% chance that at least one squirrel will visit the yard; given that at least one squirrel visits the yard, there is an 80% chance that Casey will bark; and if no squirrels visit the yard, there is a 10% chance that Casey will bark.

The first fact we can deduce is: $P(A^c) = 1 - P(A) = 0.70$. We now proceed to complete the following table.

	B	B^c	Total
A			0.30
A^c			0.70
Total			1.00

Next, we consider $P(AB)$. By Equation 16.3, $P(AB) = P(B)P(A|B)$, but this is no help, because we know neither of the numbers on the right side of this equation. Instead, we use Equation 16.4,

$$P(AB) = P(A)P(B|A) = 0.30(0.80) = 0.24. \text{ Similarly,}$$

$$P(A^cB) = P(A^c)P(B|A^c) = 0.70(0.10) = 0.07.$$

Next, we place these two numbers in our table and continue with simple subtractions and additions until we obtain the completed table, given below.

	B	B^c	Total
A	0.24	0.06	0.30
A^c	0.07	0.63	0.70
Total	0.31	0.69	1.00

There are some amazing facts revealed in this table! First, we see that $P(B) = 0.31$; i.e., there is a 31% chance that Casey will bark during a trial. Why is this amazing? Well, we started with information on whether Casey would bark conditional on squirrel behavior, and end up with the unconditional probability of Casey barking.

OK, well maybe the former was not so amazing, but this next one definitely is. It is so great that it has a name: **Bayes' formula** or **Bayes' rule**; well, two names. It is named in honor of the Reverend Thomas Bayes who did or did not discover it before his death in 1761. (The historical controversy will be ignored in this course.)

First, I will show you how to use Bayes' rule, which is very easy, and then I will give you the formula, which is quite a mess. Suppose that during my trial I hear Casey bark. It is natural for me to wonder, "Did any squirrels visit the yard?" In other words, I want to calculate $P(A|B)$. Now before Bayes nobody could answer this question; in fact, it looked impossible: we are given information about conditional probabilities of B's given A's, how could we possibly reverse them? It seemed like alchemy or some occult method would be required to obtain an answer.

But as often happens, especially in math or riding a bicycle, what appears to be illogical or impossible works if you just do it. (Ugh! This sounds like a Nike commercial; what's next? Impossible is nothing?)

Let's just calculate $P(A|B)$; but how? Well, by definition, $P(A|B) = P(AB)/P(B)$ and we can read both of these numbers from our table! Thus,

$$P(A|B) = P(AB)/P(B) = 0.24/0.31 = 0.774.$$

In words, given that Casey barks, there is a 77.4% chance that at least one squirrel visited the yard.

We see that it is easy to reverse the direction of conditioning, provided we are able to complete the table of probabilities. My advice is that in practice, just complete the table and then you can calculate any conditional probability that you desire.

For completeness, I will show you Bayes' formula in all its mathematical glory, but I do not recommend using it and you are not responsible for it in any way!

Here are elements we need for Bayes' formula.

- We need a **partition** of the sample space, denoted by the k events $A_1, A_2, \ldots, A_k$. By a partition I mean that the events are pairwise disjoint (i.e., they don't overlap in any way) and their union is the entire sample space.
- We need some other event of interest B.
- We need to know $P(A_i)$, for $i = 1, 2, \ldots, k$.
- We need to know $P(B|A_i)$, for $i = 1, 2, \ldots, k$.

Table 16.8: The table of counts for a sample of size n.

	B	B^c	Total
A	a	b	n_1
A^c	c	d	n_2
Total	m_1	m_2	n

Before I give you Bayes' formula note that these three conditions are met for my example with Casey: $k = 2$; the events are $A_1 = A$ and $A_2 = A^c$; B is as above; and both probabilities and both conditional probabilities are known.

So, here it is, **Bayes' formula**:

$$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{\sum_{j=1}^{k} P(A_j)P(B|A_j)}, \text{ for } i = 1, 2, \ldots, k. \tag{16.5}$$

16.2 Random Samples from a Finite Population

Recall our table of probabilities, reproduced below:

	B	B^c	Total
A	$P(AB)$	$P(AB^c)$	$P(A)$
A^c	$P(A^cB)$	$P(A^cB^c)$	$P(A^c)$
Total	$P(B)$	$P(B^c)$	1

In my experience, in most scientific applications these eight probabilities are unknown. Also, a scientist might well be interested in any or all of the eight possible conditional probabilities. (Admittedly, as discussed earlier, there are many mathematical relationships between these 16 unknown rates.) In this section we will discuss what is possible and what is practicable in estimating these various numbers.

We will investigate three types of random samples. Whichever way we sample, we will present the data we obtain in a table, as illustrated in Table 16.8, our ubiquitous 2×2 table for data, introduced in Chapter 15 in Table 15.1 on page 354. I introduce the three types of random samples below and I will illustrate each with our earlier example of sex and lenses.

1. **Type 1: Overall Random Sample.** Select a random sample of size n from the population. In this case all eight of the remaining numbers in Table 16.8 (excluding the n) are the observed values of random variables; i.e., their values cannot be predicted, with certainty, before data collection.

2. **Type 2: Independent Random Samples from the Rows.** Select two independent random samples: The first is of size n_1 from the population of all subjects with value A for the first response; the second is of size n_2 from the population of all subjects with value A^c for the first response. In this case, n_1, n_2 and $n = n_1 + n_2$ are all fixed in advance by the researcher; the remaining six counts are observed values of random variables.

3. **Type 3: Independent Random Samples from the Columns.** Select two independent random samples: The first is of size m_1 from the population of all subjects with value B for the second response; the second is of size m_2 from the population of all subjects with value B^c for the second response. In this case, m_1, m_2 and $n = m_1 + m_2$ are all fixed in advance by the researcher; the remaining six counts are observed values of random variables.

For our sex and lenses study, these become:

1. **Type 1: Overall Random Sample.** A random sample of size n is selected from the population of all 1,000 students.

2. **Type 2: Independent Random Samples from the Rows.** Two lists are created: one of the 600 female students and one of the 400 male students. Select a random sample of size n_1 from the female population. Then select an independent random sample of size n_2 from the male population.

3. **Type 3: Independent Random Samples from the Columns.** Two lists are created: one of the 500 students who would answer 'yes' and one of the 500 students who would answer 'no.' Select a random sample of size m_1 from the 'yes' population. Then select an independent random sample of size m_2 from the 'no' population.

Note that it is often the case that at least one of the three ways of sampling is unrealistic. In the above, I cannot imagine that the researcher would have a list of either the 'yes' or 'no' population; hence, the Type 3 sampling is not of interest for this example.

Let us consider Type 2 sampling in general. Upon reflection, you will realize that Type 2 sampling is equivalent to what we studied in Chapter 15, except that the names have been changed. In particular, population 1 in Chapter 15 consists of all subjects with feature A and population 2 in Chapter 15 consists of all subjects with feature A^c. In this context, label B a success and B^c a failure. Thus, what we earlier called p_1 and p_2 are now $P(B|A)$ and $P(B|A^c)$, respectively. Thus, our earlier methods for estimation and testing of $p_1 - p_2$ can be immediately applied to a difference of conditional probabilities: $P(B|A) - P(B|A^c)$.

Of course, Type 2 and Type 3 sampling are really the same mathematically. For example, if you have independent random samples from the columns, you may simply interchange the roles of rows and columns and then have independent random samples from the rows. In practice it is convenient to allow for both Type 2 and Type 3 sampling.

For independent random samples from the columns, identify population 1^* as all subjects with feature B and population 2^* as all subjects with feature B^c. Make the definitions $p_1^* = P(A|B)$ and $p_2^* = P(A|B^c)$. Thus, the difference in conditional probabilities $P(A|B) - P(A|B^c)$ is simply $p_1^* - p_2^*$.

The above might seem very confusing, but it's really quite simple: Sample by rows and we are interested in $p_1 - p_2$; sample by columns and we are interested in $p_1^* - p_2^*$.

There are six parameters of most interest to a researcher:

$$p_A, p_B, p_1, p_2, p_1^* \text{ and } p_2^*.$$

Now, here are the really important facts to note:

1. For Type 1 sampling, all six of these parameters can be estimated.

2. For Type 2 sampling, only p_1 and p_2 can be estimated; the other four cannot be estimated.

3. For Type 3 sampling, only p_1^* and p_2^* can be estimated; the other four cannot be estimated.

We can see the truth of these facts with a simple, but extreme, example. In the sex and lenses study, suppose that I decide to take a Type 2 sample as follows: $n_1 = 10$ of the 600 females and all 400 males. Then I will get the following data:

	B	B^c	Total
A	a	$10 - a$	10
A^c	140	260	400
Total	$140 + a$	$270 - a$	410

We know from the population counts that $P(A|B) = p_1^* = 360/500 = 0.72$. But our estimate of p_1^* from this table will be $a/(140 + a)$ which can range from a minimum of 0 when $a = 0$ to a maximum of 0.067 when $a = 10$. In other words, our estimate of 0.72 will never be larger than 0.067! **This is a very bad estimate!** Think about this example and make sure that you understand why it is happening.

Before leaving this section, I want to revisit the screening test example in the context of our three types of sampling.

Medical researchers typically do not use Type 1 sampling for the following two reasons.

1. Many diseases are quite rare, making $P(A)$, $P(AB)$ and $P(AB^c)$ very small. As a result, one would require a huge sample size to estimate these well. In the best of situations a huge sample size is expensive. Also, nobody would spend a huge amount of money on sampling just to (possibly) learn that the screening test is ineffective.

2. We skirt around the issue of how difficult it is to obtain a random sample. For many diseases, the sufferers are people who are not particularly easy to find, which, of course, is an important step in being in a sample. For example, IV drug users and sex workers are two groups that have high rates of HIV infection, but are difficult to locate for a study. And, if located, they *might* be reluctant to participate.

Type 3 sampling is not possible because researchers will not have lists of people who will test positive (negative) before they collect data!

As a result, the most realistic way to sample is Type 2: Select what you hope is a random sample from people who clearly have the disease and an independent random sample from people who seem to not have the disease and proceed.

Thus, confirming what I stated earlier, because they use Type 2 sampling, medical researchers can estimate p_1 and p_2. They cannot estimate either p_1^* or p_2^*. Thus, it is understandable why they refer to the former with the definite article 'the.'

Table 16.9: Hypothetical population counts, in thousands.

	Outcome		
Risk Factor	Bad (B)	Good (B^c)	Total
Present (A)	24	276	300
Absent (A^c)	28	672	700
Total	52	948	1,000

16.3 Relative Risk and the Odds Ratio

Let's return to the model of a finite population with two responses, as introduced at the beginning of this chapter. The following situation is common in medical studies. The first response is the presence (A) or absence (A^c) of a risk factor. The second response is a health outcome, either bad (B) or good (B^c). Here is an example. A pregnant woman is a smoker (A) or nonsmoker (A^c). Her baby's birth-weight is either low (B) or normal (B^c).

We will begin with a table of hypothetical population counts, presented in Table 16.9. Typically, a researcher is interested in comparing $P(B|A)$ to $P(B|A^c)$. These are the probabilities of the bad outcome conditional on the risk feature being present or absent, respectively. Even though B is an undesirable outcome, we label it a success because in these medical problems it is often rare. Even if it's not rare in a particular problem, we still call it a success to avoid confusion. (Having B sometimes be a success and sometimes be a failure *would* definitely confuse me!) Using Chapter 15 notation, I write $p_1 = P(B|A)$ and $p_2 = P(B|A^c)$.

We note that for the hypothetical population counts in Table 16.9, $p_1 = 24/300 = 0.08$ and $p_2 = 28/700 = 0.04$. We could compare these by subtracting:

$$p_1 - p_2 = 0.08 - 0.04 = 0.04,$$

in words, the probability of the bad outcome is larger by 0.04 when the risk factor is present compared to when the risk factor is absent. Because both p_1 and p_2 are small, we might want to compare them by dividing: p_1/p_2. This ratio is called the **relative risk**.

For the population counts in Table 16.9, the relative risk equals $p_1/p_2 = 0.08/0.04 = 2$. In words, the presence of the risk factor doubles the probability of the bad outcome.

Scientifically, it would make sense to want to estimate $p_1 - p_2$ or p_1/p_2. Recall the three types of random sampling that were introduced on page 398. Type 1 sampling is rarely used because it is difficult to get a random sample from the entire population and even if we could obtain one, with a rare bad outcome we won't get enough data for subjects with bad outcomes to learn very much. Also, it can be difficult to perform Type 2 sampling, because there won't be lists of people based on the presence or absence of the risk factor. Type 3 sampling is, however, often reasonable. One can use hospital records to obtain—one hopes—Type 3 random samples. In fact, a study based on Type 3 sampling is called a **case-control study**. As we have argued earlier, however, with Type 3 sampling we cannot estimate p_1 or p_2. We can handle this difficulty, somewhat, by introducing the notion of the **odds ratio**.

Odds are an alternative to probabilities as a measure of uncertainty. For example, consider one cast of a balanced die and suppose we are interested in the outcome '4.' We have said that the probability of the outcome '4' is $1/6$, but we could also say that the odds of the '4' are 1 (the number of outcomes that give '4') to 5 (the number of outcomes that don't give '4'), or $1/5$. In general, if the event of interest has probability p, then the odds of the event is $p/(1-p)$.

The current paragraph is enrichment. I recommend you read it, but you will not be responsible for its content. *Be careful with language!* I am talking about the *odds of* an event. Sometimes people—especially in gambling contexts—speak of the *odds against* an event, which is the inverse of of *odds of* the event. Thus, a gambler would say that, when casting a balanced die, the odds against a '4' are the inverse of $1/5$, which is 5, usually stated as 5 to 1. But gambling is actually more complicated because bookies and casinos modify the *true odds against* an event to ensure themselves a long-run profit. We saw this earlier with the roulette wheel example. The probability of red is $18/38$. Remembering that the odds of an event with probability p is $p/(1-p)$, the odds against the event is $(1-p)/p$. Thus, the odds against red are $20/18 = 10/9$. It would be a *fair bet* if the casino paid \$19 for a \$9 bet that wins. Casinos, of course, have no desire to offer fair bets. An easy way for them to deal with this is to take the actual odds against, $10/9$ for the roulette bet, and *shorten (reduce) the odds*; in the case of roulette they shorten the odds against from the true $10/9$ to the profitable 1.

For the problem of this section, the odds of B in row A is $p_1/(1-p_1)$ and the odds of B in row A^c is $p_2/(1-p_2)$. In terms of the symbols in Table 16.1, these odds are N_{AB}/N_{AB^c} and $N_{A^cB}/N_{A^cB^c}$, respectively. Thus, their ratio, called the **odds ratio**, is:

$$\frac{N_{AB}N_{A^cB^c}}{N_{AB^c}N_{A^cB}}. \tag{16.6}$$

The great thing about this formula is the following. We have defined the odds ratio in terms of the *row probabilities*; i.e., the conditional probabilities of the B's given the A's. But an examination of this formula shows that it is symmetric in the arguments A and B; hence, the odds ratio remains the same if we define it in terms of the column probabilities. Thus, and this is the important part, we can estimate the odds ratio for any of our three types of sampling.

Finally, a little bit of algebra shows that

$$\text{odds ratio} \;=\; \text{relative risk} \;\times \frac{1-p_2}{1-p_1}.$$

Thus, if both p_1 and p_2 are close to zero, the odds ratio is approximately equal to the relative risk; and we know that the relative risk is interesting.

For the population in Table 16.9, the odds ratio is

$$\frac{24(672)}{276(28)} = 2.087,$$

which is close to the relative risk, previously shown to equal 2.

We will now discuss how to estimate the odds ratio. Note that this method is valid for any of our three types of random sampling. Table 16.10 presents our notation for the data we collect, our

Table 16.10: Notation for data for estimating the odds ratio.

	Outcome		
Risk Factor	Bad (B)	Good (B^c)	Total
Present (A)	a	b	n_1
Absent (A^c)	c	d	n_2
Total	m_1	m_2	n

ubiquitous 2×2 table of data in Table 15.1. Our point estimate of the odds ratio, θ, is

$$\hat{\theta} = (ad)/(bc).$$

I placed the population in Table 16.9 in my computer and simulated a case-control study with $m_1 = m_2 = 200$. My data are in Table 16.11. My estimated odds ratio is

$$\hat{\theta} = [89(131)]/[69(111)] = 1.522,$$

which is considerably smaller than the population value, which Nature alone knows to be 2.087.

We can obtain a confidence interval estimate of θ but it's a bit involved. We actually obtain a confidence interval estimate of $\lambda = \ln(\theta)$, where by 'ln' I mean the natural logarithm that is popular in calculus and has base $e = 2.71828\ldots$.

Our point estimate of λ is $\hat{\lambda} = \ln(\hat{\theta})$. For our data, $\hat{\lambda} = \ln(1.522) = 0.4200$.

The approximate confidence interval estimate of λ is:

$$\hat{\lambda} \pm z^* \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}. \tag{16.7}$$

The 95% confidence interval estimate of λ for the data in Table 16.11 is:

$$0.4200 \pm 1.96 \sqrt{\frac{1}{89} + \frac{1}{69} + \frac{1}{111} + \frac{1}{131}} =$$

$$0.4200 \pm 1.96(0.2058) = 0.4200 \pm 0.4035 = [0.0165, 0.8235].$$

To get back to a confidence interval estimate of θ, we exponentiate the endpoints of this interval:

$$(2.71828)^{0.0165} = 1.017 \text{ and } (2.71828)^{0.8235} = 2.278.$$

Thus, finally, the 95% confidence interval estimate of the odds ratio is $[1.017, 2.278]$. This is not a very useful interval; it is correct, because it contains $\theta = 2.087$, but its lower bound is so close to 1, the risk factor being present might not be much of a danger. (An odds ratio of 1 means that $p_1 = p_2$.) More data are needed.

Table 16.11: Simulated data for estimating the odds ratio from the population given in Table 16.9.

	Outcome		
Risk Factor	Bad (B)	Good (B^c)	Total
Present (A)	89	69	158
Absent (A^c)	111	131	242
Total	200	200	400

16.4 Comparing $P(A)$ to $P(B)$

For a finite population, this section assumes we have a Type 1 random sample; i.e., a random sample from the entire population. For trials, the assumption is that we have i.i.d. dichotomous trials; i.e., the table of probabilities is the same for each trial and trials are independent. This does **not** mean that A and B are independent—see Practice Problem 2.

In this section, I will focus on the entries $P(A)$ and $P(B)$ in Table 16.3 on page 389. In many studies, it would be difficult to explain an interest in comparing these two probabilities. In my hypothetical study of sex and lenses, why would I want to compare the proportion of females in the population ($P(A)$) with the proportion of people who would answer the question with 'yes' ($P(B)$)? Similar comments are true for the studies: of a screening test; my dog barking at squirrels; the relationship between a risk factor and a bad outcome.

In my study of Larry Bird, however, $P(A)$ is the probability that he makes the first of the pair of free throws and $P(B)$ is the probability that he makes the second of the pair of free throws. As a result, it is natural to compare $P(A)$ and $P(B)$ to investigate whether he is more skilled on his first or second shot.

I will not begin this section, however, with this basketball example. Instead we will revisit a technique we used extensively in Part I of these *Course Notes*: computer simulation experiments.

16.4.1 The Computer Simulation of Power

Refer to Table 9.8 on page 203. This table presents the results of a power study for a variety of constant treatment effect alternatives for Sara's golf study. It should not be a problem if your recollection of this topic and this table are a bit hazy. My point is more about simulation experiments and not really so much about power.

Recall that I stated that the number of possible assignments for Sara's golf study is 1.075×10^{23}. I am interested in the finite population that consists of

$$N = 1.075 \times 10^{23}$$

members, with each member being a possible assignment. Each member's card contains two dichotomous pieces of information:

- A if test statistic U would reject the null hypothesis for the card's assignment; and A^c if test statistic U would fail to reject the null hypothesis for the card's assignment.

Table 16.12: Results of a computer simulation experiment on the power of test statistics U and R_1 for the alternative of a constant treatment effect of 3 yards for Sara's golf study.

	R_1		
U	Reject (B)	Fail to reject (B^c)	Total
Reject (A)	1,136	153	1,289
Fail to reject (A^c)	398	8,313	8,711
Total	1,534	8,466	10,000

- B if test statistic R_1 would reject the null hypothesis for the card's assignment; and B^c if test statistic R_1 would fail to reject the null hypothesis for the card's assignment.

Recall from Table 9.7 on page 201, that the critical regions for these tests are $U \geq 10.65$ and $R_1 \geq 1788.0$. The critical values—10.65 and 1788.0—were obtained via a computer simulation experiment with 10,000 reps and both give, approximately, $\alpha = 0.0499$ as the probability of a Type 1 error.

In Table 9.8 on page 203, I report the results of six computer simulation experiments, one each for six hypothesized values of the constant treatment effect. Thus, I have data for six different population boxes; one box for each of the effects studied. I will restrict attention for now to a constant treatment effect of 3 yards.

With the above set-up, $P(A)$ is the probability that test statistic U would correctly reject the null hypothesis given that the constant treatment effect equals 3 yards. Similarly, $P(B)$ is the probability that test statistic R_1 would correctly reject the null hypothesis given that the constant treatment effect equals 3 yards. If we could examine all 1.075×10^{23} possible assignments, then we would know the values of $P(A)$ and $P(B)$ and, hence, be able to determine which test statistic is more powerful for the alternative we are studying. Of course, we might learn that $P(A) = P(B)$, which would tell us that the tests are equally powerful for the given alternative.

Instead of examining all possible assignments, I used a computer simulation experiment with 10,000 reps; in the terminology introduced in Chapter 10, I selected a dumb random sample of size $n = 10{,}000$.

I begin by reprinting the results of my computer simulation experiment in Table 16.12. The careful reader will note that I have interchanged the rows and columns in my Chapter 9 table to obtain this Chapter 16 table, so that, in the current chapter's set-up $P(A)$ and $P(B)$ refer to power. I will now show you how to use the data in Table 16.12 to make an inference about the values of $P(A)$ and $P(B)$.

Obviously, we can consider these probabilities separately. In particular, focusing on test statistic U, we find that our point estimate of $P(A)$ is 0.1289 and the approximate 95% confidence interval estimate of $P(A)$ is:

$$0.1289 \pm 1.96\sqrt{0.1289(0.8711)/10{,}000} = 0.1289 \pm 0.0066 = [0.1223, 0.1355].$$

Similarly, our point estimate of $P(B)$ is 0.1534 and the approximate 95% confidence interval

estimate of $P(B)$ is:

$$0.1534 \pm 1.96\sqrt{0.1534(0.8466)/10{,}000} = 0.1534 \pm 0.0071 = [0.1463, 0.1605].$$

These intervals are comfortably separated; the upper bound of the former is 0.0108 smaller than the lower bound of the latter. Thus, it seems clear that R_1 is more powerful than U.

Our confidence interval formula for comparing two proportions requires us to have independent random samples. What would this entail for the current problem? We would select a dumb random sample of 10,000 assignments to evaluate the performance of U; then we would again select a dumb random sample of 10,000 assignments to evaluate the performance of R_1. In other words, we would perform a computer simulation experiment with 10,000 reps for U and then perform a computer simulation experiment with 10,000 reps for R_1. I could have done this; why didn't I?

Well, let's pretend that I had performed two computer simulation experiments and again obtained point estimates of 0.1289 and 0.1534. Let's find the 95% confidence interval estimate of $P(A) - P(B)$:

$$(0.1289 - 0.1534) \pm 1.96\sqrt{\frac{0.1289(0.8711)}{10{,}000} + \frac{0.1534(0.8466)}{10{,}000}} =$$

$$-0.0245 \pm 0.0096 = [-0.0341, -0.0249].$$

Please note the half-width of this interval: 0.0096. When we learn the correct confidence interval estimate later in this section, we will see that its half-width is only about one-half as much, 0.0046. Thus, by having one sample of assignments and using each assignment twice, we reduce the half-width, in this example, by a factor of two when compared to having two independent samples.

Let's now learn the correct way to analyze the data in Table 16.12. We will begin with a test of hypotheses.

Based on Occam's Razor, the natural null hypothesis is $P(A) = P(B)$. There are three natural possibilities for the alternative hypothesis:

$$H_1\text{: } P(A) > P(B);\text{ or } H_1\text{: } P(A) < P(B);\text{ or } H_1\text{: } P(A) \neq P(B).$$

Using the symbols in Table 16.3 on page 389,

$$P(A) = P(B) \text{ becomes } P(AB) + P(AB^c) = P(AB) + P(A^cB),$$

$$\text{which, after canceling, becomes } P(AB^c) = P(A^cB).$$

Similarly, the alternative $>$ is equivalent to

$$P(AB^c) > P(A^cB);$$

the alternative $<$ is equivalent to

$$P(AB^c) < P(A^cB);$$

and the alternative $\neq$ is equivalent to

$$P(AB^c) \neq P(A^cB).$$

Below is a partial reproduction of the table of probabilities:

	B	B^c	Total
A		$P(AB^c)$	
A^c	$P(A^cB)$		
Total			

The two probabilities in this picture are the only ones that matter for the test of hypotheses.

This is a very interesting result. Of the four cells in the population table, only two are relevant to our null hypothesis.

Let's look at the data in Table 16.12. In this table, we have the results of a dumb random sample of size 10,000, but the numbers $a = 1{,}136$ and $d = 8{,}313$ are irrelevant for our test of hypotheses. What are relevant are the numbers $b = 153$ and $c = 398$.

Define $m = b + c$ which equals $153 + 398 = 551$ for our data. I call m the **effective sample size**. The sample size is 10,000, but of these observations, only $m = 551$ are relevant for our test. We can think of m as the observed value of the random variable M; before collecting data, we define M to be the number of observations that fall into either of the cells on the *off diagonal*: AB^c or A^cB.

This next part is a mess notationally, but a simple idea. Conditional on the knowledge that an observation falls in an off diagonal cell, label the upper right cell—AB^c—a success and the lower left cell—A^cB—a failure. Thus, conditional on the value $M = m$, we have m Bernoulli trials. For these m Bernoulli trials, the probability of success is

$$P(AB^c | AB^c \text{ or } A^cB) = \frac{P(AB^c)}{P(AB^c) + P(A^cB)} \text{ which we will call } p.$$

Now here is the key point to note.

- If the null hypothesis is true, then $p = 0.5$.
- If the alternative hypothesis $>$ is true, then $p > 0.5$.
- If the alternative hypothesis $<$ is true, then $p < 0.5$.
- If the alternative hypothesis $\neq$ is true, then $p \neq 0.5$.

Thus, our test of hypotheses can be viewed as testing

$$H_0\text{: } p = 0.5 \text{ versus } H_1\text{: } p > 0.5; \text{ or } H_1\text{: } p < 0.5; \text{ or } H_1\text{: } p \neq 0.5,$$

where p is the probability of success for m Bernoulli trials. **We already know how to perform this test!** Its P-value is given by Result 12.2 on page 306 with $p_0 = 0.5$.

Because our special value of interest is 0.5, the the null sampling distribution is Bin$(m, 0.5)$, which is symmetric. As a result, the two terms that are summed for the alternative $\neq$ are the same number.

In the context of this chapter, the test from Chapter 12 is called **McNemar's test**. I will use our data on power to remind you how to use the website:

`http://stattrek.com/Tables/Binomial.aspx`

to obtain the exact P-value for McNemar's test.

After accessing the site, enter 0.5 as the **Probability of success**; enter $m = 551$ for the **Number of trials**; and enter $b = 153$ for the **Number of successes**. Click on calculate and you will obtain five probabilities. The two that are relevant are:

- $P(X \leq 153) = 1.920 \times 10^{-26}$; and
- $P(X \geq 153) = 1$.

The first of these is the exact P-value for the alternative $<$; the second of these is the exact P-value for the alternative $>$; and twice the first of these is the exact P-value for the alternative $\neq$. I should have mentioned it earlier, but the only defensible choice for the alternative for the power study is $\neq$. If one is trying to decide which test statistic is more powerful, why would one ever use a one-sided alternative!

Thus, the exact P-value for my alternative is 3.840×10^{-26}. **This is a very small P-value**! Could it be an error? Well, the mean and standard deviation of the Bin(551,0.5) distribution are

$$\mu = 551(0.5) = 275.5 \text{ and } \sigma = \sqrt{551(0.5)(0.5)} = 11.73.$$

Thus, the observed value $b = 153$ is more than ten standard deviations below the mean! As a statistician I tend to be cautious; but not in this problem! I **know** that the sum of ranks test is more powerful than the comparison of means test for the alternative we have considered.

We went to a great deal of effort to motivate the test of hypotheses for comparing $P(A)$ and $P(B)$; for a change-of-pace, I simply will give you the approximate confidence interval estimate of $P(A) - P(B)$. (Actually, the formula below is a special case of a formula you will learn in Chapter 20.) The approximate confidence interval estimate of $P(A) - P(B)$ is:

$$(\frac{b-c}{n}) \pm (z^*/n)\sqrt{\frac{n(b+c) - (b-c)^2}{n-1}}, \tag{16.8}$$

where, as usual, the value of z^* depends on one's choice of the confidence level, as given in Table 12.1. I will illustrate the use of this formula for the data of our study of power. Recall that $b = 153$, $c = 398$ and $n = 10{,}000$. First, $(b - c) = -245$; $(b + c) = 551$; and

$$n(b+c) - (b-c)^2 = 10000(551) - (-245)^2 = \ 5{,}449{,}975.$$

Thus, the approximate 95% confidence interval estimate of $P(A) - P(B)$ is:

$$-0.0245 \pm 0.000196\sqrt{\frac{5{,}449{,}975}{9{,}999}} = -0.0245 \pm 0.0046 = [-0.0291, -0.0199].$$

Note, as mentioned earlier, the half-width of this confidence interval is $h = 0.0046$.

The above computation of the confidence interval is a bit nasty! Because $n = 10{,}000$ is very large, the term under the square root sign is large and messy to obtain. When our 'data' come

from a computer simulation experiment, n is frequently very large. In these cases, there is an approximate formula which is much easier to use. In particular, consider the term under the square root. Its numerator is

$$n(b+c) - (b-c)^2.$$

For the simulation experiment above, this term becomes

$$n(b+c) - (b-c)^2 = 10000(551) - (-245)^2 = \text{ 5,449,975.}$$

The point of this argument is that the term $(b-c)^2$ has a negligible effect on the answer; with it, we obtain 5,449,975; without it, we would obtain 5,510,000 The former is only 1.1% smaller than the latter. If we exclude the $(b-c)^2$ term, then square root term in Formula 16.8 reduces to:

$$\sqrt{(n/n-1)(b+c)}.$$

For $n =$ 10,000, the ratio $n/(n-1)$ is almost one. If we ignore it, Formula 16.8 reduces to the much simpler formula:

$$(\frac{b-c}{n}) \pm (z^*/n)\sqrt{b+c}, \tag{16.9}$$

If I use this new formula for our earlier data, which had:

$$b = 153, c = 398 \text{ and } n = \text{ 10,000, we get}$$

$$-0.0245 \pm 0.000196\sqrt{153+398} = -0.0245 \pm 0.0046;$$

the same answer we obtained from Formula 16.8.

Let's apply the above inference methods to the Larry Bird data in Table 15.16 in Chapter 15. I will reproduce his data below:

First	Second Shot		
Shot:	Hit	Miss	Total
Hit	251	34	285
Miss	48	5	53
Total	299	39	338

We see that

$$b = 34; c = 48; b + c = m = 82; b - c = -14; \text{ and } n = 338.$$

For the test of $H_0 : P(A) = P(B)$, I choose the alternative $\neq$. and go to the website:

`http://stattrek.com/Tables/Binomial.aspx`.

I enter: 0.5 for the **Probability of success**; $m = 82$ for the **Number of trials**: and $b = 34$ for the **Number of successes**. I click on **Calculate** and obtain:

$$P(X \geq 34) = 0.9515 \text{ and } P(X \leq 34) = 0.0753.$$

Thus, the P-value for $>$ is 0.9515; the P-value for $<$ is 0.0753; and the P-value for my alternative, $\neq$ is $2(0.0753) = 0.1506$. There is evidence that Bird's success probability on his second shot was larger than his success probability on his first shot, but the evidence is not convincing.

This is a very wide interval; 5.2 percentage points—its half-width—is a substantial amount in free throw shooting. The point estimate, -0.041, suggests that Bird *might have been much better* on his second shots, but the data are inconclusive.

By the way, the approximate 95% confidence interval for $P(A) - P(B)$ is:

$$(-14/338) \pm (1.96/338)\sqrt{\frac{338(82) - (-14)^2}{337}} = -0.041 \pm 0.052 = [-0.093, 0.011].$$

If I use the approximate (simpler) Formula 16.9, I obtain:

$$-0.041 \pm (1.96/338)\sqrt{34 + 48} = -0.041 \pm 0.053.$$

The approximation, while excellent, is not quite as good as before because $n = 338$ instead of 10,000. As a practical matter, however, to me a half-width of 0.053 has the same meaning as a half-width of 0.052.

16.5 Paired Data; Randomization-based Inference

In the largest sense, every example in this chapter has paired data. In one study, each person's sex is *paired with* the person's response on lenses. In another study, Casey's barking behavior is paired with the foraging of squirrels. I would call neither of these, however, paired data. For me, in this chapter, I reserve the term *paired data* to situations in which I want to compare $P(A)$ and $P(B)$. Two studies fit this criterion: our revisit of the Chapter 9 study of power and our analysis of free throw data.

There is another way to view paired data that is useful: the two examples in this chapter are cases of **reusing units** (or reusing subjects or trials). For example, in the study of power our subjects are assignments and we obtain two responses/features—i.e., we reuse—from each subject. Similarly, Larry Bird goes to the line to attempt a pair of free throws; the trial gives us his response on the first attempt and then we reuse it to obtain his response on the second attempt.

I chose to analyze Bird's data with a population model. This means that the 338 trials in our data set are viewed as the realization of 338 i.i.d. trials. This extra assumption creates a **mathematical structure** in which the quantities $P(A)$ and $P(B)$ make sense; thus allowing us to derive various confidence interval estimates—for $P(A)$, $P(B)$ and $P(A) - P(B)$. Without a population model, I cannot use Bird's data for inference because randomization is not possible—as you will soon see, randomizing would mean I could randomize the order of his shots, but clearly the first shot must be taken before the second shot.

My first example below is one that has been a favorite of teachers of Statistics throughout my career, but I have no idea whether such a study has ever been conducted! Imagine that we want to compare two therapies—call them cream 1 and cream 2—as a treatment for acne. As I hope these names suggest, the therapies would be *applied to the skin* as opposed to being taken orally or given in a shot. This will be important.

Arguably, a numerical response would be natural, but let's assume that the response is a dichotomy, with possibilities *improved*—a success—and *not improved*—a failure. Suppose that we had 100 persons suffering from acne available for study. We could perform a CRD, using either randomization-based inference—Chapter 8—or population-based inference—Chapter 15. We could, however, do something else, which I will now describe.

We could reuse each subject. In particular, if Bert has acne, we could tell him to put cream 1 on the left side of his face and cream 2 on the right side of his face. After the period of treatment, we would obtain *two responses* from Bert. We might, for example, code the responses as:

- A if cream 1 yields a success;
- A^c if cream 1 yields a failure;
- B if cream 2 yields a success; and
- B^c if cream 2 yields a failure.

Here are two natural questions:

1. How did the researcher decide that cream 1 would be applied to the left side Bert's face?
2. How about the other subjects in the study; Would they all apply cream 1 to the left sides of their faces?

Let's consider the second question first. It would be bad science to have every subject put cream 1 on the left side and cream 2 on the right side. Well, perhaps I should say potentially bad science. I don't really know whether side of the face influences the response. But if I performed the study in this way, all I can legitimately claim is that I have compared left-side-cream 1 to right-side-cream-2; in other words, the effect of the side of the face would be completely **confounded** with the effect of the type of cream.

There are two solutions to this side-of-the-face issue. The first is simple randomization. Suppose that there are 100 persons available for study. A separate randomization is performed for each of the 100 persons. For example, for Bert, side left or right is selected at random; cream 1 is applied to the selected side and cream 2 is applied to the remaining side. Thus, with 100 persons the overall randomization will involve 100 small randomizations.

Suppose that we do indeed have 100 subjects for study and we perform the 100 separate randomizations as described above. Let X denote the number of subjects who will place cream 1 on the left sides of their faces. Clearly, the probability distribution of X is Bin(100,0.5). We know from earlier work—or it can be easily verified with our binomial calculator website— that $P(X = 50)$ is small; only 0.0796. Thus, there is a 92% chance that one of the creams will be applied to more left sides than the other cream. Also, $P(40 \leq X \leq 60) = 0.9648$; thus, a discrepancy of more than 20 from side-to-side is unlikely, but not unimaginable. If the researcher feels strongly that side-of-face has a strong influence on the response, the fact that randomization could lead to one cream getting an additional 20 or more *good sides* is disturbing.

A solution to the above issue—namely, that randomization will likely lead to $X \neq 50$—is given by a **cross-over design**. In a cross-over design we select only one assignment, which is a desirable

simplification over the 100 needed above. The one assignment selects 50 subjects at random from the total of 100. Each of the 50 selected subjects puts cream 1 on the left side of the face and each of the remaining 50 subjects puts cream 1 on the right side of the face.

An obvious question is: Why not **always** use a cross-over design? The cross-over design requires a more complicated analysis and we don't have time to present it in this course. (More accurately, I have made an executive decision not to present it.)

I conjecture that the above acne example has remained popular with teachers of Statistics because it is such a natural example of a medical issue that can be treated two different ways simultaneously. The next step in the hierarchy would be a medical issue which can be addressed with only one therapy at a time, but is recurrent. An obvious example—one familiar to readers of these notes—would be a study of headaches. Artificial Headache Study-2 (HS-2) was introduced in Example 8.4 on page 168 and its data are in Table 8.4. For convenience, I reproduce its data below with the names of the drugs changed to 1 and 2:

	Pain relieved?			Row Prop.	
Drug :	Yes	No	Total	Yes	No
1	29	21	50	0.58	0.42
2	21	29	50	0.42	0.58
Total	50	50	100		

The exact P-value for the alternative $\neq$ for Fisher's test is 0.1612—details not given.

We could modify this headache study to enable subject reuse. In particular, we would need two headaches per subject, with one headache treated by drug 1 and the other with drug 2. The treatment assigned to the first headache would be determined by randomization.

I will create two distinct artificial data sets to investigate the issue of whether or not sample reuse is a good strategy. I need to be careful. I want to compare each new data set to the data in HS-2. Recall that HS-2 required 100 subjects—50 assigned to each treatment—to obtain the total of 100 observations in the data table above. One hundred subjects *reused* would give us data on *200 headaches*, which seems to me to be giving an unfair advantage to subject reuse. Therefore, each of my two data sets for subject reuse has 50 subjects, giving a total of 100 headaches.

Next, I want both of my two new data sets to be comparable to the data from HS-2. Here is what I mean. With subject reuse, my 2×2 data table will look like the following.

	Drug 2		
Drug 1	Success (B)	Failure(B^c)	Total
Success (A)	a	b	n_1
Failure (A^c)	c	d	n_2
Total	m_1	m_2	50

In HS-2, drug 1 gives 58% successes and drug 2 gives 42% successes. To be fair (comparable) I must have these same numbers for both of my subject reuse data sets. Thus, their data tables will look like:

Table 16.13: The two extreme subject reuse data sets consistent with HS-2 data. For the alternative $\neq$, the P-value for $b = 8$ is 0.0078 and the P-value for $b = 29$ is 0.3222. For comparison, the P-value for the HS-2 data with the same alternative is 0.1612.

	Smallest Possible b				Largest Possible b		
	Drug 2				Drug 2		
Drug 1	Success (B)	Failure(B^c)	Total	Drug 1	Success (B)	Failure(B^c)	Total
Succ. (A)	21	8	29	Succ. (A)	0	29	29
Fail. (A^c)	0	21	21	Fail. (A^c)	21	0	21
Total	21	29	50	Total	21	29	50

	Drug 2		
Drug 1	Success (B)	Failure(B^c)	Total
Success (A)	a	b	29
Failure (A^c)	c	d	21
Total	21	29	50

Recall that in these tables, a counts the number of (reused) subjects who would achieve a success with both drugs; d counts the number of (reused) subjects who would achieve a failure with both drugs; b counts the number of (reused) subjects for whom—more picturesquely—drug 1 defeated drug 2; and c counts the number of (reused) subjects for whom drug 1 lost to drug 2.

You may verify—or you may simply trust me—that in this last table, b can take on any integer value from 8 to 29. I look at the extremes of these possibilities in Table 16.13. Let's take a few minutes to examine the information in this table.

For $b = 8$, the data are highly statistically significant, with an exact P-value of 0.0078. For $b = 29$, the exact P-value is very large, 0.3222. First, although I won't prove this, if I created similar tables for $b = 9, 10, 11, \ldots, 28$, we would find that the P-value for the alternative $\neq$—as well as the alternative $>$—would increase with the value of b. Indeed, we would find that for $b = 16$ the P-value is 0.1516 and for $b = 17$ the P-value is 0.1686. The table with $b = 17$ is:

	Drug 2		
Drug 1	Success (B)	Failure(B^c)	Total
Success (A)	12	17	29
Failure (A^c)	9	12	21
Total	21	29	50

In this table, slightly fewer than one-half of the subjects (actual count, 24 of 50) respond the same to each drug and slightly more than one-half of the subjects respond differently to the drugs. As a physician, I can't believe that the results would be dissimilar for so many subjects; thus, I would anticipate obtaining b that is smaller than 17 and, hence, opt for subject reuse because it would give me a smaller P-value; i.e., it would be more sensitive. Of course, I am not a physician; thus, the validity of my opinion is quite questionable. Someday if you are the researcher in a study like this, you will need to decide whether to have subject reuse.

Above I have talked about sample reuse for hypothetical studies of acne and headaches. If one opts for subject reuse, then one can perform population-based inference or the randomization-based inference of Part I of these notes; from either perspective, one uses McNemar's test and obtains the same P-value for any fixed alternative.

16.5.1 Maslow's Hammer Revisited

Above I talked about a hierarchy: for acne, a subject can receive two treatments simultaneously; for headaches a subject can receive two treatments serially. Intuition often suggests—and experience has often verified—that subject reuse can lead to a more efficient study. It is perhaps then simply human nature (Maslow's Hammer) that researchers seek other venues for subject reuse. In this subsection I will discuss one such situation, with lots of cautionary words for you.

Suppose that you have 200 patients with colon cancer and two therapies that you want to compare. The above ideas for acne and headaches clearly won't work. But here is an idea. Before assigning subjects to treatments, record the values of several variables on each individual and summarize these values with a single number. Let's assume that we believe that the larger the number, the better the subject's prognosis for a favorable response, be it a dichotomy or a number. In this scenario it is valid to do the following:

> Use the values of the 200 numbers to form 100 pairs of subjects in the following way. The subjects with the two largest numbers create a pair. Of the remaining 198 numbers, the subjects with the two largest numbers create a pair. And so on.

Once the 100 pairs are formed, within each pair select a subject at random to be given the first therapy; the other member of the pair will receive the second therapy.

Here is the really important point. If you proceed as above, then it is **valid to perform randomization-based inference** on the 100 pairs. I recommend against performing population-based inference in this case, but I don't have time to give my whole argument, except to note the following. Suppose that Bert and Walt are among my 200 patients and that they are paired as above. If I looked at the entire population of hundreds of thousands of people (I am guessing at the population size), then I would be amazed if Bert and Walt were paired in the population. Thus, it is not clear how sample pairs relate to population pairs, so I won't do it!

By the way, as you will see in Chapter 20, randomization is key. If one forms pairs in an observational study, the results are a disaster!

16.6 Summary

This chapter is concerned with the situation in which each unit (trial, subject) yields two dichotomous responses. We begin with units that are subjects; this situation, as before in these notes, leads us to define a finite population. We begin with the table of population counts, given in Table 16.1. If we divide each population count by the population size, N, we obtain the table of population proportions or probabilities, given in Table 16.3.

In practice, these tables—counts and probabilities—are known only to Nature. When units are trials, there is no notion of population counts, making the table of probabilities the starting point. Again, in practice, the table of probabilities is known only to Nature.

The table of probabilities allows us to define the very useful and interesting concept of conditional probabilities, of which there are eight. Thus, in addition to the eight probabilities in the table of probabilities, we have eight conditional probabilities; 16 is a lot of probabilities! But, no worries; we learn that there are really only three non-redundant (conditional or not) probabilities. There are many valid sets of three non-redundant probabilities and any set of them will yield the remaining 13 probabilities.

I introduce you to a very important use of the above ideas: screening tests for a disease. Also, you learn the important Bayes' rule (or formula), which allows us to *reverse the direction of conditioning*.

The next issue is: How to obtain data in order to perform inference on the various probabilities of interest. Three possibilities are considered; listed and described on page 398. Type 2 sampling—independent random samples from the rows—is mathematically equivalent to Type 3 sampling—independent random samples from the columns.

Of the 18 possible probabilities, most analyses focus on comparing the members of one or more of the following pairs: $P(A)$ and $P(B)$; p_1 and p_2; and p_1^* and p_2^*. Note that when I write p_1 and p_2, I am reverting to Chapter 15 notation; in Chapter 16 notation, $p_1 = P(B|A)$ and $p_2 = P(B|A^c)$. I prefer the Chapter 15 notation because it's not so messy! Also, when I write p_1^* and p_2^*, this is Chapter 15 notation in which the populations are in the columns and the success is in the first row. Again, p_1^* and p_2^* could be expressed in the messier conditional probability notation of Chapter 16.

The main result is that

1. For Type 1 sampling, all six of these parameters can be estimated.

2. For Type 2 sampling, only p_1 and p_2 can be estimated; the other four cannot be estimated.

3. For Type 3 sampling, only p_1^* and p_2^* can be estimated; the other four cannot be estimated.

I give a numerical example with Type 2 sampling that illustrates the second (and the mathematically equivalent third) of these results.

I then introduce you to a class of medical studies—the relationship between a risk factor and a bad outcome—for which we need to estimate p_1 and p_2, but neither Type 1 nor Type 2 sampling is realistic. Provided that p_1 and p_2 are relatively small, this difficulty can be overcome by focusing on the odds ratio rather than the relative risk, p_1/p_2. We can then use Type 3 sampling to estimate the odds ratio, which is the same number for columns as it is for rows.

I give two examples in which the researcher is interested in comparing $P(A)$ and $P(B)$. It is easy to find the exact P-value for the test of the null hypothesis that these two probabilities are equal. The test is called McNemar's test and is a special case of the test we learned in Chapter 12. There is also an approximate confidence interval estimate of $P(A) - P(B)$, given in Formula 16.8, that is based on a Normal curve approximation.

The chapter ends with a lengthy discussion of applications to medical studies. We consider studies with subject reuse and randomization. It makes sense that subject reuse should lead to a

more sensitive analysis, but in any given situation the researcher must decide which method to use; there are no guarantees in this area!

16.7 Practice Problems

1. Suppose that we are given the following table of population counts:

	B	B^c	Total
A	800	200	1,000
A^c	1,200	2,800	4,000
Total	2,000	3,000	5,000

   (a) Calculate the table of population probabilities.

   (b) Calculate the eight conditional probabilities.

2. Suppose that we are given the following table of population counts:

	B	B^c	Total
A	400	600	1,000
A^c	1,600	2,400	4,000
Total	2,000	3,000	5,000

   (a) Calculate the table of population probabilities.

   (b) Calculate the eight conditional probabilities.

   (c) Notice that every conditional probability is equal to the corresponding unconditional probability. Whenever this happens, we say that the two responses are statistically independent. With independence, the multiplication rule for conditional probabilities:

$$P(AB) = P(A)P(B|A), \text{ becomes } P(AB) = P(A)P(B),$$

   which corresponds to our definition of independence in Chapter 10. In words, the probability associated with cell AB is the product of its row and column marginal probabilities. It now follows that the two responses are statistically independent if, and only if, every one of the four cell probabilities is equal to the product of its row and column marginal probabilities.

   Of course, it's no fun to check this multiplication for every cell. Fortunately, it can be shown that this multiplicative relationship holds for either: all four cells or none of the cells. Thus, we need to check only one cell.

3. Below is a table of probabilities.

	B	B^c	Total
A	0.195	0.455	0.650
A^c	0.105	0.245	0.350
Total	0.300	0.700	1.000

Explain why you don't need to do any calculations to obtain the eight conditional probabilities.

4. On page 392, I stated that if we know:

> A conditional probability for each row plus one of the row marginal probabilities,

then we could obtain all eight of the probabilities in the table of probabilities. I will **not** prove this fact in all of its algebraic glory; instead, I will show you an example of *how to do it*.

For example, given $P(B|A) = 0.375$, $P(B|A^c) = 0.500$ and $P(A) = 0.800$, determine the eight probabilities in Table 16.3.

By the way, I also stated that if we know:

> A conditional probability for each column plus one of the column marginal probabilities,

then we could obtain all eight of the probabilities in the table of probabilities. I won't show you an example of this because it is just like the one I do show you, but with the rows and columns interchanged.

5. Below is the table of population counts for a disease and its screening test. (Recall that A means the disease is present and B means the screening test is positive.)

	B	B^c	Total
A	254	35	289
A^c	199	3126	3325
Total	453	3161	3614

(a) What proportion of the population would test positive?

(b) What proportion of the population is disease free?

(c) What proportion of the population is free of the disease and would test negative?

(d) What proportion of the population has the disease and would test positive?

(e) Of those who would test negative, what proportion has the disease?

(f) Of those who are free of the disease, what proportion would test positive?

(g) What proportion of the population would receive a correct screening test result?

(h) Of those who would receive an incorrect screening test result, what proportion would receive a false positive?

(i) What proportion of the population does not have the disease or would test negative?

6. My dog Casey would visit my neighbor Sally while she was shooting free throws. I could see Sally shoot, but I could not see the outcome of her shot. Because Sally was a professional poker player, she did not have a *tell*; i.e., as best I could discern, her reaction to a made shot was identical to her reaction to a missed shot. In other words, Sally and Casey could see everything; I could only see when shots were attempted. According to Sally, her free throws were Bernoulli trials with probability of success (made shot) $p = 0.80$. Also according to Sally, immediately after she made a shot, Casey would bark 70% of the time; immediately after she missed a shot, Casey would bark 40% of the time (Sally liked to feed squirrels). Casey would neither confirm nor refute Sally's numbers.

   Use the above information to answer the following questions.

   (a) Sally is preparing to shoot; what is the probability that Casey is about to bark?

   (b) Sally has shot and Casey has barked; what is the probability Sally made the shot?

   (c) Sally has shot and Casey is silent; what is the probability Sally made the shot?

7. Refer to the medical studies introduced in Section 16.3. The population counts are given by the following table, in thousands.

	B	B^c	Total
A	9	291	300
A^c	7	693	700
Total	16	984	1,000

   Calculate the values of p_1, p_2, the relative risk and the odds ratio for this population.

8. Refer to the previous problem. I used Minitab to generate data from this population using a case-control study (Type 3 sampling) and obtained the following data.

	B	B^c	Total
A	99	93	192
A^c	101	207	308
Total	200	300	500

   (a) Calculate the odds ratio for these data. Viewing this number as the point estimate of the population odds ratio, comment.

   (b) Obtain the approximate 95% confidence interval estimate of the population odds ratio. Is the interval estimate correct?

9. The data for Rick Roby shooting free throws is below:

First	Second Shot		
Shot:	Hit	Miss	Total
Hit	54	37	91
Miss	49	31	80
Total	103	68	171

(a) Find the exact P-values for each of the three possible alternatives for the test of the null hypothesis that $P(A) = P(B)$.

(b) Calculate the approximate 95% confidence interval estimate of $P(A) - P(B)$.

16.8 Solutions to Practice Problems

1. (a) I divide each count by $N =$ 5,000 and obtain the table of probabilities below:

	B	B^c	Total
A	0.16	0.04	0.20
A^c	0.24	0.56	0.80
Total	0.40	0.60	1.00

(b) The table of conditional probabilities of B's given A's is below.

	B	B^c	Total
A	$P(B\|A) = 0.80$	$P(B^c\|A) = 0.20$	1.00
A^c	$P(B\|A^c) = 0.30$	$P(B^c\|A^c) = 0.70$	1.00
	$P(B) = 0.40$	$P(B^c) = 0.60$	1.00

The table of conditional probabilities of A's given B's is below. For example, $P(A|B^c) = 0.07$.

	B	B^c	
A	$P(A\|B) = 0.40$	$P(A\|B^c) = 0.07$	$P(A) = 0.20$
A^c	$P(A^c\|B) = 0.60$	$P(A^c\|B^c) = 0.93$	$P(A^c) = 0.80$
Total	1.00	1.00	1.00

2. (a) I divide each count by $N =$ 5,000 and obtain the table of probabilities below:

	B	B^c	Total
A	0.08	0.12	0.20
A^c	0.32	0.48	0.80
Total	0.40	0.60	1.00

(b) The table of conditional probabilities of B's given A's is below.

	B	B^c	Total
A	$P(B\|A) = 0.40$	$P(B^c\|A) = 0.60$	1.00
A^c	$P(B\|A^c) = 0.40$	$P(B^c\|A^c) = 0.60$	1.00
	$P(B) = 0.40$	$P(B^c) = 0.60$	1.00

The table of conditional probabilities of A's given B's is below.

	B	B^c	
A	$P(A\|B) = 0.20$	$P(A\|B^c) = 0.20$	$P(A) = 0.20$
A^c	$P(A^c\|B) = 0.80$	$P(A^c\|B^c) = 0.80$	$P(A^c) = 0.80$
Total	1.00	1.00	1.00

3. We can see that $P(AB) = 0.195$ is equal to $P(A)P(B) = 0.650(0.300) = 0.195$. Thus, the two responses are independent (refer to the previous practice problem). As a result the marginal probabilities are equal to the conditional probabilities. For example

$$0.30 = P(B) = P(B|A) = P(B|A^c).$$

4. We proceed as follows. Given that $P(A) = 0.800$, we know that $P(A^c) = 0.200$. We put these two numbers into our table of probabilities:

	B	B^c	Total
A			0.800
A^c			0.200
Total			1.000

Next, we use my second statement of the multiplication rule for conditional probabilities, Equation 16.4, twice. First,

$$P(AB) = P(A)P(B|A) = 0.800(0.375) = 0.300.$$

Second, we use it with a slight change of names:

$$P(A^cB) = P(A^c)P(B|A^c) = 0.200(0.500) = 0.100.$$

We place these two newly acquired probabilities into our table, giving:

	B	B^c	Total
A	0.300		0.800
A^c	0.100		0.200
Total			1.000

Finally, after a flurry of additions and subtractions, we obtain:

	B	B^c	Total
A	0.300	0.500	0.800
A^c	0.100	0.100	0.200
Total	0.400	0.600	1.000

5. (a) $453/3614 = 0.125$; (b) $3325/3614 = 0.920$; (c) $3126/3614 = 0.865$; (d) $254/3614 = 0.070$; (e) $35/3161 = 0.011$; (f) $199/3325 = 0.060$; (g) $(254 + 3126)/3614 = 0.935$; (h) $199/(35 + 199) = 0.850$; (i) $(3614 - 254)/3614 = 0.930$.

6. This, of course, is a dreaded story problem. The primary challenge is to write the given information in the language of this chapter. First, we identify the trial: Sally attempts a free throw. Second, we identify the two dichotomous responses and give them labels: one response is the outcome of the shot and the other response is Casey's behavior. I will define B for Casey barking and A for Sally making her shot. This gives us the following table of probabilities with unknown probabilities missing:

	B	B^c	Total
A			
A^c			
Total			1.00

Next, we write one piece of the given information in symbols: $P(A) = 0.80$, which implies that $P(A^c) = 0.20$. We put this information in our table:

	B	B^c	Total
A			0.80
A^c			0.20
Total			1.00

Next, we use the multiplication rule for conditional probabilities twice:

$$P(AB) = P(A)P(B|A) = 0.80(0.70) = 0.56 \text{ and}$$

$$P(A^cB) = P(A^c)P(B|A^c) = 0.20(0.40) = 0.08.$$

We put this information in our table and do lots of adding and subtracting; the end result is:

	B	B^c	Total
A	0.56	0.24	0.80
A^c	0.08	0.12	0.20
Total	0.64	0.36	1.00

We are now ready to answer questions!

(a) The question asks for $P(B)$; from the table, $P(B) = 0.64$.

(b) The question asks for $P(A|B)$; from the definition of conditional probability,

$$P(A|B) = P(AB)/P(B), \text{ which equals } 0.56/0.64 = 0.875.$$

(c) The question asks for $P(A|B^c)$; from the definition of conditional probability,

$$P(A|B^c) = P(AB^c)/P(B^c), \text{ which equals } 0.24/0.36 = 0.667.$$

7. We have $p_1 = 9/300 = 0.03$ and $p_2 = 7/700 = 0.01$. The relative risk is $p_1/p_2 = 0.03/0.01 = 3$. The odds ratio is

$$\theta = \frac{9(693)}{291(7)} = 3.062.$$

8. (a) The odds ratio of these data is

$$\hat{\theta} = \frac{99(207)}{93(101)} = 2.182.$$

The point estimate is considerably smaller than the population value, 3.062.

(b) First, I compute

$$\hat{\lambda} = \ln(\hat{\theta}) = \ln(2.182) = 0.7802.$$

The 95% confidence interval estimate of λ is

$$0.7802 \pm 1.96\sqrt{1/99 + 1/93 + 1/101 + 1/207} = 0.7802 \pm 1.96(0.1886) =$$
$$0.7802 \pm 0.3697 = [0.4105, 1.1499].$$

This gives us the following 95% confidence interval estimate of θ:

$$e^{0.4105} \leq \theta \leq e^{1.1499} \text{ or } 1.508 \leq \theta \leq 3.158.$$

This interval is correct because it (barely) includes $\theta = 3.062$.

9. (a) Go to the website

`http://stattrek.com/Tables/Binomial.aspx`

to obtain the exact P-values for McNemar's test.

After accessing the site, enter 0.5 as the **Probability of success**; enter $m = 37 + 49 = 86$ for the **Number of trials**; and enter $b = 37$ for the **Number of successes**. Click on calculate and you will obtain five probabilities. The two that are relevant are:

- $P(X \leq 37) = 0.1177$
- $P(X \geq 37) = 0.9197$.

Thus, the P-value for $<$ is 0.1177; the P-value for $>$ is 0.9197; and the P-value for $\neq$ is $2(0.1177) = 0.2354$.

(b) For the confidence interval estimate, note that $(b - c) = -12$, $b + c = 86$ and $n = 171$. Thus, the interval is

$$(\frac{-12}{171}) \pm (1.96/171)\sqrt{\frac{171(86) - (-12)^2}{170}} = -0.0702 \pm 0.01146(9.255) =$$
$$-0.0702 \pm 0.1061 = [-0.1763, 0.0359].$$

In the data, Roby was a considerably better shooter on his second shot, but the confidence interval estimate is inconclusive.

16.9 Homework

1. This is a problem about a very good screening test for a very rare disease. You are given the following probabilities:

$$P(A) = 0.001, P(B|A) = 0.999 \text{ and } P(B|A^c) = 0.01.$$

Calculate $P(A^c|B)$. Comment.

Hint: Rather than work with very small probabilities, it might be easier to work with population counts. To this end, let the population size N be one million.

2. This problem is about relative risks and odds ratios. Below is a table of hypothetical population counts, in thousands.

Group	B	B^c	Total
A	12	188	200
A^c	12	788	800
Total	24	976	1000

A case-control study with 800 subjects from this population yielded the data below.

Group	B	B^c	Total
A	207	90	297
A^c	193	310	503
Total	400	400	800

(a) Calculate the relative risk and odds ratio for the population.

(b) Calculate the point estimate of the population odds ratio.

(c) Obtain the 95% CI for the population odds ratio.

3. A former student of mine, Jackie, planned to study her dog, Basia, with a total of 50 trials. Jackie wanted to study Basia's ability to catch a kernel of popped corn that has been tossed towards her. (I am guessing that Basia is a female.) Years of experience had convinced Jackie that Basia was very skilled at catching popcorn that was tossed directly at her. For her study, Jackie chose the following two treatments. For the first [second] treatment, Jackie would toss the popcorn approximately two feet to Basia's right [left]. A trial is labeled a success if, and only if, Basia catches the kernel before it hits the ground.

Jackie, of course, could perform a CRD and analyze it using the methods of Chapter 15—if she was willing to assume Bernoulli trials—or Chapter 8—if not. Here is another idea. We can take the 50 trials and form 25 pairs. Trials 1 and 2 form the first pair; trials 3 and 4 form the second pair; and so on. Jackie chose to perform this randomized pairs design; similar to the acne and headache studies, Jackie performed a separate randomization for each of her 25 pairs of trials. This is mathematically valid because Jackie it is a form of *reusing units*.

Jackie obtained the following results: Basia obtained a total of 16 successes on the first treatment; seven pairs of trials yielded two successes; and four pairs of trials yielded two failures.

(a) Use the information above to complete the following data table:

	Treatment 2		
Treatment 1:	Success	Failure	Total
Success			
Failure			
Total			25

(b) Obtain the exact P-value for each of the three possible alternatives.

(c) Now *pretend* that Jackie had performed a CRD and obtained *exactly the same data*. Obtain the exact P-value for each of the three possible alternatives. Compare these answers to your answers in (b).

4. Refer to the previous problem. Leigh performed 50 pairs of trials on her dog Attica. A trial consisted of Leigh standing near a window inside her home while Attica was reposed on the floor. For treatment 1, Leigh would *yell*, "Squirrel, Attica!" For treatment 2, Leigh would *calmly remark*, "Hey Attica, squirrel." Attica's response was classified into one of two categories—a success if she got excited and a failure if she did not move. There was a total of 37 successes on the first treatment and a total of only 16 successes on the second treatment. For only five pairs of trials did Attica give two successes.

(a) Present these data in a 2×2 table.

(b) Find the exact P-value for the alternative $>$.

(c) *Pretend* that Leigh had performed a CRD and obtained *exactly the same data*. Obtain the exact P-value for the alternative $>$.

Chapter 17

Inference for One Numerical Population

In Chapter 10 you learned about finite populations. You learned about smart and dumb random samples from a finite population. You learned that i.i.d. trials can be viewed as the outcomes of a dumb random sample from a finite population. Chapter 11 developed these ideas in the special case of a dichotomous response. This was a very fruitful development, leading to all the results not named Poisson in Chapters 12–16. And, of course, our results for the Poisson are related to our results for the binomial.

In Chapters 17–20 we mimic the work of Chapters 11–16, but for a numerical response rather than a dichotomy. First, you will see the familiar distinction between a finite population and a mathematical model for the process that generates the outcomes of trials. Second, you will see that responses that are counts must be studied differently than responses that are measurements. We begin by studying responses that are obtained by counting.

Before we get to count responses, let me lay out some notation for this chapter. Recall that either dumb random sampling from a finite population **or** the assumption that trials are i.i.d., result in our observing n i.i.d. random variables:

$$X_1, X_2, X_3, \ldots, X_n.$$

The probability/sampling distribution for each of these random variables is determined by the **population**. Recall that for a dichotomous response the population is quite simple; it is determined by the single number p. For a numerical response, as you will soon see, the population is more complex—it is a picture, not a single number. Finally, when I want to talk about a generic random variable—one observation of a trial or one population member selected at random—I will use the symbol X, without a subscript.

You may need—on occasion—to refer back to the preceding paragraph as you work through this chapter.

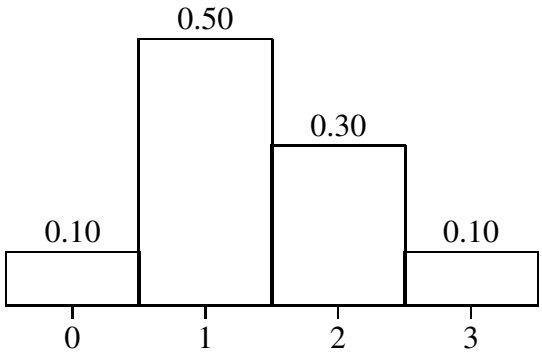
17.1 Responses Obtained by Counting

I will begin with finite populations.

Table 17.1: The population distribution for the cat population.

x	0	1	2	3	Total
$P(X = x)$	0.10	0.50	0.30	0.10	1.00

Figure 17.1: The probability histogram for the cat population.



17.1.1 Finite Populations for Counts

Please remember that the two examples in this subsection are both hypothetical. In particular, I claim no knowledge of cat ownership or household size in our society.

Example 17.1 (The cat population.) *A city consists of exactly 100,000 households. Nature knows that 10,000 of these households have no cats; 50,000 of these households have exactly one cat; 30,000 of these households have exactly two cats; and the remaining 10,000 households have exactly three cats.*

We can visualize the *cat population* as a population box that contains 100,000 cards, one for each household. On a household's card is its number of cats: 0, 1, 2 or 3. Consider the chance mechanism of selecting one card at random from the population box. (Equivalently, selecting one household at random from the city.) Let X be the number on the card that will be selected. It is easy to determine the sampling distribution of X and it is given in Table 17.1. For example, 50,000 of the 100,000 households have exactly one cat; thus $P(X = 1) = 50{,}000/100{,}000 = 0.50$. It will be useful to draw the probability histogram of the random variable X; it is presented in Figure 17.1. To this end, note that consecutive possible values of X differ by 1; thus, $\delta = 1$ and the height of each rectangle in Figure 17.1 equals the probability of its center value. For example, the rectangle centered at 1 has a height of 0.50 because $P(X = 1) = 0.50$. Either the distribution in Table 17.1 or its probability histogram in Figure 17.1 can play the role of the population. In the next section, we will see that for a measurement response the population is a picture, called the **probability density function**. (Indeed, the population **must be** a picture for mathematical reasons—trust me on this.)

Because we have no choice with a measurement—the population **is** a picture—for consistency, I will refer to the probability histogram of a count response as **the** population. Except when I don't; occasionally, it will be convenient for me to view the probability distribution—such as the one in Table 17.1—as being the population. As Oscar Wilde reportedly said,

> Consistency is the last refuge of the unimaginative.

It can be shown that the mean, μ, of the cat population equals 1.40 cats per household and its standard deviation, σ, equals 0.80 cats per household. I suggest you trust me on the accuracy of these values. Certainly, if one imagines a fulcrum placed at 1.40 in Figure 17.1, it *appears that the picture will balance*. If you really enjoy hand computations, you can use Equations 7.1 and 7.3 on pages 147 and 148 to obtain $\mu = 1.40$ and $\sigma^2 = 0.64$. Finally, if you refer to my original description of the cat population in Example 17.1, you can easily verify that the median of the 100,000 population values is 1. (In the sorted list, positions 10,001 through 60,000 are all home to the response value 1. Thus, the two center positions, 50,000 and 50,001 both house 1's; hence, the median is 1.) For future use it is convenient to have a Greek letter to represent the median of a population; we will use ν, pronounced as *new*.

You have now seen the veracity of my comment in the first paragraph of this chapter; the population for a count response—a probability histogram—is much more complicated than the population for a dichotomy—the number p.

Thus far with the cat population, I have focused exclusively on Nature's perspective. We now turn to the view of a researcher.

Imagine that you are a researcher who is interested in the cat population. All you would know is that the response is a count; thus, the population is a probability histogram. But *which* probability histogram? It is natural to begin with the idea of using data to *estimate* the population's probability histogram. How should you do that?

Mathematically, the answer is simple: Select a random sample from the population of 100,000 households. Provided that the sample size, n, is 5% or fewer of the population size, $N = 100{,}000$, whether the sample is smart or dumb matters little and can be ignored. For the cat population, this means a sample of 5,000 or fewer households. (It is beyond my imagination—see Wilde quote above—that a cat population researcher would have the energy and resources to sample more than 5,000 households!)

In practice, a researcher would attempt to obtain a sample for which the WTP assumption (Definition 10.3 on page 240) is reasonable.

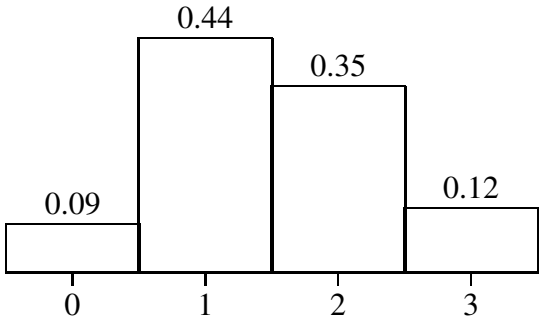
Because the cat population is hypothetical, I cannot show you the results of a real survey. Instead, I put the cat population in my computer and instructed my favorite statistical software package, Minitab, to select a random sample of $n = 100$ households. (I chose a dumb sample because it is easier to program.) The data I obtained are summarized in Table 17.2. In this table, of course, the researcher would not know the numbers in the *Unknown Probability* column; but Nature would. Nature can see that the researcher's relative frequencies are somewhat close to the unknown probabilities. Given that the population is a picture, it seems reasonable to draw a picture of the data. Which picture?

The idea is that I want to allow Nature to compare the picture of the data to the picture of the population. Thus, the dot plot is not a good idea—it's like comparing apples to music. A histogram

Table 17.2: Data from a simulated random sample of size $n = 100$ from the cat population in Table 17.1.

Value	Frequency	Relative Frequency	Unknown Probability
0	9	0.09	0.10
1	44	0.44	0.50
2	35	0.35	0.30
3	12	0.12	0.10
Total	100	1.00	1.00

Figure 17.2: The density histogram for the data in Table 17.2. This is our *picture-estimate* of the population in Figure 17.1.



seems reasonable, but which one? The natural choice is the density histogram because, like the probability histogram, its total area is one.

For a count response—but not a measurement response—we need one modification of the density histogram before we use it. Remembering back to Chapter 2, one of the reasons for drawing a histogram of data is to group values into class intervals—sacrificing precision in the response values for a picture that is more useful. In Chapter 7, we **never** grouped values for the probability histogram. Thus, when we use a density histogram to estimate the probability histogram of a population, we **do not** group values together. Without grouping, there is no need for an endpoint convention; thus, we modify slightly our method for drawing a density histogram. The modification is presented in Figure 17.2, our density histogram of the data in Table 17.2.

In Chapter 12 our population is a number, p. Our estimate of it, $\hat{p}$, is called a *point estimate* because it is a point/number estimating a point/number. In the current chapter, we estimate a picture—the probability histogram—by a picture—the density histogram; thus, it is natural to refer to the density histogram as the **picture-estimate** of the population.

Table 17.3: Data from a simulated random sample of size $n = 5{,}000$ from the cat population in in Table 17.1.

Value	Frequency	Relative Frequency	Unknown Probability
0	504	0.1008	0.1000
1	2,478	0.4956	0.5000
2	1,478	0.2956	0.3000
3	540	0.1080	0.1000
Total	10,000	1.0000	1.0000

A picture-estimate can be quite useful, especially if it is based on a large amount of data. For example, because the cat population is completely hypothetical, it is easy to generate a random sample of any size we want. I decided to select a dumb random of size $n = 5{,}000$ households from the cat population; my data are in Table 17.3. Even a cursory comparison of the numbers in the third and fourth columns of this table indicates that for a sample of size 5,000, the picture-estimate of the cat population is very nearly perfectly accurate.

Now I need to give you the bad news. In Chapter 12, you learned how to estimate the population **with confidence**. We called the result the confidence interval estimate of p because it consisted of an interval of numbers—in other words, it consisted of a bunch of *potential populations*. I wish that I could estimate a probability histogram with confidence, but that goal is unattainable. As a result, as a researcher, you must **choose either or both** of the following strategies.

1. Create the picture-estimate—the density histogram—and be happy with it, despite our inability to estimate with confidence.

2. Estimate some **feature** of the probability histogram with confidence. The feature most often estimated is the mean, μ; we will learn about another possible feature in Chapter 18.

Thus far in this chapter, I have restricted attention to a count response for a finite population; indeed, I have presented only one example! Nevertheless, the above two strategies are the choices you face for every numerical response, be it count or measurement.

I end this subsection with another example of a count response on a finite population. I want to remind you that the following example is totally hypothetical; I have chosen it for two reasons.

1. It will be convenient to have second and third specific examples—the cat population was the first—of a skewed population.

2. This example will be useful in Chapter 18 when I discuss various common errors in estimation.

Example 17.2 (The family size population.) *A community consists of exactly 7,000 households. The variable of interest is the number of children in the household who are attending public school. Population counts and two population distributions are given in Table 17.4. The two probability histograms are given in Figure 17.4.*

Table 17.4: The population counts and two distributions for the family size population.

Value:	0	1	2	3	4	5	6	7	Total
Counts:	2,800	1,260	840	714	546	420	294	126	7,000
				Population 1: All 7,000 households					
x :	0	1	2	3	4	5	6	7	Total
$P(X = x)$:	0.400	0.180	0.120	0.102	0.078	0.060	0.042	0.018	1.000
				Population 2: The 4,200 households with a positive response					
y :	0	1	2	3	4	5	6	7	Total
$P(Y = y)$:	0.000	0.300	0.200	0.170	0.130	0.100	0.070	0.030	1.000

Figure 17.3: The probability histograms for the two family size populations.

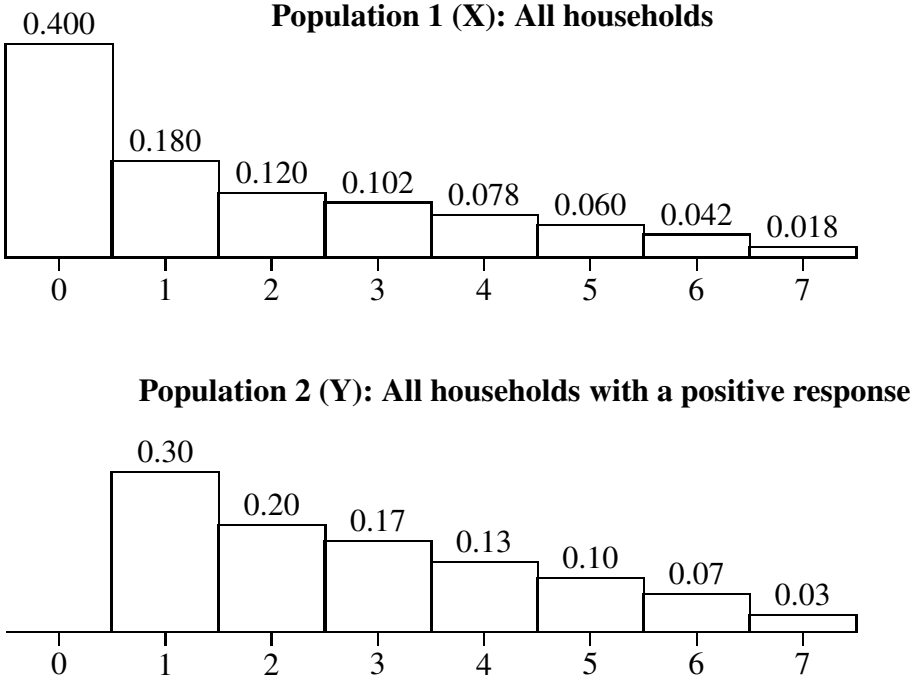


Table 17.5: The 36 quartets in solitaire mahjong.

Category	Number	Quartets
Numerals	9	1, 2, 3, 4, 5, 6, 7, 8 and 9
Letters	7	B, C, E, F, N, S and W
Circles	8	2, 3, 4, 5, 6, 7, 8 and 9
Bamboo	8	2, 3, 4, 5, 6, 7, 8 and 9
Miscellaneous	4	Woman, platter, plant and bird
Total	36	

We will revisit the family size population in the Practice Problems and Homework.

17.1.2 A Population of Trials

In Example 12.1, I introduced you to my friend Bert and his 100 games of online solitaire mahjong. Below I will show you data from another friend of mine, Walt, who plays a different version of online solitaire mahjong. Walt's version is definitely more difficult than Bert's; thus, I am not particularly interested in comparing their performances.

In my presentation of Bert's data, I made the outcome of a game a dichotomy: win or lose. I now want to make the outcome of Walt's game a count. To do this, I need to tell you a bit about online solitaire mahjong. First, for ease of presentation, I will say simply mahjong instead of online solitaire mahjong. Feel free to ignore the description below and jump ahead to the beginning of Example 17.3.

If you are interested in learning more about mahjong, you can play Walt's version at

`http://freegames.ws/games/boardgames/mahjong/freemahjong.htm`.

Mahjong begins with 144 tiles, consisting of 36 quartets. The quartets are described in Table 17.5. The 144 tiles are arranged in three dimensions. The player studies the arrangement and clicks on a pair of tiles of the same quartet type. (There are rules governing which tiles one may click on; again, if you are interested, play it a few times.) For example, the player might click on two birds. The tiles that have been clicked disappear, leaving two fewer tiles in the arrangement. The game ends in a victory if all tiles—72 pairs—are removed and ends in a loss if no legal moves remain. (A nice feature of the game is it tells you when no moves are available.) Let T denote the number of tiles remaining in the arrangement when the game ends. If $T = 0$ the game ends with a victory; if $T > 0$ the game ends with a loss. I define the response to be $X = T/2$, the number of pairs of tiles remaining when the game ends. Thus, the possible values of X are the integers between 0 and 72, inclusive.

Example 17.3 (Walt playing mahjong.) *Walt played 250 games of mahjong. A score of $x = 0$ means he won the game; a score of $x > 0$ means the game ended with x pairs remaining. The smaller the value of x, the better Walt performed.*

Table 17.6: Data from Walt's 250 games of mahjong. The value is the number of pairs remaining when the game ended. A value of 0 means that Walt won the game.

Value	Freq.	Rel. Freq.	Value	Freq.	Rel. Freq.	Value	Freq.	Rel. Freq.	Value	Freq.	Rel. Freq.
0	34	0.136	13	5	0.020	25	8	0.032	37	5	0.020
1	5	0.020	14	7	0.028	26	6	0.024	39	4	0.016
2	3	0.012	15	5	0.020	27	9	0.036	40	1	0.004
4	1	0.004	16	8	0.032	28	6	0.024	41	2	0.008
5	3	0.012	17	3	0.012	29	5	0.020	42	3	0.012
6	1	0.004	18	6	0.024	30	4	0.016	44	2	0.008
7	4	0.016	19	6	0.024	31	13	0.052	45	1	0.004
8	4	0.016	20	8	0.032	32	4	0.016	47	2	0.008
9	1	0.004	21	7	0.028	33	4	0.016	55	1	0.004
10	7	0.028	22	8	0.032	34	6	0.024	57	1	0.004
11	3	0.012	23	14	0.056	35	4	0.016	Total	250	1.000
12	5	0.020	24	8	0.032	36	3	0.012			

We assume that Walt's games are the outcomes of i.i.d. trials with probabilities given by an unknown probability histogram. Table 17.6 presents Walt's data.

It is **not easy** to learn from this table! Notice that the data range from a low of 0 to a high of 57, with a number of gaps; in particular, none of Walt's games ended with x equal to: 3, 38, 43, 46, 48–54, 56 or 58–72. I won't draw the density histogram of Walt's data; i.e., the picture-estimate of the unknown probability histogram. I will note, however, that the histogram is strongly skewed to the right with a tall peak above 0 and a number of minor peaks. Suffice to say, looking at the data in Table 17.6, I conclude that 250 observations are not enough to obtain a good picture-estimate of the population.

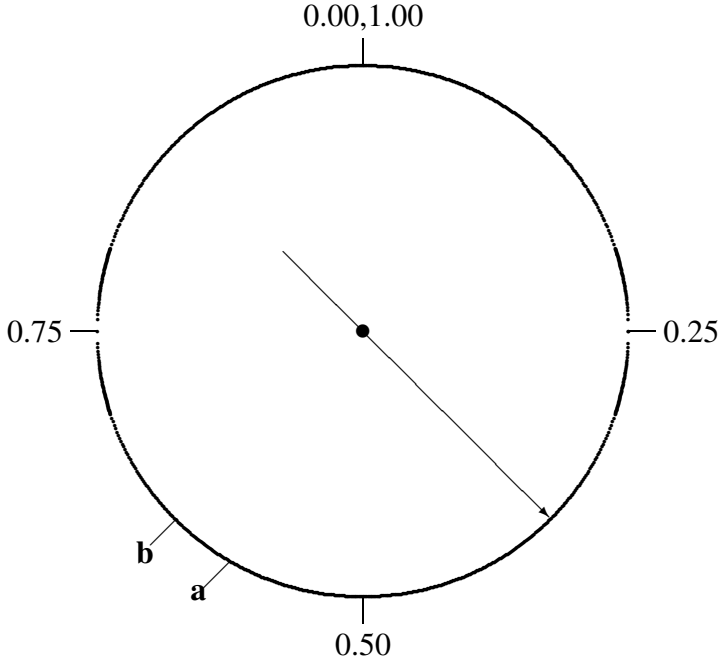
For future reference, I will note that for these 250 observations, the mean is $\bar{x} = 19.916$, the standard deviation is $s = 12.739$ and the median is $\tilde{x} = 21$. (This is another example of the surprising result that the mean is *smaller* than the median even though the data are strongly skewed to the right.)

Later in these *Course Notes* I will look at the 216 observations that remain after deleting the 34 games Walt won. (My reason for doing this will be explained at that time.) For future reference, I will note that for these remaining 216 observations, the mean is $\bar{x} = 23.051$, the standard deviation is $s = 10.739$ and the median is $\tilde{x} = 23$. Also, the distribution of these remaining 216 observations is close to symmetric; thus, it is no surprise that the mean and median nearly coincide.

17.2 Responses Obtained by Measuring

Under the entry, *Measurement*, Wikipedia lists seven basic measurement quantities:

Figure 17.4: The balanced spinner.



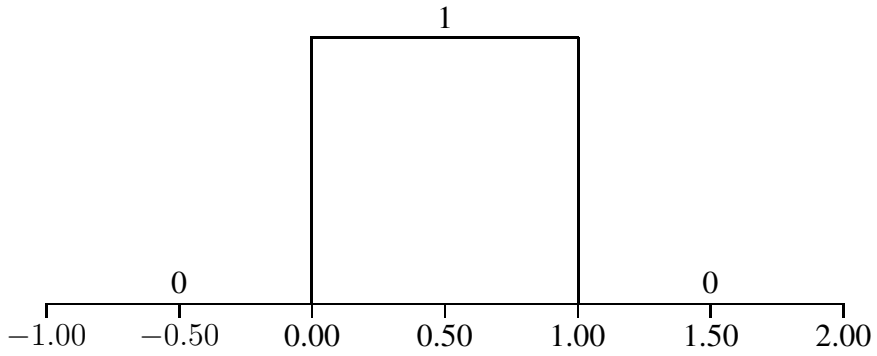
time, length, mass, temperature, electric current, amount of substance and luminous intensity.

All of the examples in these *Course Notes* are limited to the first four of these and a select few mathematical functions of one or more of these; for example, area, volume and velocity. Also, I tend to prefer weight and speed to their closely related mass and velocity.

If you reflect on our development of probability in these *Course Notes*, you will realize that almost everything we have done has grown from the notion of the equally likely case. For a CRD, we assume that all possible assignments are equally likely to occur. For a finite population, we assume that all cards in the population box are equally likely to be selected. For the i.i.d. trials we have studied, the outcome of a trial could be viewed as selecting a card from a box. The only exception is the Poisson Process, which is derived from a mathematical model for randomness in time. But even the Poisson distribution is tied to the equally likely case in the sense that it approximates the binomial.

In order to develop the mathematical theory we use for a numerical response, we need a new basic chance mechanism; one that, indeed, is similar to the equally likely case. This new chance mechanism is the **(balanced) spinner model**. I will usually suppress the adjective *balanced* because, in these notes, we will not consider any unbalanced spinner models. The spinner model is a refinement of the balanced roulette wheel we discussed earlier in these notes. Figure 17.4 presents a circle with circumference equal to 1. If you think of this as a pre-digital-clock clock face, put the numbers: 0.00 at the 12:00, 0.25 at the 3:00, 0.50 at the 6:00, 0.75 at the 9:00 and 1.00 at the 12:00, as I have done in this figure. Ignore the fact, for now, that 0.00 and 1.00 share the same point on this circle. The figure also contains a spinner with its arrow pointing at what appears to be at or near 0.375. The modifier *balanced* in the *balanced spinner model* reflects the assumption that

Figure 17.5: The uniform or rectangular pdf on the interval [0,1].



if one flicks the spinner with adequate force, then the arrow shows no preference in its stopping point.

One needs to be careful with this notion of *showing no preference*. We **cannot** say that every stopping point—every number between 0 and 1—is equally likely to occur, because there are an infinite number of stopping points. If one allows an infinite number of equally likely outcomes, then all of the results of probability theory that we use will collapse. (Trust me on this.) Instead, somebody figured out a clever way to look at this. The idea is that instead of assigning probabilities to points on the circle, we assign probabilities to arcs of the circle. The rule of the spinner model is:

> The probability that the arrow lands in an arc is equal to the length of the arc.

For example, in Figure 17.4 find the arc from point a, moving clockwise, to point b. By the rule, the probability that the arrow lands within this arc is equal to $b - a$, which, from the figure, appears to be a bit smaller than 0.05.

The balanced spinner model states that successive operations of it yield i.i.d. random variables with probabilities given by the rule above. The spinner model is important because it gives us a physical device to think about for measurements.

Let X be a random variable whose observed value is obtained by an operation of the spinner model; i.e., by flicking the spinner. I will now show you how to use a mathematical function to calculate probabilities for X. Figure 17.5 presents the graph of a very important function in Statistics. It is called the **uniform probability density function on the interval** $[0, 1]$. It is an example of a very important class of functions that are called **probability density functions**, abbreviated pdf for the singular and pdfs for the plural, pronounced simply as pea-dee-eff(s). Let me point out some features of this function and its graph.

1. If we denote the function by $f(x)$, we see that $f(x) = 0$ if $x < 0$ or if $x > 1$. This means that the graph of $f(x)$ coincides with the horizontal axis for $x < 0$ and for $x > 1$; I emphasize

this feature in the figure by typing a '0' to represent the height of the graph. In future graphs, I won't bother with the '0' anymore.

2. The function $f(x)$ equals 1 for all $0 \leq x \leq 1$, as noted in the figure.

3. The total area under the pdf (and above the horizontal axis, which is always implied) equals 1.

This pdf is called the uniform pdf because—except for where it equals zero—it has a uniform height. It is sometimes called the rectangular pdf because its graph—again ignoring where it equals zero—has the shape of a rectangle. It is my impression that uniform is the more popular of the two names.

Remember that just before I first referred you to Figure 17.5, I defined the random variable X whose observed value is obtained by an operation of the spinner model. Statisticians say that **probabilities for X are given by the uniform pdf on** $[0, 1]$. Let me show you why.

Suppose that we have any two numbers a and b that satisfy $0 \leq a < b \leq 1$. We are interested in the probability of the event $(a \leq X \leq b)$. From the spinner model, we know that the probability of this event equals the length of the arc that begins at a and moves clockwise to b; i.e., the probability equals $(b - a)$. **We can also obtain this answer from the pdf.** We simply calculate the area under the pdf between the numbers a and b. This is the area of a rectangle with base equal to $(b - a)$ and height equal to 1; thus, it equals

$$(b - a) \times 1 = b - a.$$

In other words, we obtain probabilities for the spinner model by calculating areas under the uniform pdf.

If this were a course for math majors, I would show you that the above demonstration for the event $(a \leq X \leq b)$ can be extended to all possible events, but this is too tedious for my purposes for this course.

17.2.1 The General Definition of a Probability Density Function

I have shown you one extended example of a measurement random variable. For the random variable X with observed value given by one operation of the spinner model, we have found that X has a pdf which will yield its probabilities by computing areas. The general result is that if we have a chance mechanism that yields a measurement random variable X, then it will have its own pdf, which will not necessarily be the uniform pdf on the interval $[0, 1]$. The pdf for X is given in the following definition.

Definition 17.1 *The pdf of a measurement random variable X is a function which satisfies the following equation for every pair of real numbers a and b with $a \leq b$.*

$$P(a \leq X \leq b) \textit{ equals the area under the graph of the pdf between the numbers } a \textit{ and } b. \quad (17.1)$$

The above definition has two important consequences.

1. The total area under X's pdf equals one because the total probability is one.

2. The graph of a pdf may not fall below the horizontal axis because, if it did, there would be some negative areas and probabilities cannot be negative.

Now we get to a subtle issue. I have been looking at this situation from the perspective of a scientist. I have a chance mechanism that arises in a scientific problem and it yields a measurement random variable X. I next try to find X's pdf. (We had a successful quest for the spinner model; can we be successful for other situations?) For example, think back to Sara's study of golf, introduced in Chapter 1. Consider Sara's trials with the 3-Wood; if we assume that these are i.i.d. trials, what is the pdf?

Here is a surprise. At this time, we are not going to try to answer this question for Sara. Instead, we switch perspective to that of a mathematician. Rather than try to find the pdf for a particular X, **we simply study functions that could be pdfs**. As a mathematician, I am interested in properties of functions that could be pdfs; whether there is a scientific application for a particular function does not concern me.

(**Aside:** The following link will take you to what passes for humor in this area:

`http://www-users.cs.york.ac.uk/susan/joke/3.htm#real`.)

The above ideas lead to the following definition.

Definition 17.2 *A function f* **could** *be a pdf for some measurement random variable if it satisfies the following conditions.*

1. *The value of $f(x)$ is nonnegative for all real numbers x.*
2. *The total area under the function $f(x)$ equals 1.*

We get the following result, which is quite useful and comforting.

Result 17.1 *Let a and b be any numbers with $a < b$. Let X be a measurement random variable; hence, it has a pdf. Then the following is true:*

$$P(a \leq X \leq b) = P(a \leq X < b) = P(a < X \leq b) = P(a < X < b).$$

In words, this result tells us that for a measurement random variable X, the probability of that X falls in an interval of numbers does not depend on whether the interval includes either or both of its endpoints. (We know that this is **not** true for a count random variable.) This result is actually very easy to prove; thus, I will prove it for you.

We can write

$$P(a \leq X \leq b) = P(a \leq X < b) + P(X = b).$$

Thus, the result will follow when I prove that $P(X = b) = 0$. We write,

$$P(X = b) = P(b \leq X \leq b).$$

By Definition 17.1, the probability of the latter term is the area under the pdf between b and b. But this is simply the area of a line; thus, it equals zero.

As a mathematician, I want to study functions that could be pdfs; i.e., that satisfy Definition 17.2. Not surprisingly, it is most efficient if I can develop properties for **families of pdfs**.

Let's pause for a moment. You might be thinking,

Families of pdfs? I barely know what a pdf is!

Actually, you have already met the most famous family of pdfs. Dating back to Chapter 7, we have found several uses for the family of Normal curves; namely, these uses involved using a particular Normal curve to obtain an approximate probability. Notice that a Normal curve **could** be a pdf; its total area is one and it is never negative. Indeed, the family of Normal curves is usually called the family of Normal pdfs.

During the 20th century, many clever mathematical statisticians exerted a great deal of time and effort to obtain many wonderful features of the family of Normal pdfs, some of which you will learn in the remainder of these *Course Notes*. Because of the plethora of useful results for Normal pdfs, it is predictable—recall Maslow's hammer—that scientists and statisticians love to state that the pdf of a scientifically important random variable X is a Normal curve. In many cases, a Normal pdf may be a good **approximation** to the true and unknown pdf, but, sadly, in many cases it is not. My advice is that when you come across such a claim, be skeptical; think about whether it makes sense scientifically. For example, if your response is the age for a population of college students, you *know* that the pdf will be strongly skewed to the right and, hence, it will not be well approximated by a Normal curve.

17.2.2 Families of Probability Density Functions

Almost always in these notes, when faced with a measurement random variable I will simply assume that it has an unknown pdf. On occasion, however, it will be useful to at least entertain the possibility that the unknown pdf is the member of some family. There are four families that we will consider; they are listed below, with comments.

1. The family of **Normal pdfs**. You saw the graph of a Normal pdf in Figure 7.4 on page 151. If you believe that the unknown pdf is exactly or approximately symmetric, you might entertain the possibility that it is a Normal curve.

2. The family of **Laplace pdfs**; also called the family of double exponential pdfs. Graphs of some Laplace pdfs are given at:

   `http://en.wikipedia.org/wiki/Laplace_distribution`.

   (Note: Unless you really love this material, do not attempt to read the entire passage. I simply want you to look at the graphs of select Laplace pdfs at the top of the page on the right. Similar comments apply to the sites below.) If you believe that the unknown pdf is exactly or approximately symmetric, you might entertain the possibility that it is a Laplace pdf. In other words, a Laplace pdf is a natural competitor to a Normal pdf.

3. The family of **exponential pdfs**. Graphs of some exponential pdfs are given at:

   `http://en.wikipedia.org/wiki/Exponential_distribution`.

   An exponential pdf is skewed to the right and its most important feature is its connection to the Poisson Process; this connection is explored in a Practice Problem in this chapter.

4. The family of log-normal pdfs. Graphs of some log-normal pdfs are given at:

   `http://en.wikipedia.org/wiki/Log-normal_distribution`.

   If the random variable of interest takes on positive values only—i.e., zero or negative measurements are not allowed—then you might consider using a log-normal pdf. The log-normal is a rich family, including pdfs that are nearly symmetric as well as those that are strongly skewed to the right.

   Why the name log-normal? Well, if the random variable X has a log-normal pdf, then the random variable Y which equals the natural log of X has a normal pdf. (This is why X must be strictly positive; one cannot take the log of zero or a negative number.) Reversing this, if Y has a Normal pdf, and we let X be equal to $\exp(Y)$, then X has the log-normal distribution. Bonus points if you have spotted that the name is dumb; it should be called the exponential-normal distribution, but the inaccurate name log-normal has persisted.

The family of Weibull pdfs is a competitor to the family of log-normal pdfs, but we won't consider it in these *Course Notes*.

17.3 Estimation of μ

As I mentioned earlier, for a count response, the population—a probability histogram—can be estimated by the density histogram of the data. Sadly, estimation with confidence is not possible. The story is similar for a measurement response. A density histogram of the data can be used to estimate the pdf. (Grouping values, as we did in Chapter 2 is allowed.) If you want a smoother estimate, you can employ a kernel estimate, as discussed in Chapter 2. Again sadly, estimation with confidence is not possible.

Even though we cannot estimate with confidence, I believe that it is always a good idea to use the data to obtain a picture-estimate of the population picture. In addition, often a scientist wants to estimate the mean, μ, of the population, for either counts or measurements. The estimation of the mean is the topic of this section. As you will learn, there are some really useful results on estimating the mean of a population.

Statisticians, as a group, have been criticized for being a bit too enthusiastic in their interest in the mean. I remember seeing a textbook in which the author stated, more or less, the following:

> Tom owns a store and is interested in the income of his customers.

So far, this is fine, but then the author, without any explanation, proceeded to say:

> Let μ be the mean income of his population of customers.

Huh? How did *interest* translate into the mean? Before I introduce the technical results, I want to share three brief stories; the first two are in the category of *jokes* and the third is from real life.

1. When I first studied Statistics in the late 1960s, here is the first joke I heard about my area of interest:

> A statistician is a person who, while standing with one foot in a vat of boiling water and one foot in a vat of ice water, will say, "On average, I feel fine!"

2. I told the story of the statistician and the water every semester for many years until some undergraduate math majors provided me with an updated version which takes into account the popularity of bow hunting in Wisconsin. Their story went as follows:

   > Three statisticians are bow hunting. They locate a deer and take aim. The first statistician fires and his arrow lands 40 feet to the left of the deer. The second statistician fires and her arrow lands 40 feet to the right of the deer. The third statistician jumps up and down shouting, "We hit it!"

3. Quoting from

   > `http://en.wikipedia.org/wiki/Aloha_Airlines_Flight_243`:
   > On April 28, 1988, a Boeing 737-297 ... suffered extensive damage after an explosive decompression in flight, but was able to land safely at Kahului Airport on Maui.
   > The safe landing of the aircraft despite the substantial damage inflicted by the decompression established Aloha Airlines Flight 243 as a significant event in the history of aviation, with far-reaching effects on aviation safety policies and procedures. ...
   > ... the United States National Transportation Safety Board (NTSB) concluded that the accident was caused by metal fatigue exacerbated by crevice corrosion. The plane was 19 years old and operated in a coastal environment, with exposure to salt and humidity.

   I cannot overemphasize what a big deal this was in America. It inspired a 1990 television movie, *Miracle Landing*. In real life, there was an anxiety among fliers concerning metal fatigue and, especially, flying in an older airplane.

   So, what does this have to do with Statistics? Reportedly, an airline sent a memo to flight attendants (see [1]), which stated:

   > To avoid increasing a potentially high level of customer anxiety, please use the following responses when queried by customers. Question: How old is this aircraft? Answer: I'm unaware of the age of this particular aircraft. However, the average age of our aircraft is 13.5 years.

   The airline proposed responding to a question about a *particular airplane* with an answer about their *fleet of airplanes*. Is this really very different from my earlier reference to what the textbook author said about the merchant Tom?

Remember: Science trumps (is more important than) Statistics. For an extremely depressing example, consider the distribution of the sizes (mass or diameter) of all meteors/asteroids that enter Earth's atmosphere during the next 100 years. I don't care about the mean or median of the distribution; it's the maximum that we need to worry about!

17.3.1 The Assumptions

Recall that we plan to observe random variables

$$X_1, X_2, X_3, \ldots X_n.$$

We assume that these are i.i.d. random variables from an unknown population. Our goal is to estimate the mean of the population, denoted by μ. Also, for later use, let σ denote the unknown standard deviation of the population. We will be interested in the following summaries—themselves random variables too—of the random variables $X_1, X_2, X_3, \ldots X_n$:

$$\bar{X} \text{ and } S, \text{ the mean and standard deviation of the random variables.}$$

The observed values of these random variables are $\bar{x}$ and s.

The obvious point estimator [estimate] of μ is $\bar{X}$ [$\bar{x}$]. We now face a new problem. When we studied the binomial or the Poisson, we could calculate the exact sampling distribution of our point estimator. This allowed us to obtain exact—actually conservative—confidence intervals for the parameter of interest, either p or θ. In addition, we could use a Normal curve and Slutsky's Theorem to obtain approximate confidence intervals for either p or θ and we had a pretty good idea when the approximation was good. The problem of estimating μ is much more difficult.

It is more difficult because we are studying a much more general problem. Above I state that our random variables come from *an unknown population*, without specifying any features of the population. The exact sampling distribution of $\bar{X}$ will vary from population to population and—with one notable exception, which is discussed later—is a mess to derive. Thus, we will work almost exclusively—I would say exclusively, but some statisticians like to split hairs on this issue—on finding approximate answers, with the added annoyance that it is very difficult to tell whether the approximation is good.

The mathematical theory needed for our results begins with a famous result, called the **Central Limit Theorem**. See

`http://en.wikipedia.org/wiki/Central_limit_theorem`

if you are interested in details of its history. Let's examine this three word name. *Theorem*, of course, means it is an important mathematical fact. *Limit* means that the truth of the theorem is achieved only as n grows without bound. In other words, for any finite value of n the result of the theorem is only an approximation. This should remind you of our earlier presentation of the Law of Large Numbers, which is also a limit result. *Central*, finally, refer to its importance; i.e., it is central to all of probability theory.

There are two parts to the Central Limit Theorem. The first part tells us how to standardize $\bar{X}$ and is given in the equation below. If we define Z by:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \tag{17.2}$$

then Z is the standardized version of $\bar{X}$. The second part of the Central Limit Theorem addresses the following issue. I want to calculate probabilities for Z, but I don't know how to do this.

I decide to use the N(0,1) curve to obtain approximate probabilities for Z. The Central Limit Theorem shows that in the limit, as n grows without bound, any such approximate probability will converge to the true (unknown) probability. This is a fairly amazing result! It does not matter what the original population looks like; in the limit, the N(0,1) curve gives correct probabilities for Z.

Of course, in practice, it is never the case that 'n grows without bound;' we have a fixed n and we use it. But if n is *large* we can **hope** that the N(0,1) curve gives good approximations. Ideally, this **hope** will be verified with some computer simulations and we will do this on several occasions.

If we expand Z in Equation 17.2, we get the following result, that, while scientifically useless, will guide us to better things.

Result 17.2 (The approximate confidence interval for μ when σ is known.) *As usual, the target confidence level determines the value of z^*, as given in Table 12.1 on page 296. The confidence interval estimate of μ is*

$$\bar{x} \pm z^*(\sigma/\sqrt{n}). \tag{17.3}$$

Recall the cat population, introduced in Example 17.1 with probability histogram given in Figure 17.1. Recall that Nature knows $\mu = 1.40$ and $\sigma = 0.80$. Imagine a researcher who does not know μ, but somehow knows that $\sigma = 0.80$. Scientifically, this is ridiculous, but many mathematical statisticians seem to love the television show *Jeopardy*: any answer—Equation 17.3—must have a question!

Thus, please humor me while I illustrate some computations. In my role as Nature, I put the cat population into my computer. Then, switching gears, I become the researcher who knows that $\sigma = 0.80$, but does not know the value of μ; again, a neat trick, given that the value of μ is needed to compute σ! Anyways, as the researcher I select a random sample of $n = 25$ households (smart or dumb doesn't matter) and obtain the following data:

1	1	1	1	2	0	1	0	1	3	2	0	0
2	1	1	3	1	1	3	1	3	2	1	2	

For 95% confidence, $z^* = 1.96$. Thus, the 95% confidence interval estimate of μ, given $\sigma = 0.8$ and $n = 25$, is:

$$\bar{x} \pm 1.96(0.80/\sqrt{25}) = \bar{x} \pm 0.314.$$

For the 25 observations above, you may verify that $\bar{x} = 1.36$; also, although we don't need it for the confidence interval, the standard deviation of these 25 numbers is $s = 0.952$

Nature can see that the point estimate, 1.36, is almost correct because $\mu = 1.40$. Both Nature and the researcher can see that the standard deviation of the data is quite a bit larger—0.952 is 19% larger than 0.80—than the standard deviation of the population. The 95% confidence interval for μ is:

$$\bar{x} \pm 0.314 = 1.36 \pm 0.314 = [1.046, 1.674].$$

Nature can see that this particular confidence interval is correct because it includes $\mu = 1.40$; the researcher, ignorant of the value of μ, would not know this fact.

By looking at one random sample of size $n = 25$, we have had the experience of evaluating the formula, but we don't know how it *performs*. The Central Limit Theorem approximation, which gives us the 95%, is accurate as n grows without bound, but is it any good for $n = 25$?

To answer this last question, I repeated the above activity another 9,999 times. To be precise, I had Minitab generate a total of 10,000 random samples of size $n = 25$. For each generated random sample I calculated the interval

$$\bar{x} \pm 0.314.$$

I obtained the following results:

- A total of 286 simulated confidence intervals were too small. Recall that this means that the upper bound of the interval, u, is smaller than $\mu = 1.40$.

- A total of 318 simulated confidence intervals were too large. Recall that this means that the lower bound of the interval, l, is larger than $\mu = 1.40$.

- A total of $286 + 318 = 604$ confidence intervals were incorrect. This is larger than the target of 500. If we calculate the nearly certain interval for the probability of an incorrect 95% confidence interval, we get:

  $$0.0604 \pm 3\sqrt{(0.0604)(0.9396)/10000} = 0.0604 \pm 0.0071.$$

  Thus, the Central Limit Theorem approximation, while not horrible, is clearly not exact; The true probability of an incorrect confidence interval is definitely larger than 0.0500.

I repeated the above simulation study on the cat population, but took $n = 100$. I obtained the following results:

- A total of 253 simulated confidence intervals were too small.

- A total of 260 simulated confidence intervals were too large.

- A total of $253 + 260 = 513$ confidence intervals were incorrect. This is larger than the target of 500, but well within the bounds of sampling error. Thus, it appears that the Central Limit Theorem approximation is quite good for $n = 100$.

In science, σ is always unknown. We need a way to deal with this. The obvious idea works; replace the unknown σ with the s computed from the data. More precisely, define Z' as follows.

$$Z' = \frac{\bar{X} - \mu}{S/\sqrt{n}} \tag{17.4}$$

According to the work of Slutsky, the Central Limit Theorem conclusion for Z is also true for Z'; i.e., we can use the N(0,1) curve for Z' too. This leads to our second confidence interval formula, which we will call **Slutsky's confidence interval estimate of** μ:

Table 17.7: Results from three simulation experiments. Each simulation had 10,000 reps, with a rep consisting of a random sample of size n from the **cat population** in Table 17.1. For each sample, **Slutsky's** approximate 95% confidence interval estimate of μ, Formula 17.5, is computed and Nature classifies it as too small, too large or correct.

Sample size (n)	Number of Too Small Intervals	Number of Too Large Intervals	Number of Incorrect Intervals
10	451	371	822
20	377	286	663
40	363	245	608

Result 17.3 (Slutsky's approximate confidence interval estimate of μ.) *As usual, the target confidence level determines the value of z^*, as given in Table 12.1 on page 296. Slutsky's approximate confidence interval estimate of μ is:*

$$\bar{x} \pm z^*(s/\sqrt{n}). \tag{17.5}$$

Playing the role of Nature, I put the cat population into my computer. Then, switching roles to the researcher, I selected a random sample of size $n = 10$. I obtained the following data:

$$2, 1, 2, 2, 1, 2, 1, 1, 3, 2.$$

These data yield $\bar{x} = 1.700$ and $s = 0.675$. Thus, for these data, Slutsky's 95% confidence interval estimate of μ is

$$1.700 \pm 1.96(0.675/\sqrt{10}) = 1.700 \pm 0.418 = [1.282, 2.118].$$

Reverting to Nature, I note that, in fact, $\mu = 1.40$. Thus, the point estimate is incorrect, but the 95% confidence interval is correct.

I performed three simulation experiments for the cat population, each with 10,000 reps, to investigate the performance of Slutsky's approximate 95% confidence interval estimate of μ. The first simulation experiment was for random samples of size $n = 10$; the second was for $n = 20$; and the third was for $n = 40$. My results are presented in Table 17.7. Take a minute to look at the numbers in this table; below is their most important feature.

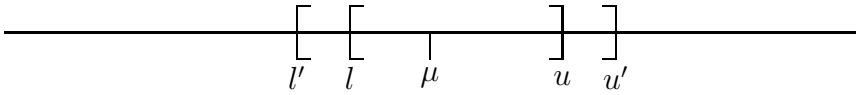
- For each sample size, there are too many incorrect intervals. Note, however, that as n increases, the number of incorrect intervals declines. This is in agreement with the math theory which states that as n grows without bound, the 95% confidence level becomes accurate.

Imagine that you are a mathematical statistician thinking about the results in Table 17.7. The only formula you have is Slutsky's and it is yielding too many incorrect intervals! What can you do? Let's look at the confidence interval formula again:

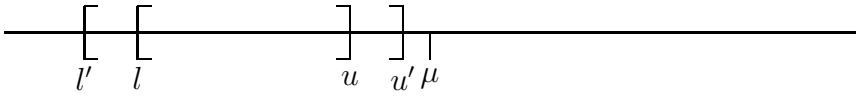
$$\bar{x} \pm z^*(s/\sqrt{n}).$$

Figure 17.6: Some possible effects of making Slutsky's confidence intervals wider. Slutsky's interval is $[l, u]$; its wider version is $[l', u']$.

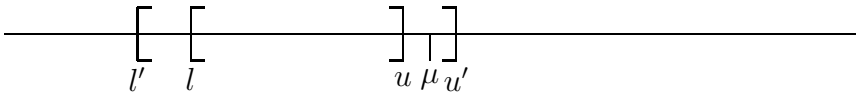
Both intervals are correct:



Both intervals are too small:



Slutsky's interval is too small, but its wider version is correct:



The values of $\bar{x}$, s and n come from the data; you can't change them. All that you can possibly change is the number z^*. You realize that if you replace z^* by a larger number, the interval will become wider. Think back to any one of my simulation studies that are reported in Table 17.7. If I make every one of the 10,000 intervals wider, this is what will happen:

1. Intervals that **had been correct** will remain correct; because the center ($\bar{x}$) does not change, the new interval will include the original interval, which includes μ. See Figure 17.6.

2. Some of the intervals that **had been incorrect** will remain incorrect. See Figure 17.6 for a picture of a too small interval that remains too small.

3. Intervals that **had been incorrect might become correct**. Figure 17.6 illustrates this possibility for a too small confidence interval.

So, how much wider should we make Slutsky's intervals? We will replace z^* by a larger number that we will denote by t^*. Table 17.7 suggests that we will need a bigger correction for $n = 10$ than we will need for $n = 40$. Also, for n sufficiently large, we won't need any correction. In other words, the value of t^* should depend on the sample size n. Actually, statisticians prefer to say that t^* is a function of the number of degrees of freedom in the deviations, which is $(n - 1)$.

A solution to this difficulty—finding t^*—was obtained by William Sealy Gosset in 1908, who published his findings under the name Student. Gosset invented a new family of pdfs, called the t-curves or, sometimes, Student's t-curves. A specific t-curve is determined by one parameter, called its degrees of freedom (df). Possible values for the degrees of freedom are the positive

integers: 1, 2, 3, ... Every t-curve has its one peak at 0 and is symmetric. As the number of degrees of freedom increases, the t-curves have less spread; visually, the peak becomes taller and the symmetric tails shorter. As the degrees of freedom grow, the t-curves converge to the N(0,1) curve; thus, in an abuse of language, the N(0,1) curve is sometimes referred to the t-curve with $df = \infty$. Pictures of the t-curves for $df = 1, 2, 5$ and ∞ (i.e., the N(0,1) curve) are available at

`http://en.wikipedia.org/wiki/Student's_t-distribution`

Note the following. For degrees of freedom as small as 5, visually it is difficult to distinguish between a t-curve and the N(0,1) curve. For $df > 5$, even though we might not be able to **see** the differences between the two curves, the differences are important!

We will use the following website for calculating areas under a t-curve for you.

`http://stattrek.com/online-calculator/t-distribution.aspx`

(I used a different site during the Summer 2013 class, but it has exploded; well, the internet equivalent of exploding. Don't use it!) We are going to use this website quite a lot; thus, it will be convenient for me to present a facsimile of it in Table 17.8. In this facsimile, I have replaced the website's three boxes by three horizontal line segments because it is easier to create the latter with my word processor.

Above the three boxes in the website is the default option *t-score* for the first user specification: *Describe the random variable*. There is another option of which you will learn later; for now, **do not change the default**.

Next, you need to enter the number of degrees of freedom for the t-curve/distribution of interest. In my illustrative examples below I will use $df = 10$. Your next choice is to enter a value for *t score* or *Cumulative probability:* $P(T < t)$ **but not both**. If you:

- Opt for *t score*, you may enter any real number; negative, zero or positive. Alternatively, if you
- Opt for *Cumulative probability:* $P(T < t)$ then you must enter a number that is strictly between 0 and 1.

Let me do a few examples to illustrate the use of this site. Remember in each example below, I have entered $df = 10$.

1. I want to obtain the area under the t-curve with $df = 10$ to the left of -1.367.
   - I enter -1.367 in the box next to *t score* and click on the box *Calculate*.
   - My answer, 0.1008, appears in the *Cumulative probability:* $P(T < -1.367)$ box.

   Note that in order to remind me of what I am doing, in the *Cumulative probability:* box, the site replaces the generic t with the value I am using, -1.367.
2. I want to find the number, let's call it t, with the following property: the area under the t-curve with $df = 10$ to the left of t is equal to 0.05.

Table 17.8: Facsimile of t-curve calculator website.

Describe the random variable	t-score
Degrees of freedom	_____
t score	_____
Cumulative probability: $P(T < t)$	_____

- I enter 0.05 in the box next to *Cumulative probability:* $P(T < t)$ and click on the box *Calculate*.
- My answer, -1.812, appears in the *t-score* box.

(**Aside:** Do you need a break from this exciting presentation? If so go to

`http://www-users.cs.york.ac.uk/susan/joke/3.htm#boil`.

The mathematician's behavior in this joke is relevant to the next two examples. If you can't stand to take a break or don't have the time, carry on.)

3. I want to obtain the area under the t-curve with $df = 10$ to the right of 2.863. For this problem, the site is annoying! It only gives areas to the left. There are two ways to attack this problem; I will use the fact that the total area under any pdf equals 1. (The other way uses the fact that the t-curves are symmetric around 0; you don't need to know it.)

   Proceeding as in (1) above, I find that the area under the curve to the **left** of 2.863 is equal to 0.9916. As a result, the area I seek is $1 - 0.9916 = 0.0084$.

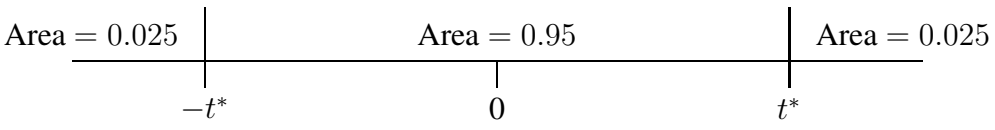
4. I want to find the number, let's call it t, with the following property: the area under the t-curve with $df = 10$ to the right of t is equal to 0.15. This problem is similar to the problem in (2) above, except that left has been changed to right.

   As in (3) above, I rewrite the problem. The number I seek, t, has the property that the area to the left of it equals $1 - 0.15 = 0.85$. Proceeding as in (2) above, I obtain the answer $t = +1.093$.

Now that you are familiar with how this site operates, I will use it to solve a very specific problem.

Recall that for the cat population and $n = 10$, Slutsky's confidence interval did not perform well. I am now going to show you Gosset's interval for these data. (I will give you the general method shortly.) I go to the t-curve website calculator and enter $n - 1 = 9$ for the degrees of freedom.

Figure 17.7: Finding the value of t^* for Gosset's 95% confidence interval estimate.



Our goal is to find the number t^* with the property:

The area under the t-curve between $-t^*$ and t^* is equal to 0.95.

Figure 17.7 illustrates this idea (except that I did not bother to draw the t-curve). We see that because of the symmetry of the t-curve, the areas to the left of $-t^*$ and to the right of t^* both equal 0.025. Because our website calculator handles only areas to the left, I change my problem slightly: My new goal is to find the number t^* with the property that the area under the t-curve to the left of it equals $0.95 + 0.025 = 0.975$. I enter 0.975 into the box *Cumulative probability* .... I click on calculate and obtain the answer 2.262. In other words, the number 2.262 plays the same role for the t-curve with $df = 9$ as 1.96 does for the N(0,1) curve. Thus, I replace $z^* = 1.96$ Slutsky's 95% confidence interval by $t^* = 2.262$, and obtain:

$$\bar{x} \pm 2.262(s/\sqrt{10}).$$

For $n = 20$, Gosset's 95% confidence interval is:

$$\bar{x} \pm 2.093(s/\sqrt{20}).$$

Finally, for $n = 40$, Gosset's 95% confidence interval is:

$$\bar{x} \pm 2.023(s/\sqrt{40}).$$

Note that as n increases from 10 to 20 to 40, the value of t^* for Gosset's interval becomes closer to the $z^* = 1.96$ for Slutsky's interval. This is because as n grows, Slutsky performs better and, thus, less correction is needed.

I simulated 10,000 of Gosset intervals for the cat population and each of the values of $n = 10$, 20 and 40; my results are in Table 17.9. Look at the numbers in this table for a few minutes and note the following.

1. For $n = 10$, Gosset's interval performs awesomely! (Bieberly?) The number of incorrect intervals, 510, almost matches the target number, 500. This discrepancy—between 510 and 500—is well within the bounds of chance.

2. Be careful! For $n = 20$, Gosset gives more incorrect intervals than it does for $n = 10$; does this mean its performance is declining as n grows? **No!** The difference between 510 and 540 is well within the random variation of a simulation study. If Gosset (or Slutsky) performs well for a particular value of n, it will also perform well for any larger value of n.

Table 17.9: Results from three simulation experiments. Each simulation had 10,000 reps, with a rep consisting of a random sample of size n from the **cat population** in Table 17.1. For each sample, **Gosset's** approximate 95% confidence interval estimate of μ is computed and Nature classifies it as too small, too large or correct.

Sample size (n)	Number of Too Small Intervals	Number of Too Large Intervals	Number of Incorrect Intervals
10	362	148	510
20	332	208	540
40	320	209	529

I will now give you Gosset's general approximate confidence interval for μ.

Result 17.4 (Gosset's approximate confidence interval estimate of μ.) *Gosset's confidence interval is:*

$$\bar{x} \pm t^*(s/\sqrt{n}). \tag{17.6}$$

The value of t^* depends on the sample size and the desired confidence level, as described below.

1. *Select the desired confidence level and write it as a decimal; e.g., 0.95 or 0.99.*
2. *Subtract the desired confidence level from one and call it the error rate. Divide the error rate by two and subtract the result from one; call the final answer c; e.g., 0.95 gives $c = 1 - 0.05/2 = 0.975$ and 0.99 gives $c = 1 - 0.01/2 = 0.995$.*
3. *Go the website*

   `http://stattrek.com/online-calculator/t-distribution.aspx`.

   Next, enter $n - 1$ for degrees of freedom; enter c in the Cumulative probability ... *box; and click on* Calculate. *The value t^* will appear in the* t-score *box.*

Take a moment and verify that you know how to obtain $t^* = 2.093$ for 95% confidence and $n = 20$.

17.3.2 Gosset or Slutsky?

If $n = 500$ and you want 95% confidence, you may use Slutsky with $z^* = 1.960$ or Gosset with $t^* = 1.965$. For $n =$ 1,000 and 95% confidence, Gosset's $t^* = 1.962$. In words, for n sufficiently large, the two confidence interval formulas are essentially identical. When n is small enough for z^* and t^* to differ substantially, I recommend using Gosset's t^*. In other words, for the problem of estimation of μ, you can forget about Slutsky's formula; I recommend that you always use Gosset.

Do not read too much into my endorsement of Gosset. All I am saying is that Gosset is preferable to Slutsky; I am not saying that Gosset always performs well. Let me give you an example in which Gosset performs poorly.

Figure 17.8: The log-normal pdf with parameters 5 and 1.

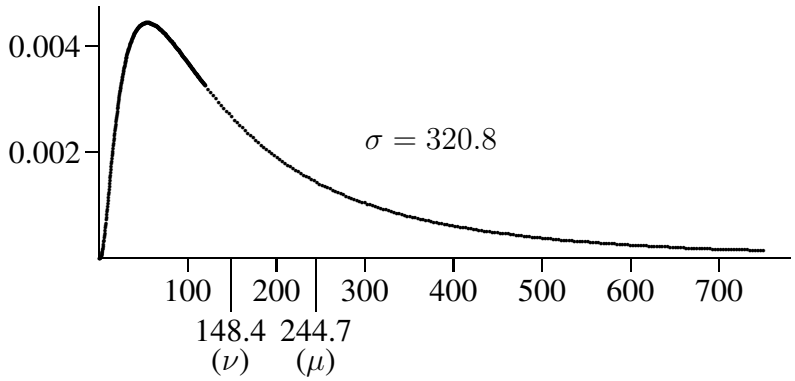


Figure 17.8 presents the log-normal pdf with parameters 5 and 1. Recall that this means if X has the log-normal pdf with parameters 5 and 1, then Y equal to the natural log of X has the N(5,1) curve for its pdf. We can see that this log-normal pdf is strongly skewed to the right. It can be shown that its mean is $\mu = 244.7$ and its standard deviation is $\sigma = 320.8$. (The value $\nu = 148.4$ is the **median of the pdf**; it divides the pdf's total area of one into two equal areas. We will discuss the estimation of ν in Chapter 18.)

Suppose that a researcher is interested in a measurement random variable X which takes on positive values. Nature knows that the pdf of X is given in Figure 17.8, but the researcher does not know this. Based on the scientific goals of the study, the researcher decides to estimate μ with 95% confidence. Thus, the researcher will use Gosset's interval to estimate μ with 95% confidence. I will now explore the performance of Gosset's interval for the population in Figure 17.8.

I performed eight simulation experiments, each with 10,000 reps. As in our earlier examples, each rep yielded Gosset's 95% confidence interval for a simulated random sample of size n from the pdf in Figure 17.8. Each simulated confidence interval is classified as: too small, too large or correct. The eight simulations are for different values of n; my results are presented in Table 17.10. Remember: Because I want 95% confidence, the target number of incorrect intervals is 500.

Let me make a few comments about this massive simulation study.

1. For $n = 10$ Gosset's interval performs very poorly. The nominal probability of obtaining an incorrect interval is 5%; the estimated probability is almost 16%.

2. As I kept doubling n in my simulations, the performance of Gosset improved, but very slowly. At $n = 640$, the error rate might be ok—the nearly certain interval, details not given, 0.0564 ± 0.0069, barely includes the target 0.0500—but it might be a bit large. For $n = 1{,}280$, finally, the error rate seems to be very close to 0.0500. Remember that the math theory says that in the limit as n goes to infinity the error rate will be 0.0500.

3. Note that, for all values of n in the table, and especially $n \leq 40$, the number of intervals that are too small is vastly larger than the number of intervals that are too large. We will explore

Table 17.10: Results from eight simulation experiments. Each simulation had 10,000 reps, with a rep consisting of a random sample of size n from the **log-normal** pdf in Figure 17.8. For each sample, **Gosset's** approximate 95% confidence interval estimate of $\mu = 244.7$ is computed and Nature classifies it as too small, too large or correct.

Sample size (n)	Number of Too Small Intervals	Number of Too Large Intervals	Number of Incorrect Intervals
10	1,582	13	1,595
20	1,355	12	1,367
40	1,122	22	1,144
80	823	44	867
160	674	54	728
320	533	90	623
640	440	124	564
1,280	376	136	512

this issue later.

Sometimes a fancy math argument allows us to use one simulation experiment for an entire family of populations. Such a family is the exponential. I performed three simulation experiments on an exponential pdf. As usual, each of the 10,000 reps: generated random sample from the exponential pdf; calculated Gosset's 95% confidence interval; checked to see how the interval compared to μ. My results are given in Table 17.11. The exponential distribution is skewed to the right and Gosset's interval does not work well for $n \leq 40$. At $n = 80$, the nearly certain interval for the true error rate includes the target, 0.0500, but just barely.

17.3.3 Population is a Normal Curve

I have told you that if the population is an exponential pdf, then one simulation experiment on confidence intervals covers every member of the family. The same is true for the family of Laplace pdfs, and I will give you my simulation results for it in a Practice Problem. The log-normal family is not so amenable; a different simulation experiment is needed for every member of the family.

The family of Normal pdfs also has the property that one simulation covers all curves. In fact, I performed a simulation experiment with 10,000 reps to study the performance of Gosset's 95% confidence interval for a Normal pdf population and $n = 5$. The results were:

- Two hundred twenty-seven of the simulated intervals were too large.
- Two hundred forty-five of the simulated intervals were too small.

Table 17.11: Results from eight simulation experiments. Each simulation had 10,000 reps, with a rep consisting of a random sample of size n from the **exponential** pdf with parameter $\lambda = 1$. For each sample, **Gosset's** approximate 95% confidence interval estimate of $\mu = 1/\lambda = 1/1 = 1.$ is computed and Nature classifies it as too small, too large or correct.

Sample size (n)	Number of Too Small Intervals	Number of Too Large Intervals	Number of Incorrect Intervals
10	928	29	957
20	738	45	783
40	584	96	680
80	472	93	565

- A total of $227 + 245 = 472$ of the simulated intervals were incorrect. This is a bit smaller than the target of 500, but well within the bounds of sampling variation.

Actually, I did not need to perform a simulation experiment for a Normal pdf because:

Result 17.5 *If the population is a Normal pdf, then the confidence level in Gosset's confidence interval is exact.*

This result is true because in Gosset's original work, he assumed a Normal pdf in order to derive the t-curves.

17.4 Lies, Damned Lies and Statistics Texts

Quoting from

`http://en.wikipedia.org/wiki/Lies,_damned_lies,_and_statistics`,

> "Lies, damned lies, and statistics" is a phrase describing the persuasive power of numbers, particularly the use of statistics to bolster weak arguments. It is also sometimes colloquially used to doubt statistics used to prove an opponent's point.
>
> The term was popularized in the United States by Mark Twain (among others), who attributed it to the 19th-century British Prime Minister Benjamin Disraeli (1804-1881): "There are three kinds of lies: lies, damned lies, and statistics." However, the phrase is not found in any of Disraeli's works and the earliest known appearances were years after his death. Other coiners have therefore been proposed, and the phrase is often attributed to Twain himself.

In the 1970s, a favorite television show of mine was *Kung Fu*. starring the late David Carradine. (You might know him as the title character Bill in two Quentin Tarantino films.) Carradine's character traveled the American West of the 19th-century; his purpose? As he put it,

I don't seek answers. I seek a better understanding of the questions.

I believe that this is a good goal for so-called higher education in general. This has been my (attempted) approach in dealing with the researcher's question:

When should I use Gosset's confidence interval formula?

If the population is a Normal pdf, then Gosset is exact. Thus, if Nature were to tell you that the population you are studying is a Normal pdf, then you can feel very comfortable using Gosset's formula. Unfortunately, Nature is not so forthcoming and—skeptics that we academics are—if you announce that Nature has told you this, we won't believe you.

In practice, a researcher does not know what the population is. Thus, **if you choose to use Gosset's formula, you must accept that there is a uncertainty about whether it will perform as advertised**. The **only** way to reduce that uncertainty is to be a **good scientist**; i.e., to be knowledgeable about the phenomenon under study. As the simulation studies have shown, as a scientist you need to be able to judge how much skewness there is in the population. If you anticipate little or no skewness, then you will anticipate that Gosset will perform as advertised. If, however, you anticipate a great deal of skewness in the population, then you should worry about using Gosset for small values of n, where—as you have seen—small is a function of the amount of skewness.

One of my firm beliefs is:

It is the scientist, not the statistician, who should assess the likely shape of the population.

If you understand human nature, especially the desire to be an expert, you will be unsurprised to learn that this belief of mine is unpopular with many statisticians; they believe that their expertise in Statistics makes them best equipped to make such decisions. One of their most popular statements is to, "Look at the data." I will address this issue shortly.

I suspect you would agree that the answer to the researcher's question:

When should I use Gosset's confidence interval formula?

that I have presented in these notes is **nuanced**. By contrast, the answer given in many introductory Statistics texts is quite definitive:

1. If $n \leq 30$, then you should use Gosset only if the population is a Normal pdf.
2. If $n > 30$, Slutsky always performs as advertised.

A word of explanation is needed. For degrees of freedom larger than 29, these texts make no distinction between t^* and z^*; thus, for them, for $n > 30$, Slutsky's and Gosset's intervals are identical. Also, how does a student know whether the population is a Normal pdf? Easy; the textbook author, the instructor or some other authority figure tells the student that the population is a Normal pdf! (As the writer Dave Barry frequently types, "I am not making this up!") Call me a radical, but I think that education—even in a math-related field—should consist of more than learning to obey, without question, an authority figure!

Also, note that I have conclusively demonstrated that the two-part advice found in these texts is untrue. For the cat population, which is not a Normal pdf, Gosset's confidence interval performed as advertised for n as small as 10. For the log-normal pdf we examined, for $n \leq 320$ Gosset's intervals performed poorly.

I have had many conversations with Statistics' educators in which I have asked them about this lie that textbooks tell students. Their response? They channel Col. Nathan R. Jessup, a character in the movie *A Few Good Men*, and shout—well, sometimes they don't raise their voices—**They can't handle the truth!** I have a much higher opinion of you.

`https://www.youtube.com/watch?v=futzi-bYW0E`

17.4.1 More on Skewness

As I mentioned above, if the population is a Normal curve, then Gosset's formula is exact. Actually, I can say a bit more than that. If the population is a Normal curve, then the probability that a Gosset confidence interval will be too small is the same as the probability that it will be too large. By contrast, for the family of exponential pdfs and our one example of a log-normal pdf, intervals that are too small vastly outnumber the intervals that are too large (see Tables 17.10 and 17.11). I will show you why.

I performed a simulation experiment with 100 reps. For each rep I generated a random sample of size $n = 20$ from the N(0,1) pdf. From each sample, I obtained two numbers: $\bar{x}$ and s. I have drawn a **scatterplot** of these 100 pairs of numbers in Figure 17.9. Each pair is represented by a un upper case 'Oh,' with s on the vertical axis and $\bar{x}$ on the horizontal axis. Note that I have deliberately not given you much information on the scales of the axes in this picture. On the horizontal axis I have marked $\mu = 0$ and you can see that the 100 values of $\bar{x}$ are placed (left-to-right) approximately symmetrically around 0. Similarly, on the vertical axis I have marked $\sigma = 1$. The important feature of this plot, however, is that there is no relationship between the values of $\bar{x}$ and s; knowing, for example, that $\bar{x} < \mu$ or that $\bar{x} > \mu$ tells us nothing about the value of s.

In fact, it's an important property of random samples from a Normal pdf that the value of $\bar{X}$ is statistically independent of the value of S.

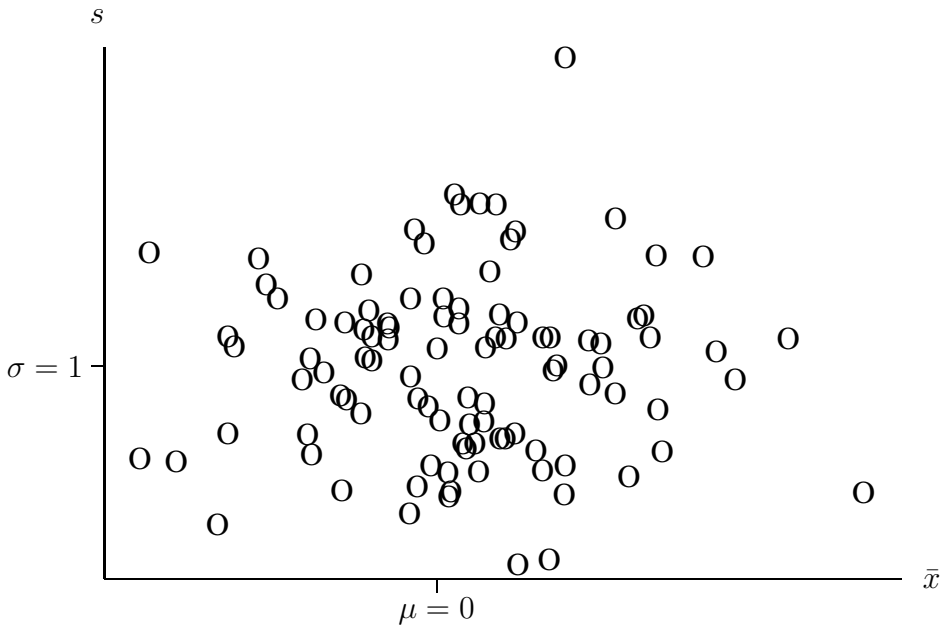
I will now contrast the above situation for Normal pdfs with samples from the log-normal pdf with parameters 5 and 1, pictured in Figure 17.8, which was the topic of a simulation experiment, with results presented in Table 17.10. Recall from this table that for $n = 20$, Gosset's 95% confidence interval performs very poorly, with 1,355 [12] simulated intervals that are too small [large].

Figure 17.10 presents a scatterplot of s versus $\bar{x}$ for 100 random samples of size $n = 20$ from the log-normal pdf in Figure 17.8. This scatterplot looks nothing like our previous one! In this scatterplot there is a very strong increasing relationship between the values $\bar{x}$ and s. In particular, if $\bar{x}$ is substantially smaller than $\mu = 244.7$, then s tends to be much smaller than $\sigma = 320.8$; and if $\bar{x}$ is substantially larger than μ, then s tends to be much larger than σ. I will now explain why this matters.

The half-width of Gosset's 95% confidence interval with $n = 20$ is

$$2.093s/\sqrt{20}.$$

Figure 17.9: The scatterplot of the standard deviation, s, versus the mean, $\bar{x}$, for 100 random samples of size $n = 20$ from a N(0,1) pdf. Note that there is no relationship between these two values. If you know about the correlation coefficient (see Chapter 21), it equals 0.021 for this picture.



Thus, if $\bar{x}$ is substantially smaller than $\mu = 244.7$, then, based on our scatterplot, this half-width will be very small. This means that the interval will be very narrow and, hence, many of these intervals will be incorrect by being too small. On the other hand, if $\bar{x}$ is substantially larger than μ, then the half-width will be very large and it will be unusual for the interval to be too large. If you question my reasoning, look at the simulation results: 1,355 intervals were too small, but only 12 were too large.

One last comment on this situation with the log-normal. A popular textbook claims that *if your data look normal, then Gosset's confidence interval performs as advertised*. Here I face a difficulty. In order for me to show your that this advice is wrong, I would need to spend time teaching you about *Normal quantile plots* and we simply do not have the time. If you take another Statistics course, you will likely learn that Normal quantile plots can be very useful, but not for the current situation. For more details, see

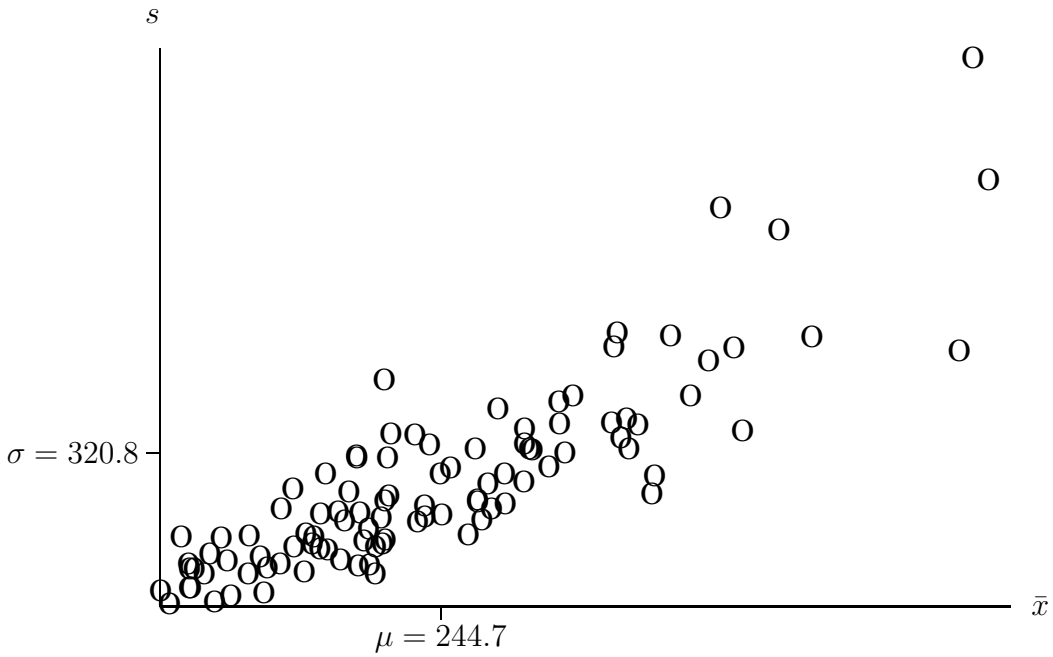
`http://www.stat.wisc.edu/~wardrop/papers/tr1008.pdf`.

Unfortunately, this paper is perhaps a bit advanced for the new student of Statistics.

17.5 Computing

In the examples of this chapter, I have given you n, $\bar{x}$ and s and you have learned how to find t^* from a website and calculate Gosset's confidence interval estimate of μ. In this short section, I will

Figure 17.10: The scatterplot of the standard deviation, s, versus the mean, $\bar{x}$, for 100 random samples of size $n = 20$ from a the log-normal pdf with parameters 5 and 1. Note that there is a strong increasing relationship between these two values. If you know about the correlation coefficient (see Chapter 21), it equals 0.862 for this picture.



show you what to do with raw data; i.e., the n numbers

$$x_1, x_2, x_3, \ldots, x_n.$$

For example, recall Reggie's $n = 15$ dart scores from a distance of 10 feet, reported in Chapter 1 and reproduced below:

181 184 189 197 198 198 200 200 205 205 206 210 215 215 220

Let's assume that these are 15 observations from i.i.d. trials. This is a count population and with such a small amount of data I cannot expect to obtain an accurate picture-estimate of the population. Instead, I will use Reggie's data to estimate the mean, μ, of the population with confidence.

You begin by going to a familiar website:

`http://vassarstats.net`.

The left-side of the page lists a number of options; click on *t-Tests & Procedures*. (You might remember doing exactly this in Chapter 1.) This takes you to a new set of options; click on the bottom one, *.95 and .99 Confidence Intervals for the Estimated Mean of a Population*. (Their use of language is poor—who am I to talk! We obtain confidence intervals for a mean, not an estimated mean.) Next, enter the data, by typing or pasting, and click on *calculate*.

The site gives lots of output, including:

$$N = 15;\ \text{mean } = 201.5333;\ \text{std. dev. } = 11.1986;\ df = 14;$$

$$\text{tcrit(.05) } = 2.14;\ \text{and tcrit(.01)} = 2.98.$$

As you may have guessed, in our language, these signify:

$$n = 15; \bar{x} = 201.5333; s = 11.1986; df = 14;\ \text{and the } t^* \text{ for } 95\%[99\%] \text{ is } 2.14[2.98].$$

I used Minitab to verify these values for $\bar{x}$ and s. I went to the website

`http://stattrek.com/online-calculator/t-distribution.aspx`

and obtained $t^* = 2.145$ for 95% and $t^* = 2.977$ for 99%.

The *vassarstats* website also reports:

- *Confidence Intervals for Estimated Mean of Population:*

  For .95 CI: 201.5333 ± 6.1878.

  For .99 CI: 201.5333 ± 8.6167.

I verified these two intervals as being correct. Thus, it appears that we can trust this website.

17.6 Summary

In Chapters 17 and 18 our data consist of i.i.d. random variables

$$X_1, X_2, X_3, \ldots, X_n,$$

from a numerical population. If the values of these random variables are obtained by counting, then the population is a probability histogram. If the values of these random variables are obtained by measuring, then the population is a probability density function (pdf). In words, in either case, the population is a picture which either consists of rectangles or is a smooth curve. The picture must be nonnegative with total area equal to one.

The population picture is used to calculate probabilities for the random variables. For counting, we focus on the random variable taking on **exactly** the number x. In particular, $P(X = x)$ equals the area of the rectangle centered at x; if there is no rectangle centered at x, then x is an impossible value for X and $P(X = x) = 0$.

For a measurement, we focus on the random variable **falling between two numbers**; i.e., lying in some interval. In particular, for any numbers a and b with $a \leq b$,

$$P(a \leq X \leq b) = \text{ the area under the pdf between the numbers } a \text{ and } b.$$

Remember that for any measurement random variable X and any number b, $P(X = b) = 0$. This implies—Result 17.1—that for any a and b with $a < b$.

$$P(a \leq X \leq b) = P(a \leq X < b) = P(a < X \leq b) = P(a < X < b).$$

Several families of pdfs are important in Statistics, including: the normal and Laplace families—both symmetric—and the exponential and log-normal families—both skewed to the right.

The population picture is known to Nature, but is unknown to the researcher. The population picture can be estimated by a density histogram of the data. In addition, the kernels introduced in Chapter 2 can serve as an estimate of a pdf. Sadly, it is impossible, in general, to estimate a picture with confidence. (For a family of pdfs, if one estimates all of the parameters with confidence, then one can obtain a confidence estimate of the pdf, but this topic is beyond the scope of this course.)

Researchers often focus on estimating a feature of the population picture. The most popular feature in science and statistics is the center of gravity, or mean, of the picture. The great thing about estimating the mean is that we have an approximate method that works for **any population picture**. (This is a bit inaccurate; there are some mathematical situations in which our method does not work, but they tend to be—for better or worse—ignored in practice and are definitely way beyond the scope of a first course.) The not-so-great thing is that determining whether the approximation is good can be very tricky.

The method we use is Gosset's approximate confidence interval estimate of μ. Its formula is

$$\bar{x} \pm t^*(s/\sqrt{n}),$$

with the method of determining t^* explained in Result 17.4.

Gosset derived this formula by assuming that the population picture is a normal pdf. Thus, Gosset's confidence level is **exact** if the population is a Normal pdf. Gosset's work was a solid piece of mathematics; by assuming a normal population, he was able to derive the family of t-curve pdfs which made his formula easy to obtain. Strictly speaking, Gosset's math result is not useful to scientists. As a scientist, how could you ever be sure that the unknown population was **exactly a Normal curve?**

Thus, a serious scientist must always wonder whether Gosset's confidence interval is valid—i.e., performs as advertised—for the true unknown population picture. Statisticians have been studying this issue for more than 50 years using a combination of fancy math arguments and computer simulations. In this chapter you saw the results of several computer simulation experiments that explored the behavior of Gosset's confidence interval. The issue of the performance of Gosset's confidence interval is too huge and complicated to lend itself to a one sentence answer, but the following summary is accurate, while not very precise.

1. Gosset's big weakness is skewness in the population picture. We saw that for the small amount of skewness in the cat population, Gosset worked fine for n as small as 10. For more pronounced skewness—see the log-normal and exponential examples—even for values of n much larger than 30, Gosset might perform very poorly.

2. This item is more complicated mathematically. A thorough presentation is beyond the scope of this course, but I want to mention it. Look at the formula for the Normal curve in Equation 7.6; x's that are much larger or much smaller than the mean, μ are said to be in the tail of the curve. For any Normal curve, the tails behave like

$$\exp[-(x-\mu)^2].$$

   In math terms, this is a very rapid decay; i.e., as x moves away from μ this number becomes close to zero very rapidly. By contrast, for the Laplace family of pdfs, the tails behave like

$$\exp[-|x-\mu|].$$

   which is a much slower decay that in the Normal curves. In everyday language, we say that the Laplace pdfs have heavier tails than the Normal pdfs. Heavier tails affect the performance of Gosset's confidence interval. This last idea is explored a bit in a Practice Problem and in Chapter 18.

17.7 Practice Problems

1. Recall the community of 7,000 households that was introduced in Example 17.2. The variable of interest is the number of children in the household who are attending public school. I used this community to create two populations: population 1 consisting of all 7,000 households; and population 2 consisting of the 4,200 households that have at least one child attending public school.

   (a) Calculate the mean, μ and median, ν of the two populations.

   (b) Use the probability histograms in Figure 17.3 to obtain $P(2 \leq X \leq 4)$ and $P(2 \leq Y \leq 4)$. Explain your reasoning.

   (c) It is possible to obtain probabilities for Y (population 2) from population 1 and the definition of conditional probability, Equation 16.2. For example,

   $$P(Y = 2) = P(X = 2|X > 0) = \frac{P(X = 2 \text{ and } X > 0)}{P(X > 0)} =$$

   $$P(X = 2)/0.600 = 0.120/0.600 = 0.20.$$

   Use a similar argument to determine $P(Y = 3)$ from $P(X = 3)$ and $P(X > 0) = 0.600$.

2. Intelligence quotients (IQs) for members of a population are generally manipulated (a less judgmental term is transformed) in a non-linear manner so that they have approximately a N(100,15) pdf. For the sake of this question, let's take this to be true. Use the website to answer the following questions. In parts (a)–(c) you are asked to determine the proportion of the population with the given feature. In (d) you will learn a new way to use the website.

   (a) An IQ of 90 or less.

   (b) An IQ of 130 or more.

   (c) An IQ between 95 and 120.

   (d) An IQ such that 83 percent of the population have a lower IQ.

3. Example 17.3 introduced you to data obtained by my friend Walt. For his 250 games, Walt obtained $\bar{x} = 19.916$ pairs of tiles unmatched and $s = 12.739$. Use these data to obtain Gosset's 95% confidence interval estimate of the mean, μ, of his population.

4. Refer to the previous question. Walt's 250 games can be viewed as a *mixture* of two distributions: games he wins, which always give $x = 0$; and games he loses which give $x > 0$. It is reasonable to ask, "Given that Walt loses, what is the population mean for the process that generates his observed values of X?" For this problem, the data consist of $n = 216$ games, with $\bar{x} = 23.051$ and $s = 10.739$. Use these data to obtain Gosset's 95% confidence interval estimate of the mean, μ, of the new population.

5. I performed seven—by now—familiar simulation experiments to investigate the performance of Gosset's 95% confidence interval when sampling from a Laplace pdf. As with the family of exponential pdfs, for a given value of n, one simulation covers the entire family. The results of my seven simulations are presented in Table 17.12.

   Write several sentences that describe what this table reveals.

6. In this problem I will explore the connection between a Poisson Process with parameter λ per unit time and the exponential distribution. If you are not interested in this connection, feel free to jump (skip, hop) ahead to the next practice problem.

   Consider the following function. For any $\lambda > 0$,

$$\begin{aligned} f(x) &= \lambda \exp(-\lambda x) \quad \text{for } x \geq 0 \\ &= 0 \quad \text{for } x < 0 \end{aligned}$$

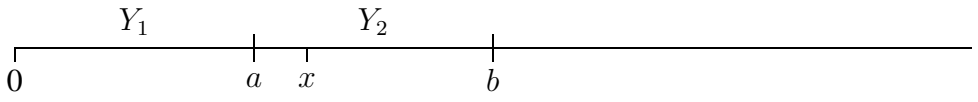
   If you have studied calculus, you can verify that the total area under this curve equals 1, for every value of $\lambda > 0$. Thus, for any fixed value of $\lambda > 0$, it is a pdf. If we let $\lambda > 0$ vary, we get a family of pdfs, called the family of exponential pdfs with parameter $\lambda > 0$. (By the way, it is the family of exponential distributions; don't call it the exponential family, because that means something else!) The parameter λ is called the rate of the exponential. The mean and standard deviation of an exponential equal the same number, the inverse of the rate:

$$\mu = \sigma = 1/\lambda.$$

   If X is a random variable with the exponential pdf with parameter λ, then for two numbers, $a < b$, it can be shown—using calculus—that

$$P(a < X \leq b) = \exp(-\lambda a) - \exp(-\lambda b).$$

   Now suppose that we have a Poisson Process with rate λ per unit time. For the numbers $a < b$ in the previous paragraph, let Y_1 be the number of successes in the Poisson Process in the interval $[0, a]$ and let Y_2 be the number of successes in the Poisson Process in the interval $(a, b]$. In addition, let X be the time of the first success in the Poisson Process. In the timeline below, x denotes the observed value of the random variable X. As drawn below, $a < x \leq b$; i.e., the event $(a < X \leq b)$ has occurred.



   Looking at this picture we can see that

$$a < X \leq b \text{ is the same event as } (Y_1 = 0 \text{ and } Y_2 > 0).$$

From Chapter 13: Y_1 has a Poisson distribution with parameter λa; Y_2 has a Poisson distribution with parameter $\lambda(b-a)$; and Y_1 and Y_2 are independent. Furthermore,

$$P(Y_1 = 0) = \exp(-\lambda a) \text{ and } P(Y_2 > 0) = 1 - P(Y_2 = 0) = 1 - \exp(-\lambda(b-a)).$$

Thus,

$$P(Y_1 = 0 \text{ and } Y_2 > 0) = \exp(-\lambda a) \times [1 - \exp(-\lambda(b-a))] = \exp(-\lambda a) - \exp(-\lambda b).$$

In words, the distribution of X is exponential with parameter λ.

Thus, we have another way to view a Poisson Process. We begin at time 0; at random time X_1 a success occurs; at random time $X_1 + X_2$ a second success occurs; at random time $X_1 + X_2 + X_3$ a third success occurs; and so on. The **times between arrivals**

$$X_1, X_2, X_3 \ldots$$

are i.i.d. random variables with the exponential distribution with parameter λ.

This is helpful for studying Poisson Processes because now—through the use of i.i.d. exponential random variables—we are able to simulate a realization of a Poisson Process for any length of time that we desire.

7. This last practice problem is aimed at students who are interested in computer programming. Feel free to skip this problem if you are not interested.

   Let Q be any number, strictly between 0 and 1: $0 < Q < 1$. For a measurement random variable X, consider the equation $P(X \leq x_Q) = Q$. The idea of this equation is that given Q, the unknown in this equation is x_Q. For example, in part (d) of Practice Problem 2, $Q = 0.83$ and X has a N(100,15) distribution. We found that $x_{0.83} = 114.312$; i.e., $P(X \leq 114.312) = 0.83$. We call the number x_Q the Q-quantile of the distribution of X, or, more briefly, the Q-quantile of X.

   Computer programmers are good at simulating i.i.d. trials from the uniform distribution on the interval $[0, 1]$. How do they simulate from other pdfs? As it turns out, the answer is quite simple.

   To be precise, suppose you want to generate an observation X from the exponential pdf with $\lambda = 2$. Simply generate a value u from the uniform distribution on the interval $[0, 1]$ and transform it to $x = X_u$; i.e., the generated observation is equal to the u-quantile of X. With this method, the generated x is from the exponential pdf with $\lambda = 2$. Contact me if you want more details. The important point is that if we can generate i.i.d. trials from the uniform distribution on the interval $[0, 1]$, then we can generate i.i.d. trials from any pdf!

Table 17.12: Results from seven simulation experiments. Each simulation had 10,000 reps, with a rep consisting of a random sample of size n from a Laplace pdf. For each sample, Gosset's approximate 95% confidence interval estimate of μ is computed and Nature classifies it as too small, too large or correct.

Sample size (n)	Number of Too Small Intervals	Number of Too Large Intervals	Number of Incorrect Intervals
10	179	195	374
20	224	207	431
40	229	245	474
80	251	245	496
160	263	256	519
320	246	243	489
640	266	270	536

17.8 Solutions to Practice Problems

1. (a) I will use the population counts given in Table 17.4. For both populations, the total number of children is:

$$1260(1) + 840(2) + 714(3) + 546(4) + 420(5) + 294(6) + 126(7) = \ 12{,}012.$$

Thus, the mean for population 1 is $12012/7000 = 1.716$ and the mean for population 2 is $12012/4200 = 2.860$.

Also from the population counts, we can see that the median of population 1 is 1 and the median of population 2 is 2.5.

(b) For both population pictures, $\delta = 1$; thus, the area of each rectangle equals its height. Reading from the pictures,

$$P(2 \leq X \leq 4) = 0.120 + 0.102 + 0.078 = 0.300 \text{ and}$$

$$P(2 \leq Y \leq 4) = 0.20 + 0.17 + 0.13 = 0.50.$$

(c) We note that

$$P(Y = 3) = P(X = 3|X > 0) = P(X = 3)/P(X > 0) = 0.102/0.600 = 0.17.$$

2. For parts (a)–(c) we use the default *Area from a value* option. For all parts we enter 100 in the *Mean* box and 15 in the *SD* box.

(a) We want $P(X \leq 90)$. Click on the *Below* option and enter 90 in its box. Click on *Recalculate* and the answer 0.2525 appears in the *Results* box. In words, just over one-quarter of the population have an IQ of 90 or less. Note that even though IQs are usually (?) reported as an integer, I do not use a continuity correction. I choose to treat IQ as a measurement.

(b) We want $P(X \geq 130)$. Click on the *Above* option and enter 130 in its box. Click on *Recalculate* and the answer 0.0228 appears in the *Results* box. In words, approximately one-in-44 population members have an IQ of 130 or larger.

(c) We want $P(95 \leq X \leq 120)$. We have not done a problem like this before. Click on the *Between* option and enter 95 [120] in its left [right] box. Click on *Recalculate* and the answer 0.5393 appears in the *Results* box. Be careful! If you place 120 in the left box and 95 in the right box, the site gives the answer -0.5393.

(d) This one's a bit tricky. We want the number a which satisfies

$$P(X \leq a) = 0.83.$$

First, change the option from *Area from a value* to *Value from an area*. This will change the site's labels. Enter 0.83 in the *Area* box and reenter your values of 100 for *Mean* and 15 for *SD*. Click on the *Below* option and the answer, 114.311 appears in its box; i.e., I did **not** need to click on *Recalculate*. If I wanted a new area, then I would need to click on *Recalculate*. For example, I changed the area to 0.75, clicked on *Recalculate* and obtained the answer 110.113 in *Below's* box.

3. Go to the t-curve calculator website,

   `http://stattrek.com/online-calculator/t-distribution.aspx`

   Following the presentation in these *Course Notes*, enter $n - 1 = 250 - 1 = 249$ in the degrees of freedom box and 0.975 in the *Cumulative probability* box. Click on *Calculate* and $t^* = 1.970$ appears in the *t score* box. Note that with this large number of degrees of freedom, t^* barely exceeds $z^* = 1.96$ for the N(0,1) curve.

   Gosset's 95% confidence interval estimate of μ is:

$$19.916 \pm 1.97(12.739/\sqrt{250}) = 19.916 \pm 1.587 = [18.329, 21.503].$$

4. I proceed as in (2) above, but enter $n - 1 = 216 - 1 = 215$ in the degrees of freedom box and 0.975 in the *Cumulative probability* box. Click on *Calculate* and $t^* = 1.971$ appears in the *t score* box.

   Gosset's 95% confidence interval estimate of μ is:

$$23.051 \pm 1.971(10.739/\sqrt{216}) = 23.051 \pm 1.440 = [21.611, 24.491].$$

5. For $n \geq 40$, Gosset's confidence interval appears to behave as advertised. For example, with $n = 40$, Gosset gives 474 incorrect intervals, which is within sampling error of the target 500. The surprising feature in this table is that for $n \leq 20$, Gosset's intervals **give too few incorrect intervals**! We have not seen this situation before!

   This is not the worst situation that could occur, but it does mean that the Gosset's intervals are wider than they need to be. The reason this happens is because the Laplace pdfs have a heavier tail than the Normal pdfs. (See comments in item 2 on page 458.)

17.9 Homework Problems

1. Refer to the population presented in Table 17.13. Let X denote a random variable whose distribution is given by this population.

   (a) Draw the probability histogram for this population.

   (b) What is the mean, μ, of this population. (No computation is required.)

   (c) Suppose we have a random sample of size $n = 20$ from this population. Do you think that Gosset's 95% confidence interval will perform well? Briefly explain your answer.

   (d) Calculate $P(2 < X \leq 5)$.

2. The random variable X has a uniform distribution on the interval 0 to 20. This means that the pdf of X is

$$\begin{aligned} f(x) &= 0.05, \quad \text{for } 0 \leq x \leq 20 \\ &= 0, \quad \text{otherwise} \end{aligned}$$

   (a) Draw the graph of this pdf.

   (b) Use your graph to calculate $P(X \leq 5)$; $P(7 < X \leq 15)$; $P(X > 35)$.

3. In Example 12.1, I introduced you to my friend Bert and his 100 games of online solitaire mahjong. Recall that Bert won 29 of his 100 games. Recall that in Example 17.3 I showed how to report the outcome of a mahjong game as a count. For the 71 games Bert lost, his mean number of pairs of tiles remaining was 23.93, with a standard deviation of 11.33.

   Assuming that these 71 numbers are the realizations of 71 i.i.d. trials, calculate Gosset's 95% confidence interval estimate of μ.

4. Refer to the previous problem. Briefly compare your results for Bert with my results for Walt in Practice Problem 4.

5. Recall that in Section 17.5 I used Reggie's dart data from 10 feet to illustrate the use of the *vassarstats* website. With the assumption of i.i.d. trials, calculate Gosset's 95% confidence interval estimate of μ for Reggie's data from 12 feet, reproduced below.

   163 164 168 174 175 186 191 196 196 197 200 200 201 203 206

6. There is a family of uniform pdfs, with two parameters: the left and right boundaries of the rectangle, A and B. For example, if $A = 0$ and $B = 1$, we get the uniform pdf on the interval $[0, 1]$; $A = 0$ and $B = 20$, we get the uniform pdf on the interval $[0, 20]$. I want you to explore the performance of Gosset's 95% confidence interval estimate of $\mu = (A + B)/2$ for a i.i.d. data from the uniform pdf on the interval $[A, B]$.

   I performed three simulation experiments of the usual sort, with 10,000 reps for each experiment. My results are in Table 17.14. Briefly describe what this table reveals.

Table 17.13: The population distribution for Homework problem 1.

x	0	1	2	3	4	5	6	Total
$P(X = x)$	0.05	0.10	0.20	0.30	0.20	0.10	0.05	1.00

Table 17.14: Results from three simulation experiments. Each simulation had 10,000 reps, with a rep consisting n i.i.d. trials from a **uniform pdf**. For each sample, **Gosset's** approximate 95% confidence interval estimate of μ is computed and Nature classifies it as too small, too large or correct.

Sample size (n)	Number of Too Small Intervals	Number of Too Large Intervals	Number of Incorrect Intervals
10	250	268	518
20	272	247	519
40	239	240	479

7. I performed a simulation experiment with 10,000 reps. I used the family size population 1 in Figure 17.3. Each rep generated a random sample of size $n = 20$ and constructed Gosset's 95% confidence interval estimate of μ. The experiment yielded 641 incorrect intervals, with 120 of one type and 521 of the other type.

   What is your opinion as to how many of the confidence intervals—120 or 521—were too large? Explain your answer.

Bibliography

[1] "Perspectives: Overheard", *Newsweek*, March 20, 1989, page 19.

Chapter 18

Inference for One Numerical Population: Continued

Chapter 17 did most of the *heavy lifting* for inference for one numerical population. By comparison, this chapter is pretty user-friendly.

18.1 A Test of Hypotheses for μ

This section is very similar to Section 12.5, which presented a test of hypotheses for a binomial p. Again, the idea is that of all of the possible values of μ, there is one value of special interest to the researcher. This known special value of interest is denoted by μ_0 and the null hypothesis is that $\mu = \mu_0$. As in Chapter 12, the justification for the value μ_0 is: *history*; *theory*; or *contracts or law*. As in Chapter 12, the test of this section is not terribly useful in science. Recall that the very important McNemar's test of Chapter 16 was a special case of the not-so-important test of Chapter 12. Similarly, Chapter 20 will present an important use for the test of this section.

As usual, there are three possibilities for the alternative:

$$\mu > \mu_0; \qquad \mu < \mu_0; \qquad \text{or} \qquad \mu \neq \mu_0.$$

As in Chapter 17, we assume that our data will consist of n i.i.d. random variables:

$$X_1, X_2, X_3, \ldots, X_n,$$

with summary random variables $\bar{X}$ and S, the mean and standard deviation of these variables. The observed values of these guys are:

$$x_1, x_2, x_3, \ldots, x_n, \bar{x} \text{ and } s.$$

Because our hypotheses involve the mean of the population, the obvious and natural choice for the test statistic is $\bar{X}$, with observed value $\bar{x}$. In order to obtain an approximate sampling distribution we standardize $\bar{X}$ and obtain

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}.$$

We don't yet have our test statistic; there is a flaw inherent in this Z: we don't know the values μ and σ. Handling σ is easy enough; we replace it in Z with S, giving

$$Z' = \frac{\bar{X} - \mu}{S/\sqrt{n}}.$$

We will follow our approach of Chapter 17 and use Gosset's t-curve with $df = n - 1$ to obtain approximate probabilities for Z'. But Z' is not a test statistic because we don't know the value of μ. Just in time, we recall that we want to know how the test statistic behaves **on the assumption that the null hypothesis is true.** Given that the null hypothesis is true, we can replace the unknown μ in Z' with the known μ_0. The result is our test statistic:

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}; \tag{18.1}$$

after data are collected, the observed value of T is

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}. \tag{18.2}$$

The three rules for finding the P-value are similar to earlier rules and are summarized in the following result. The website we are using in these *Course Notes* gives areas to the left under a t-curve. In the items listed below, I include an equivalent *area to the right* rule.

Result 18.1 *In the formulas below, t is given in Equation 18.2 and areas are computed under the t-curve with $df = n - 1$.*

1. *For the alternative $\mu > \mu_0$, the approximate P-value equals the area to the right of t. If you prefer, the approximate P-value equals the area to the left of $-t$.*
2. *For the alternative $\mu < \mu_0$, the approximate P-value equals the area to the left of t. If you prefer, the approximate P-value equals the area to the right of $-t$.*
3. *For the alternative $\mu \neq \mu_0$, the approximate P-value equals twice the area to the right of $|t|$. If you prefer, the approximate P-value equals twice the area to the left of $-|t|$.*

I will illustrate the use of these rules.

Suppose that we have

$$\mu_0 = 20, n = 16, \bar{x} = 23.00 \text{ and } s = 8.00.$$

First, we use Equation 18.2 to obtain the observed value of the test statistic:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{23.00 - 20}{8.00/\sqrt{16}} = 3/2 = 1.50.$$

Using the website introduced in Chapter 17:

`http://stattrek.com/online-calculator/t-distribution.aspx`

and the rules above:

- For the alternative $>$: enter $n - 1 = 16 - 1 = 15$ for the degrees of freedom; enter $-t = -1.50$ in the *t score* box; and click on *Calculate*. The approximate P-value, 0.0772, appears in the *Cumulative probability* box.
- For the alternative $<$: leave 15 for the degrees of freedom; enter $t = 1.50$ in the *t score* box; and click on *Calculate*. The approximate P-value, 0.9228, appears in the *Cumulative probability* box.
- For the alternative $\neq$: the approximate P-value equals twice the area to the left of $-|t| = -1.50$. From the above, we know that this area equals 0.0772. Thus, the approximate P-value equals $2(0.0772) = 0.1544$.

If you believe that the population is symmetric or approximately symmetric, then the approximate P-values given above should be reasonably accurate, even for relatively small values of n.

If you suspect that the population is strongly skewed and your alternative is two-sided ($\neq$), my advice is to use the above rules if your n is *very large*. Of course, *very large* is vague; the guidelines we had in Chapter 17—i.e., how large depends on how skewed—are fine here too.

If, however, you suspect that the population is strongly skewed and your alternative is one-sided ($>$ or $<$), then my advice is to **never** use the rules above. I don't have the time to explain why, but it's related to the fact that for a population that is strongly skewed to the right [left] the incorrect confidence intervals are too small [large] much more often than they are too large [small]. This translates into the approximate P-value being either much too large or much too small.

18.2 Estimating the Median of a pdf

Recall that a fundamental feature of a pdf is that the total area under it is equal to 1. It thus follows that there exists a number ν (pronounced new) with the following property.

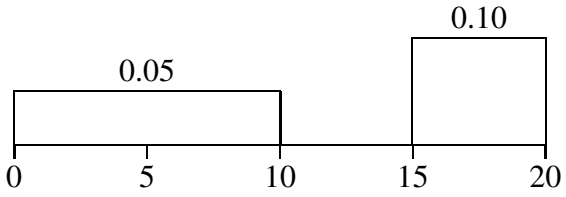
> The area under the pdf to the left (and right) of ν is equal to 0.5.

The number ν is called the median of the pdf, for rather obvious reasons. Note my use of the definite article: the median. I am being a bit dishonest here. Let me explain. For every pdf we have seen, including all the families of pdfs mentioned in Chapter 17, there is a unique number, ν, with the property that:

> the area under the pdf to the left (and right) of ν is equal to 0.5.

It is possible mathematically, however, for there to be an interval of numbers with this property. Figure 18.1 presents a pdf (a *combination* of two rectangles) for which all numbers in the closed interval $[10, 15]$ are medians. Notice how this happens. It requires a gap between two collections of possible measurement values **and exactly one-half** of the area is on each side of the gap. Such

Figure 18.1: A bizarre pdf with an interval of medians. In particular, every real number between 10 and 15 inclusive is a median of this pdf.



a picture is perfectly reasonable to a mathematician, but I am still waiting for someone to suggest a scientific situation for which it would be the pdf. As a result, I label such a pdf to be **bizarre**. In these *Course Notes* I exclude such bizarre pdfs although, with a bit of awkwardness, we could include them. I just don't want to do so!

There is a really amazing exact result for estimating ν. Recall that we assume that we will observe i.i.d. random variables:

$$X_1, X_2, X_3, \ldots, X_n.$$

The observed values of these random variables are denoted by:

$$x_1, x_2, x_3, \ldots, x_n.$$

We take these n numbers and sort them, from smallest to largest and denote these sorted data by:

$$x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq \ldots, \leq x_{(n)}.$$

Following the notation of Chapter 10, the median of these data is denoted by $\tilde{x}$ and this median is our point estimate of ν.

Here is the amazing part. Without any assumptions about the pdf, we can obtain an exact confidence interval estimate of ν. For $n \leq 20$ the exact confidence interval estimate of ν can be found in Table 18.1. Here is an example of its use.

Suppose we have a random sample of size $n = 10$ from a pdf. Using our table, we have the choice of the following exact confidence interval estimates of ν:

- $[x_{(1)}, x_{(10)}]$ is the exact 99.8% confidence interval estimate of ν;
- $[x_{(2)}, x_{(9)}]$ is the exact 97.9% confidence interval estimate of ν; and
- $[x_{(3)}, x_{(8)}]$ is the exact 89.1% confidence interval estimate of ν.

It will be instructive for you to see how to obtain the levels of the intervals in Table 18.1.

For each measurement random variable, X_i, define the random variable Y_i as follows:

$$\begin{array}{rcrl} Y_i & = & 1 & \text{if } X_i < \nu \\ & = & 0.5 & \text{if } X_i = \nu \\ & = & 0 & \text{if } X_i > \nu \end{array}$$

Table 18.1: Confidence interval estimates of the median of a pdf. The confidence levels are exact..

n	Confidence Interval	Exact Confidence Level	n	Confidence Interval	Exact Confidence Level
2	$[x_{(1)}, x_{(2)}]$	50.0%	14	$[x_{(2)}, x_{(13)}]$	99.8%
			14	$[x_{(3)}, x_{(12)}]$	98.7%
3	$[x_{(1)}, x_{(3)}]$	75.0%	14	$[x_{(4)}, x_{(11)}]$	94.3%
			14	$[x_{(5)}, x_{(10)}]$	82.0%
4	$[x_{(1)}, x_{(4)}]$	87.5%			
			15	$[x_{(3)}, x_{(13)}]$	99.3%
5	$[x_{(1)}, x_{(5)}]$	93.8%	15	$[x_{(4)}, x_{(12)}]$	96.5%
			15	$[x_{(5)}, x_{(11)}]$	88.2%
6	$[x_{(1)}, x_{(6)}]$	96.9%			
6	$[x_{(2)}, x_{(5)}]$	78.1%	16	$[x_{(3)}, x_{(14)}]$	99.6%
			16	$[x_{(4)}, x_{(13)}]$	97.9%
7	$[x_{(1)}, x_{(7)}]$	98.4%	16	$[x_{(5)}, x_{(12)}]$	92.3%
7	$[x_{(2)}, x_{(6)}]$	87.5%	16	$[x_{(6)}, x_{(11)}]$	79.0%
8	$[x_{(1)}, x_{(8)}]$	99.2%	17	$[x_{(3)}, x_{(15)}]$	99.8%
8	$[x_{(2)}, x_{(7)}]$	93.0%	17	$[x_{(4)}, x_{(14)}]$	98.7%
			17	$[x_{(5)}, x_{(13)}]$	95.1%
9	$[x_{(1)}, x_{(9)}]$	99.6%	17	$[x_{(6)}, x_{(12)}]$	85.7%
9	$[x_{(2)}, x_{(8)}]$	96.1%			
9	$[x_{(3)}, x_{(7)}]$	82.0%	18	$[x_{(4)}, x_{(15)}]$	99.2%
			18	$[x_{(5)}, x_{(14)}]$	96.9%
10	$[x_{(1)}, x_{(10)}]$	99.8%	18	$[x_{(6)}, x_{(13)}]$	90.4%
10	$[x_{(2)}, x_{(9)}]$	97.9%			
10	$[x_{(3)}, x_{(8)}]$	89.1%	19	$[x_{(4)}, x_{(16)}]$	99.6%
			19	$[x_{(5)}, x_{(15)}]$	98.1%
11	$[x_{(1)}, x_{(11)}]$	99.9%	19	$[x_{(6)}, x_{(14)}]$	93.6%
11	$[x_{(2)}, x_{(10)}]$	98.8%	19	$[x_{(7)}, x_{(13)}]$	83.3%
11	$[x_{(3)}, x_{(9)}]$	93.5%			
			20	$[x_{(4)}, x_{(17)}]$	99.7%
12	$[x_{(2)}, x_{(11)}]$	99.4%	20	$[x_{(5)}, x_{(16)}]$	98.8%
12	$[x_{(3)}, x_{(10)}]$	96.1%	20	$[x_{(6)}, x_{(15)}]$	95.9%
12	$[x_{(4)}, x_{(9)}]$	85.4%	20	$[x_{(7)}, x_{(14)}]$	88.5%
13	$[x_{(2)}, x_{(12)}]$	99.7%			
13	$[x_{(3)}, x_{(11)}]$	97.8%			
13	$[x_{(4)}, x_{(10)}]$	90.8%			

The argument below is invalid for a count response. Thus, in particular, the confidence levels in Table 18.1 are invalid for a count response. See Result 18.3 below for a fact of a positive nature for a count response.

The Y_i's are n random variables with the following features:

1. A researcher who does not know the value of ν will not be able to observe the Y_i's. Nature, however, will be able to observe the Y_i's.

2. For every value of i,
$$P(Y_i = 1) = 0.5 \text{ and } P(Y_i = 0) = 0.5.$$
This follows from the definition of the median of a pdf.

3. The Y_i's are independent. This is true because the X_i's are independent.

4. The Y_i's are Bernoulli trials with $p = 0.5$. Thus, if we define
$$Y = Y_1 + Y_2 + \ldots + Y_n,$$
then Y has the binomial distribution with parameters n and $p = 0.5$. In words, Y counts the number of response values that will be smaller than ν.

Let's look at the confidence interval $[x_{(3)}, x_{(8)}]$ for $n = 10$. As with all confidence intervals, it can be too small, too large or correct. Let's look at the first two of these possibilities:

- The confidence interval is too small if, and only if, its upper bound, $x_{(8)}$, is $< \nu$. This occurs if, and only if, $y \geq 8$; i.e., the eighth **ordered observation** is less than ν if, and only if, at **least eight unordered observations** are smaller than ν. Thus, the probability that the confidence interval is too small can be obtained easily from our binomial calculator website:

  `http://stattrek.com/Tables/Binomial.aspx`.

  If you go to this site and enter: 0.5, 10 and 8 and then click on *Calculate*, you will obtain
$$P(Y \geq 8) = 0.05469.$$

- The confidence interval is too large if, and only if, its lower bound, $x_{(3)}$ is $> \nu$. This occurs if, and only if: $y \leq 2$. Because the binomial distribution with $p = 0.5$ is symmetric,
$$P(Y \leq 2) = P(Y \geq 8) = 0.05469.$$
(If you don't like this argument, you may use the website a second time.)

- The probability that the confidence interval is correct is:
$$1 - P(\text{The CI is too small}) - P(\text{The CI is too large}) = 1 - 2(0.05469) = 0.89072,$$
which I have rounded to 0.891, or 89.1%, in Table 18.1.

For $n > 20$ it is possible to obtain an exact CI for ν with a modification of my argument above. In this class, however, we will be happy with the approximation given in the following result.

Result 18.2 (Approximate confidence interval estimate of the median, ν, of a pdf.) *Proceed as follows:*

1. *First, obtain the usual z^*—see Table 12.1 on page 296—for your target confidence level.*
2. *Calculate*
$$k' = \frac{n+1}{2} - \frac{z^*\sqrt{n}}{2}.$$
Round k' down *to the nearest integer and call the result k. (If k' is an integer, then $k = k'$; but if k' is an integer, you were probably sloppy in calculating it!)*
3. *The approximate confidence interval estimate of ν is $[x_{(k)}, x_{(n+1-k)}]$. Note that $x_{(n+1-k)}$ is the kth largest observation in the sorted list. Thus, this confidence interval is symmetric* in position*; it extends from the kth smallest observation to the kth largest observation.*

I will illustrate this method with a sample size of $n = 50$. I will choose 95% for the confidence level, which gives $z^* = 1.96$. I calculate

$$k' = \frac{51}{2} - \frac{1.96\sqrt{50}}{2} = 25.50 - 6.93 = 18.57,$$

which I round down to $k = 18$. Thus, the approximate 95% confidence interval estimate of ν is $[x_{(18)}, x_{(33)}]$.

For the purposes of this course, if $n > 20$ you may simply use Result 18.2 to obtain a confidence interval estimate of ν for a pdf. I will, however, present a brief digression on the quality of the approximation.

Consider the interval above for $n = 50$: $[x_{(18)}, x_{(33)}]$. For my definition of Y above, this interval is too small if, and only if, $Y \geq 33$, where $Y \sim$ Bin(50,0.5). With the help of the binomial calculator website,

`http://stattrek.com/Tables/Binomial.aspx`,

we find

$$P(Y \geq 33) = 0.0164.$$

Because of the symmetry in the binomial for $p = 0.5$, this is also the probability that the interval will be too large. Thus, the exact probability that this interval will be correct is

$$1 - 2(0.0164) = 0.9672.$$

This probability is a bit larger than our target, 0.95; thus, we might look at a narrower interval: $[x_{(19)}, x_{(32)}]$. Again, using the website, we find

$$P(Y \geq 32) = 0.0324.$$

As a result, the probability that the narrower interval will be correct is only $1-2(0.0324) = 0.9352$. I would stick with the original interval.

18.2.1 Examples with Real Data

I introduced you to Brian's study of running in Problem 1 of Section 1.8. Below are his ten sorted times to run one mile in combat boots:

321 323 329 330 331 332 337 337 343 347

and below are his ten sorted times to run one mile in jungle boots:

301 315 316 317 321 321 323 327 327 327

I will add the assumption that Brian's times, with either footwear, are the result of observing $n = 10$ i.i.d. trials from a pdf with unknown median ν. From Table 18.1, I choose the confidence interval with exact level 89.1%: $[x_{(3)}, x_{(8)}]$. For Brian's data the 89.1% confidence interval estimate of ν is:

$$[x_{(3)}, x_{(8)}] = [329, 337] \text{ for combat boots; and } [316, 327] \text{ for jungle boots.}$$

In Chapter 2 you learned of Sara's study of golf. She had two samples of size $n = 40$. If we assume that each sample is the result of observing i.i.d. trials, then we may use Result 18.2 to obtain an approximate 95% confidence interval estimate of each population median.

For $n = 40$ and $z^* = 1.96$, we get

$$k' = \frac{40+1}{2} - \frac{1.96\sqrt{40}}{2} = 20.50 - 6.20 = 14.30.$$

Thus, $k = 14$ and the approximate 95% confidence interval estimate of ν is $[x_{(14)}, x_{(27)}]$. Sara's data, sorted by club, are in Table 2.2 on page 29. Using this table, we find that

$$[x_{(14)}, x_{(27)}] = [107, 122] \text{ for the 3-Wood; and } [x_{(14)}, x_{(27)}] = [92, 108] \text{ for the 3-Iron.}$$

For Kymn's study of rowing in Chapter 2, $n = 5$ for each treatment. If we assume that each sample is the result of observing i.i.d. trials, then from Table 18.1, the 93.8% (exact) confidence interval estimate of ν is:

$$[x_{(1)}, x_{(5)}] = [489, 493] \text{ and } [479, 488] \text{ for treatments 1 and 2, respectively.}$$

These numerical values can be found in Figure 2.1 on page 28.

Finally, for Cathy's study of running in Chapter 2, $n = 3$ for each treatment. If we assume that each sample is the result of observing i.i.d. trials, then from Table 18.1, the 75.0% (exact) confidence interval estimate of ν is:

$$[x_{(1)}, x_{(3)}] = [521, 539] \text{ and } [520, 528] \text{ for the high school and park routes, respectively.}$$

These numerical values can be found in Table 2.5 on page 41.

Table 18.2: The population distribution for the cat population.

x	0	1	2	3	Total
$P(X = x)$	0.10	0.50	0.30	0.10	1.00

18.2.2 Estimating the Median of a Count Response

The confidence intervals given in Table 18.1 and Result 18.2 clearly state that the population must be a pdf. And there is a sign at a favorite park of mine: No Skateboards Allowed. Right. People *always* obey signs.

In this subsection I will explore what happens if the population is a probability histogram. I will begin with a particular example.

Suppose that a researcher selects a dumb random sample of size $n = 20$ from the cat population presented in Table 17.1 and reproduced in Table 18.2. After looking at Table 18.1, I decide on the exact confidence level of 95.9% which gives the interval $[x_{(6)}, x_{(15)}]$. Recall that the median, ν, of the cat population is 1. I will calculate the probability that the interval $[X_{(6)}, X_{(15)}]$ will be correct. (Are you thinking: Why? It's 95.9%. Recall, however, that the 95.9% comes from the assumption that the population is a pdf, which is no longer the situation.)

The confidence interval estimate will be too small if, and only if, its upper bound is smaller than $\nu = 1$. This happens if, and only if, at least 15 of the 20 observations are smaller than 1. I go to

`http://stattrek.com/Tables/Binomial.aspx`

and enter 0.1, 20 and 15; I click on *Calculate* and find that the probability the interval will be too small is 9.48×10^{-12}. Let's call this zero.

The confidence interval estimate will be too large if, and only if, its lower bound is larger than $\nu = 1$. This happens if, and only if, at most 5 of the 20 observations are 1 or 0. I go to

`http://stattrek.com/Tables/Binomial.aspx`

and enter 0.6, 20 and 5; I click on *Calculate* and find that the probability the interval will be too large is 0.0016.

As a result of the two above computations, the probability that the interval $[X_{(6)}, X_{(15)}]$ will be correct is $1 - 0.0016 = 0.9984$. This is way too large! The probability of an incorrect interval is more than 25 times smaller than I wanted ($0.041/0.0016 = 25.625$).

The above example for the cat population illustrates the following result.

Result 18.3 (Confidence interval for the median for a count population.) *The actual probability of a correct confidence interval is larger for a probability histogram than it is for a pdf. Sometimes it is very much larger.*

There is another—in my mind, more important—issue with the cat population. The possible values of X are 0, 1, 2 and 3. Below are two additional possibilities for the cat population. (Given that the cat population is totally hypothetical, we should not become attached to one possibility!)

u	0	1	2	3	Total
$P(U=u)$	0.100	0.401	0.309	0.190	1.000

w	0	1	2	3	Total
$P(W=w)$	0.100	0.399	0.311	0.190	1.000

These two populations are almost identical, but the median for U is 1 and the median for W is 2. This is a **huge difference**, especially when you note the values of U and W have the very small range of 0 to 3. By contrast, the mean of U is 1.589 and the mean of W is 1.591 (details not shown). Thus, the means are far superior to the medians for summarizing the extremely small difference between the two populations.

Thus, my general guideline is: **If the number of possible values of a count response is *small* do not use the median, either to summarize the data or describe the population.**

The remaining situation I will discuss is when the number of possible values of a count response is *large*. Admittedly, there is a big gap between *small* and *large*, but I need to limit the time we spend on this topic.

Earlier I derived the confidence level for one of the intervals in Table 18.1. The derivation was based on the following facts for a pdf:

$$P(X < \nu) = 0.50; P(X = \nu) = 0; \text{ and } P(X > \nu) = 0.50.$$

Almost always, one or more of these equations is **not** true for a probability histogram. (See the Practice Problems for the exception.) For example, for the cat population,

$$P(X < \nu) = P(X = 0) = 0.10; P(X = \nu) = P(X = 1) = 0.50;$$

$$\text{and } P(X > \nu) = P(X > 1) = 0.40.$$

The result we saw above—for $n = 20$—is that the actual error rate of the interval was more than 25 times smaller than the nominal error rate. This huge discrepancy was due to the fact that $P(X = \nu)$ is so large. If the researcher believes that $P(X = \nu)$, while not literally zero, is close to zero, then the actual confidence levels are only a bit larger than what they are for a pdf. I will illustrate these ideas in the following example.

Example 18.1 (Bob playing Tetris, circa 1990.) *Years ago, I enjoyed playing the video game Tetris. The score on the game I played was equal to the number of lines I completed before the screen overflowed. (If this makes no sense, don't worry; the score was a count and higher counts were better.) My sorted scores from 25 games are in Table 18.3.*

Even if you exclude my uncharacteristic small outlier, 51, there is a lot of variation in these responses. Neither the nature of Tetris nor my scores suggest that $P(X = \nu)$, whatever ν might be, is very large. Thus, I personally feel comfortable to use our designed-for-pdf confidence intervals for this count response. For completeness, I will calculate the 95% confidence interval estimate.

First,

$$k' = \frac{25+1}{2} - \frac{1.96\sqrt{25}}{2} = 13 - 4.9 = 8.1,$$

Table 18.3: Twenty-five sorted Tetris scores from 1990.

51	70	73	74	75	81	85	90	90	93	94	94	95
98	100	100	101	103	106	106	107	110	111	112	114	

giving $k = 8$. The approximate confidence interval is

$$[x_{(8)}, x_{(18)}] = [90, 103].$$

18.3 Prediction

In Chapter 14 you learned how to predict the total number of successes in m future Bernoulli trials and the total number of successes in a future observation of a Poisson Process. We will consider again prediction in Chapters 21 and 22 when we study regression. This brief section introduces you to two prediction methods for i.i.d. trials with a measurement response; i.e., for which the population is a pdf.

Here is our mathematical model. We plan to observe $(n + 1)$ i.i.d. random variables:

$$X_1, X_2, \ldots, X_n, X_{n+1}.$$

Our goal is to use the values of the first n of these:

$$X_1, X_2, \ldots, X_n,$$

to predict the value of the last one X_{n+1}. I will give you two methods for doing this:

1. Assume the pdf is a Normal curve with both the mean μ and the standard deviation σ unknown.
2. Making no assumptions about the form of the pdf. This is the so-called *distribution-free* method.

18.3.1 Prediction for a Normal pdf

Summarize the variables

$$X_1, X_2, \ldots, X_n$$

with their mean $\bar{X}$ and their standard deviation S. The observed values of these summaries are $\bar{x}$ and s. Below is the main result. I won't give any proof or motivation of it.

Result 18.4 (Prediction interval for X_{n+1} for a Normal pdf.) *With the framework described above, the prediction interval is*

$$\bar{x} \pm t^* s \sqrt{1 + (1/n)} \tag{18.3}$$

The value of t^* depends on the sample size and the desired probability of the interval being correct, as described below. (This is exactly the same procedure we had in Chapter 17 for Gosset's confidence interval.)

1. *Select the desired probability of a correct prediction and write it as a decimal; e.g., 0.95 or 0.99.*

2. *Subtract the desired probability from one and call it the error rate. Divide the error rate by two and subtract the result from one; call the final answer c; e.g., 0.95 gives $c = 1-0.05/2 = 0.975$ and 0.99 gives $c = 1 - 0.01/2 = 0.995$.*

3. *Go the website*

   `http://stattrek.com/online-calculator/t-distribution.aspx`.

   Next, enter $n-1$ for degrees of freedom; enter c in the Cumulative probability ... *box; and click on* Calculate. *The value t^* will appear in the* t-score *box.*

I will illustrate this result with data from Chapter 17.

Table 17.6 on page 432 presents the scores from 250 games of mahjong played by my friend Walt. I want to focus on the $n = 216$ games he lost ($x > 0$) and use that data to predict his score on the next game he loses. For these $n = 216$ games, we have the following summary statistics:

$$\bar{x} = 23.051 \text{ and } s = 10.739.$$

In addition, I commented on the fact—but did not demonstrate it—that the distribution of these 216 scores is approximately symmetric. Thus, even though a count can never have exactly a Normal pdf, I will use a Normal pdf as an approximation to the unknown probability histogram.

I want the probability that my prediction interval is correct to equal (approximately) 0.95. With the help of the t-curve website, I find $t^* = 1.971$. Thus, the (approximate) 95% prediction interval is

$$23.051 \pm 1.971(10.739)\sqrt{1 + (1/216)} = 23.051 \pm 21.216 = [1.835, 44.267] = [2, 44],$$

after rounding. This is a very wide interval!

Walt's next game, a loss, yielded $x = 5$; thus, the prediction interval was correct. Indeed, Walt played six more losing games—with one win mixed in—and his scores—12, 24, 6, 26, 9, 20—all were in the 95% prediction interval. Thus, the 95% prediction interval gave seven out of seven correct predictions. Hardly conclusive, but better than zero out of seven!

18.3.2 Distribution-Free Prediction

Let's assume that our population is a pdf, but otherwise we make no assumption about it. This is why we use the name **distribution-free**.

In this situation, we take our n variables:

$$X_1, X_2, X_3, \ldots, X_n,$$

and sort them from smallest to largest:

$$X_{(1)} \leq X_{(2)} \leq X_{(3)} \leq \ldots \leq X_{(n)},$$

Actually, because we assume that the population is a pdf, the probability that any adjacent values are equal is 0. Thus, when I present math arguments, I may assume that the sorted random variables satisfy:

$$X_{(1)} < X_{(2)} < X_{(3)} < \ldots < X_{(n)},$$

These sorted random variables yield the following sorted observed data:

$$x_{(1)} < x_{(2)} < x_{(3)} < \ldots < x_{(n)}.$$

My method below is messy for general n, so let's look at a fairly simple example.

First, I need to show you a useful consequence of assuming our random variables are i.i.d. from a pdf. Suppose that $n = 3$ and our sorted data are: 2, 4, and 9. There are six possibilities for unsorted data:

$$2, 4, 9 \quad 2, 9, 4 \quad 4, 2, 9 \quad 4, 9, 2 \quad 9, 2, 4 \quad \text{and} \quad 9, 4, 2$$

Here is the result: Because we assume that the $n = 3$ random variables are i.i.d., these six arrangements are equally likely to have occurred. (We had a special case of this result earlier for Bernoulli trials, which allowed us to develop the runs test and analyze the lengths or runs of successes.)

Visually, before we observe $n = 3$ i.i.d. random variables from a pdf, we know that our sorted variables will look like the following:

Now suppose that we plan to observe X_1, X_2, X_3 and X_4. We plan to sort **only the first three of these** which will give us a picture like the one above. It is true—and follows from the above, but don't worry about it if this is getting too mathematical—that each of the following four possibilities are equally likely to occur:

- $X_4 < X_{(1)}$;
- $X_{(1)} < X_4 < X_{(2)}$;
- $X_{(2)} < X_4 < X_{(3)}$; and
- $X_{(3)} < X_4$.

In the context of this section—prediction— we see that if we use $[x_{(1)}, x_{(3)}]$ as a prediction interval for X_4, the probability that the interval will be correct is $2/4$ (see the above listing), or 50%.

Let's catch our breath. In the terminology of this section, I have shown you that if we want to use $n = 3$ observations to predict the value of observation number $n + 1 = 4$, then the prediction interval

$$[X_{(1)}, X_{(3)}] \text{ with observed value } [x_{(1)}, x_{(3)}]$$

has probability 0.50 of being correct. Except for the fact that our probability of being correct, 0.50, is disappointingly small, this is a great method!

I will spare you the details, but the above argument can be generalized from $n = 3$ to any value of n. The result is given below.

Result 18.5 (Our first result on distribution-free prediction.) *Given we have n i.i.d. random variables from a pdf, the probability that the prediction interval*

$$[X_{(1)}, X_{(n)}] \tag{18.4}$$

will contain the observed value of X_{n+1} is equal to

$$1 - \frac{2}{n+1}. \tag{18.5}$$

Below are three evaluations of the probability given in Formula 18.5.

1. For $n = 3$, the probability is
$$1 - (2/4) = 0.50,$$
as we had found earlier.

2. For $n = 99$, the probability is
$$1 - (2/100) = 0.98.$$

3. For $n = 999$, the probability is
$$1 - (2/1000) = 0.998.$$

As the above computations show, for $n = 99$—and certainly for $n = 999$—the probability that the prediction interval is correct is perhaps larger than we require. And, obviously, for a large value of n, the interval from the minimum to the maximum of the observations will usually be extremely wide and, naturally, sensitive to even one extreme outlier. This is not a practical difficulty, because Result 18.5 can be generalized quite easily. (But don't worry about proving or even *seeing* the generalization; just be able to use it.)

Result 18.6 (The general result on distribution-free prediction.) *Let k be any positive integer with $k < (n/2)$. Given we have n i.i.d. random variables from a pdf, the probability that the prediction interval*

$$[X_{(k)}, X_{(n+1-k)}] \tag{18.6}$$

will contain the observed value of X_{n+1} is equal to

$$1 - \frac{2k}{n+1}. \tag{18.7}$$

I will look at the probability for this general result—Formula 18.7—for $n = 99$ and several choices of k.

1. For $n = 99$ and $k = 1$, the probability is
$$1 - (2/100) = 0.98,$$
which agrees with our earlier result.

2. For $n = 99$ and $k = 2$, the probability is
$$1 - (4/100) = 0.96.$$

3. For $n = 99$ and $k = 3$, the probability is
$$1 - (6/100) = 0.94.$$

4. For $n = 99$ and $k = 5$, the probability is
$$1 - (10/100) = 0.90.$$

Before I illustrate this result with real data, let me just comment that if the actual population is a probability histogram, then the actual probability of a correct prediction is greater than or equal to—perhaps much greater than—the value in Formula 18.7.

Let's revisit using the scores from Walt's 216 losing games of mahjong to predict his score on his next losing game. For our first method—assuming a Normal pdf—we obtained $[2, 44]$ for the 95% prediction interval.

For the distribution-free method we need to figure out the value of k. A good starting point is to compute
$$k' = 217(0.025) = 5.425.$$
(If you want a probability other than 95%, replace 0.025 in this equation by one-half of: one minus your desired probability.) This value of k' suggests trying $k = 5$ or $k = 6$.

1. With $k = 5$ the prediction interval is
$$[x_{(5)}, x_{(212)}] = [1, 45],$$
from the data in Table 17.6. The probability associated with this interval is
$$1 - (10/217) = 0.9539.$$

2. With $k = 6$ the prediction interval is
$$[x_{(6)}, x_{(211)}] = [2, 44],$$
which coincides with the Normal curve interval. The probability associated with this interval is
$$1 - (12/217) = 0.9447.$$

Of course, because Walt's response is a count, the above probabilities, 0.9539 and 0.9447, might be a bit smaller than the actual (unknowable) probabilities.

In summary, I would say that the Normal curve method and the distribution-free method give similar results.

18.3.3 Which Method Should be Used?

I performed a simulation experiment with $m = 1{,}000$ reps. For each rep I had Minitab generate $n = 999$ i.i.d. observations from a Normal pdf. (It does not matter which Normal pdf I use; for the work below, one experiment covers the entire family.) I then calculated two 95% prediction intervals for observation number 1,000:

- I used the Normal pdf interval method given in Formula 18.3.
- I used the distribution-free method given in Result 18.6. (Note that $k = 25$ gives exactly 0.95 for the probability of a correct interval.)

It is fairly easy to compare these two methods because they both give exactly 95% for the probability of a correct interval. So, what do we compare? The idea is that the narrower the prediction interval the more useful it is to the researcher. What I found in my simulation was pretty amazing: The mean width of the Normal pdf prediction intervals was 0.4%—yes, this is not a typo; less than one-half of one percent—narrower than the mean width of the distribution-free prediction intervals!

Thus, my recommendation is that for n large, use the distribution-free prediction interval. The Normal curve interval, however, might be preferred for small n. For example, if $n = 24$, from Result 18.6 we see that the largest possible probability of a correct distribution-free interval is for $k = 1$ and it equals

$$1 - (2/25) = 0.92.$$

There are two difficulties with this answer. First, I might want a larger probability of obtaining a correct interval. Second, I really hate to use $k = 1$ because it makes the prediction sensitive to even one outlier!

Finally, a published source might provide the mean and standard deviation of a set of data, but not a listing of the data. In this situation prediction assuming a Normal pdf can be used, but the distribution-free method cannot be used.

18.4 Some Cautionary Tales

Chapters 17 and 18 have been pretty technical. Lots of formulas and—for my taste—perhaps too much algebra. This section is a change-of-pace, but very important. I begin with a story from my life. (I know you love these!)

18.4.1 You Need More Than a Random Sample

In view of the fact that nearly every inference method in nearly every introductory Statistics text begins:

> Assume you have a random sample from a population

we should, perhaps, forgive novice researchers who seem to focus only on obtaining a random sample. I don't mean to be too blunt, but:

> **Having a random sample does not salvage a stupid study.**

Below is the promised story from my past.

Several years ago, the student government at UW–Madison published a booklet with the approximate title *The 100 Best Professors at UW–Madison*. I was very impressed that a colleague and friend of mine, Professor Wei-Yin Loh of Statistics, was included. Indeed, Wei-Yin was and is a wonderful teacher: thorough, creative, demanding. I think he deserved such an honor; but I am surprised he received it. Let me explain why.

The booklet proudly described the methodology. The student government selected a **random sample of undergraduate students** and asked each one to list the three best professors from whom he/she had taken a class at UW–Madison. The researchers compiled the results and the 100 professors whose names were mentioned most often were placed in the booklet.

Before you read further, think about this methodology. Can you spot its fatal flaws? I would imagine that I have failed to spot all of the flaws, but here are two huge flaws.

1. Full-time students typically take four–five courses each semester. Thus, a first semester freshman in the sample would have at most—many courses are taught by TAs and this was a compilation of professors—five professors to choose from, compared to up to 40 or more for a senior! Thus, a professor who teaches freshmen will have a huge advantage over a professor who teaches seniors.

2. A teacher who teaches a class of size 300 obviously has an advantage over a professor who teaches classes of size 15.

When I now tell you that my friend, Wei-Yin Loh, teaches smallish classes to juniors and seniors, you see why I was amazed that he received the honor of being among the *100 Best Professors at UW–Madison.*

18.4.2 Cross-sectional Versus Longitudinal Studies

I searched the web looking for a good explanation of cross-sectional and longitudinal studies and could not find one. If you know of one, please let me know. Lacking a better source, I will give you two stories from my career.

Early in my career as a teacher of introductory Statistics, I followed whatever textbook the *course coordinator* told me to use. After doing this a few years I noticed that many students were giving essentially the same comment in my student evaluations; below is one example of this collection of similar comments:

> If Wardrop does one more example about dice or decks of cards, I am going to ....
> (Note to reader: In an uncharacteristic display of restraint, I will delete the remainder of the comment.)

This surprised the math-person I was then; after all, when I was a student all of the examples involved selecting balls out of an urn. I survived, even though I had never seen an urn, except in an art museum; and those urns were too old and valuable to fill with balls!

Such experiences led me to decide to search the local newspaper for a real data set. As luck would have, the first study I happened upon was monumentally bad! Let me describe it to you.

The headline read,

> **Men Reach Their Peak Weight Before Age 55, But Women Keep Growing Heavier!**

I might be mistaken in remembering an exclamation point, but the headline was oozing with judgment. Oh, those virtuous men; those, well, the opposite-of-virtuous women. Let me now describe the basis for this headline.

Data on age, height, weight and sex were available for 20,000 persons. The data were collected over a brief period of time—as I recall, less than two years—during routine physical examinations in a physician's office.

First, the data were divided into two data sets: one for men and one for women. Within each sex, the data sets were further divided based on height, to the nearest inch. Thus, for example, there were data sets for men who were $\ldots 67, 68, 69, \ldots$ inches tall and similarly for women. The data for each sex-height combination was further divided into six age groups:

> 18–25, 25–34, 35–44, 45–54, 55–64, and 65–74.

Finally, for every combination of sex-height-age, the mean weight of the persons in the data set was calculated. The first thing to note is that the patterns were not as consistent as the headline implied; i.e., whoever wrote the headline was not being honest. Nevertheless, let me give you two examples that do agree with headline.

For women who were five feet, four inches tall the means, by age group, were as follows:

Age:	18–24	25–34	35–44	45–54	55–64	65–74
Mean Wt.:	135	142	152	154	157	154

Indeed, if you read from left-to-right the means consistently increase before decreasing for the last age group.

For men who were five feet, six inches tall the means, by age group, were as follows:

Age:	18–24	25–34	35–44	45–54	55–64	65–74
Mean Wt.:	150	160	163	164	163	160

Again, if you read from left-to-right the means increase until they begin to decrease in age group 55–64.

Think about this story and these data for a few moments. ...Nothing in these data tell us whether any particular person gained weight or lost weight or stayed the same weight throughout his/her life. All that the data showed is that for people of a fixed sex and height, the mean weight in one age group differs from the mean weight in an adjacent age group. What are some possible explanations for this?

1. The headline *could be correct*. And then everybody dies at age 75 when a meteor lands on their head.

2. Perhaps everybody gains weight throughout their lives, but for men a heavier weight leads to a larger mortality rate, beginning in their fifties. To put it bluntly, the dip in mean weight for men might be due to the heavier ones dying.

In technical terms, the study reported in the newspaper is a cross-sectional study. It took a sample of persons and measured them at a fixed point in time. By contrast, a longitudinal study would select a sample of persons and recorded their weights over a period of years. To put it succinctly, if you want to study the effect of time, you must study your units over time.

Here is a similar story. I read a study that sampled senior women and found that 75 year-old women ate, on average, a healthier diet than 60 year-old women. The analyst wrote, "This shows that as women age, their eating habits improve." What do you think?

18.4.3 Another Common Difficulty

Consider the community described in Example 17.2 on page 429. To briefly summarize, there are 7,000 households in the community; and the response of interest is the number of children in a households who attend public school. I defined the first population to be all 7,000 households and the second population to be the 4,200 households with at least one child in public school. Consider the following three possible ways to sample.

1. Researcher A selects n_1 households at random (dumb sample) from population 1.

2. Researcher B selects n_2 households at random (dumb sample) from population 2.

3. In the solution to Practice Problem 1 in Chapter 17, I showed that the total number of children in public school in the community is 12,012. Researcher C selects n_3 children in public schools at random (dumb sample) from the community.

Now, I will discuss these three samples, ignoring the issue of nonresponse; i.e., I will assume that every household or child sampled will supply the requested information (the number of children in the household attending public school).

The data from Researcher A may be used to estimate the entire probability histogram of population 1 or estimate its mean with confidence. If Researcher A discards from the sample the households that responded '0,' then the remaining data may be viewed as a random sample from population 2 and, hence, may be used to estimate the entire probability histogram of population 2 or estimate its mean with confidence.

The data from Researcher B may be used to estimate the entire probability histogram of population 2 or estimate its mean with confidence. These data should not be used to make any estimates of population 1 because it is biased in that it refuses to sample any household with the smallest response, 0.

My recommendation is that neither Researcher A nor B should estimate ν, the population median, even though the confidence levels will be exact for population 2. (Remember, the median

for population 2 is $\nu = 2.5$; thus, exactly one-half of the population responses are smaller [larger] than ν.) My recommendation is based on my opinion that for so few possible responses, the median is a poor way to summarize either population or sample.

The data from Researcher C should **not** be used to estimate any features of either population. Let me explain why for population 2.

Going back to Chapter 1 or 10, the notion of the **unit** or population member seemed dull, even by the standards of these notes! But it is important! In population 2, the population members—also called sampling units—are households. When Researcher C collects data, it is **not** the case that all households are equally likely to be selected because he/she is sampling children. In particular, on any given selection a particular family with response 7 is seven times as likely to be sampled as a particular family with response 1! (Do you see why?)

I opine that when the situation is described as I have done above—carefully specifying what each researcher is doing—then most people will see the difficulty with the sampling method of Researcher C. Real life, however, is not always so forthcoming. I will illustrate this idea with an example shared with me by Professor Stephen Stigler of the Department of Statistics at the University of Chicago. Any errors in the following are due to failings in my memory and **not** the work of Steve. In particular, I do **not** remember the means reported by Steve, only that the first mean was substantially smaller than the second mean.

Many years ago a survey of 50 *prominent men* was taken. Each man was asked two questions:

1. Question 1: How many children do you have? The 50 responses were combined and the mean was found to be 1.8.

2. Question 2: Including yourself, how many children were in your family when you were a child. The 50 responses were combined and the mean was found to be 3.3.

The conclusion of the study: This is a disaster! The size of families that produce prominent men is shrinking! Something needs to be done! Take a moment. Can you see the flaw in the above study? Of course, an obvious flaw is that these prominent men might have more children in the future; thus, the mean of 1.8 is likely a bit small. Here is a hint: Think about the sampling units.

For question 1, the researcher is sampling households, by sampling the male head. For question 2, the researcher is again collecting data on households, but is *sampling children* from the households, much like Researcher C above. Thus, households with more children are more likely to be sampled; this makes me opine that the mean of 3.3 is too large.

The above are sampling issues. Underlying the entire study is this notion that the children of prominent men are more likely to be, if not prominent, at least extra-special. Is this true? I have no idea. The interested reader is referred to, "Singapore's Patrimony (and Matrimony)" in *The Flamingo's Smile* by Stephen Jay Gould. This wonderful essay describes and criticizes a recent scheme in Singapore to encourage successful people to have more children and to reward those who do have more children.

I will leave you with one more, admittedly vague, example that combines some of the ideas of this section. I am sorry that I don't have a specific reference for what follows. I am **not** unduly troubled by this, because my goal is to encourage you to think about statistics you hear reported.

It is **not** my goal to critique America's welfare system because, frankly, I am not at all qualified to do so!

I heard the following exchange between two analysts. Analyst A stated:

> Welfare works! Data show that 50% of welfare recipients receive benefits for three months or less.

Analyst B stated:

> Welfare is a disaster! Data show that 50% of people currently on welfare, have received benefits for more than six months!

What do you think?

I can't say that I am a big fan of Analyst A. Yes, it's good that one-half of the recipients have a short stay on welfare, but an honest analysis should **not** ignore the half of the data set that costs the most to support.

As a statistician, I find myself to be more offended, however, by Analyst B who is taking cross-sectional data and implying—without actually stating it—that it gives a valid picture of what happens longitudinally. (A person on welfare one year will have a much higher chance of being in a sample than one on welfare for a week.)

18.5 Summary

As in Chapter 17, we plan to observe n i.i.d. random variables, denoted by:

$$X_1, X_2, X_3, \ldots, X_n,$$

with summary random variables $\bar{X}$ and S. These variables come from a population with mean μ and standard deviation σ, both of which are unknown. After the data are collected, we have the observed values of these random variables:

$$x_1, x_2, x_3, \ldots, x_n,$$

with observed values of the summary statistics $\bar{x}$ and s.

Section 18.1 presented the test of the null hypothesis that $\mu = \mu_0$, where μ_0 is a known number specified by the researcher. There are three options for the alternative:

$$\mu > \mu_0; \qquad \mu < \mu_0; \qquad \text{or} \qquad \mu \neq \mu_0.$$

The test statistic is given in Formula 18.1, reproduced below:

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}.$$

After the data are collected, the observed value of T is denoted by t and is given in Formula 18.2, reproduced below:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}.$$

Result 18.1 presented the rules for using the value of t and the website:

`http://stattrek.com/online-calculator/t-distribution.aspx`

to obtain the P-value for each of the three possible alternatives.

If the population is actually a member of the family of Normal pdfs, then the P-value is exact. As in Chapter 17, the approximate P-values are reasonable accurate if the population is symmetric or has a small amount of skewness. For a skewed population, I recommend against using this test with a one-sided alternative; the approximate P-values are very inaccurate, even for large values of n.

Section 18.2 presented the confidence interval estimate of the median, ν, of a population. With the assumption that the population is a pdf—and no assumption about its shape—the confidence level is exact. For $n > 20$, I give an approximate method, but its level can be made exact by using the website

`http://stattrek.com/Tables/Binomial.aspx`,

but we won't worry about that in this course.

Subsection 18.2.2 examined the performance of the confidence interval for the median for a count response.

Section 18.3 presented two methods for creating a prediction interval for the value of a future trial X_{n+1} from the same process that generated the data. The first method assumes that the population is a Normal pdf, with unknown mean and unknown standard deviation. The prediction interval is given in Formula 18.3, reproduced below:

$$\bar{x} \pm t^* s \sqrt{1 + (1/n)},$$

where t^* is the same number we used for Gosset's confidence interval.

The second method makes no assumptions about the population, other than it is a pdf. It is called the distribution-free interval. The interval is given in Formula 18.6, reproduced below:

$$[X_{(k)}, X_{(n+1-k)}].$$

The probability that this interval will contain the observed value of X_{n+1} is equal to

$$1 - \frac{2k}{n+1}.$$

Finally, Section 18.4 presents a number of cautionary tales on the use of statistical methods.

18.6 Practice Problems

1. I enjoy playing four-suit spider solitaire online. If you have ever played this game, you know that a large proportion of the games end in losses. In fact, for most games you can tell from the original display of cards that the game is essentially hopeless. Thus, I often quit a game without even trying! As a result, I will not study my *probability of winning a game*; suffice to say, it is small.

   I do, however, win quite often. When I win, the website reports the time I needed to complete the game, in seconds. Table 18.4 presents the sorted responses for $n = 49$ games that I won. The summary statistics for these $n = 49$ numbers are $\bar{x} = 1070.6$ and $s = 168.8$, both measured in seconds.

   (a) Obtain Gosset's approximate 95% confidence interval estimate of μ.

   (b) Obtain the approximate P-value for testing the null hypothesis that the mean equals 19 minutes, versus all three possible alternatives.

   (c) Obtain the approximate 95% confidence interval estimate of ν.

   (d) Assuming that the pdf is a Normal curve, calculate the 92% prediction interval for the time it takes to finish a future winning game. (When you do the homework, you will see why I make the unusual choice of 92%.)

   (e) Calculate the 92% distribution-free prediction interval for the time it takes to finish a future winning game.

   (f) Compare your answers to parts (d) and (e). Comment.

   (g) In seven future winning games, my sorted times to finish are: 960, 1049, 1058, 1372, 1448, 1448 and 2160. Comment on your answers to parts (d) and (e).

2. Below is yet another version of the cat population.

x	0	1	2	3	Total
$P(X = x)$	0.100	0.400	0.300	0.200	1.000

   (a) Verify that the median of the population, ν, equals 1.5 cats; i.e., describe the sorted response values in the 100,000 households.

   (b) For *this version* of the cat population, we have:

   $$P(X < \nu) = 0.50; P(X = \nu) = 0; \text{ and } P(X > \nu) = 0.50.$$

   Thus, the confidence levels in Table 18.1 are exact for this version of the cat population.

   (c) In this course we have dealt with many kinds of uncertainty. We are especially good at quantifying the uncertainty in the equally likely case. Now, I want your opinion on uncertainty related to the *behavior of the world*, which is, admittedly, much trickier. In particular, please comment on the following:

Table 18.4: The sorted times, in seconds, required to win a game of four-suit spider solitaire.

775	776	804	827	898	899	909	914	919	923
934	935	941	944	958	976	996	1019	1019	1021
1023	1033	1037	1045	1048	1052	1073	1085	1089	1094
1108	1116	1117	1130	1133	1142	1154	1156	1174	1182
1206	1269	1269	1286	1321	1350	1443	1466	1472	

How likely is it that **exactly** 50,000 households—not 49,999 and not 50,001—give a count response smaller than ν? (And, by logical implication, that **exactly** 50,000 households give a count response larger than ν?)

18.7 Solutions to Practice Problems

1. (a) This first question checks to see whether you remember what you learned in Chapter 17. Following the method presented in Result 17.4 on page 448, we obtain $t^* = 2.011$. Thus, Gosset's 95% confidence interval estimate of μ is:

$$1070.6 \pm 2.011(168.8/\sqrt{49}) = 1070.6 \pm 48.5 = [1022.1, 1119.1].$$

   (b) First, remember to convert $\mu_0 = 19$ minutes to $\mu_0 = 1140$ seconds. The observed value of the test statistic is

$$t = \frac{1070.6 - 1140}{168.8/\sqrt{49}} = -69.4/24.114 = -2.878.$$

   Next, go to the website

   `http://stattrek.com/online-calculator/t-distribution.aspx`

   and follow the rules given in Result 18.1.

   - For the alternative $<$, the approximate P-value is the area to the left of $t = -2.878$; this area equals 0.0030.
   - For the alternative $>$, the approximate P-value is the area to the left of $-t = 2.878$; this area equals 0.9970.
   - For the alternative $\neq$, the approximate P-value is twice the area to the left of $-|t| = -2.878$. Thus, it is $2(0.0030) = 0.0060$.

   (c) For 95% confidence, $z^* = 1.96$. Thus,

$$k' = \frac{49+1}{2} - \frac{1.96\sqrt{49}}{2} = 25 - 6.86 = 18.14,$$

which gives $k = 18$. Thus, the confidence interval is

$$[x_{(18)}, x_{(32)}] = [1019, 1116].$$

Note that this interval is almost identical to the interval for μ in part (a).

(d) Following the method in Result 18.3 on page 479, we to the website

`http://stattrek.com/online-calculator/t-distribution.aspx`;

enter 48 for the degrees of freedom; and place $c = 1 - 0.08/2 = 0.96$ in the *Cumulative probability* box. Click on *Calculate* and obtain $t^* = 1.789$. Thus, the 92% prediction interval is:

$$1070.6 \pm 1.789(168.8)\sqrt{1 + (1/49)} = 1070.6 \pm 305.0 = [765.6, 1375.6].$$

Because the computer reports data to the nearest second, I will round-off these endpoints to obtain $[766, 1376]$.

(e) First, I need to determine the value of k by solving:

$$1 - 0.92 = 0.08 = 2k/50; \text{ this gives } k = 2.$$

Thus, the 92% prediction interval is:

$$[x_{(2)}, x_{(48)}] = [776, 1466].$$

(f) The distribution-free interval 690 seconds wide and the Normal curve interval is 610 seconds wide. The distribution-free interval is greatly influenced by the three large outliers.

(g) The Normal curve prediction intervals contains only only four out of seven future games; the distribution-free method does better, containing six out of seven. For some unknown reason, my times increased dramatically after win number 49. In particular, the very large outlier of 2160 exceeded my second largest response by $2160 - 1472 = 688$ seconds! This trend—some very large observations—continued in subsequent data. Paradoxically, I believe this happened because I became *better* at playing the game. I am now able to win some particularly difficult games—which require much more time—whereas previously I would just quit (and lose) and the game would never make it into the data set.

2. (a) Recall that the population consisted of 100,000 households. Based on the table, exactly 50,000 households had either 0 or 1 cats, and the remaining exactly 50,000 households had either 2 or 3 cats. Thus, in the sorted list of population values, the value 1 is in position 50,000 and the value 2 is in position 50,001. Thus, $\nu = (1 + 2)/2 = 1.5$.

(b) Part (b) does not ask you to do anything.

(c) I would be amazed if this happened in real life!

Table 18.5: The sorted times, in seconds, required to win a game of four-suit spider solitaire.

827	919	923	976	996	1021	1023	1033
1052	1085	1089	1108	1116	1117	1130	1156
1174	1206	1269	1269	1321	1350	1443	1472

18.8 Homework Problems

1. Refer to Practice Problem 1. Table 18.5 presents the sorted times of my first 24 winning games. Essentially, I want you to mimic what we did in Practice Problem 1 for this new, smaller, data set.

   For the $n = 24$ numbers in Table 18.5, the summary statistics are $\bar{x} = 1128.1$ and $s = 163.0$

   (a) Obtain Gosset's approximate 95% confidence interval estimate of μ. Compare your answer to the answer we obtained with $n = 49$ times.

   (b) Obtain the approximate P-value for testing the null hypothesis that the mean equals 19 minutes, versus all three possible alternatives. Compare your answers to the answers we obtained with $n = 49$ times.

   (c) Obtain the approximate 95% confidence interval estimate of ν. Compare your answer to the answer we obtained with $n = 49$ times.

2. Refer to Homework Problem 1. Do the following for the $n = 24$ numbers in Table 18.5.

   (a) Assuming that the pdf is a Normal curve, calculate the 92% prediction interval for the time it takes to finish a future winning game. Compare this answer to the interval we obtained in Practice Problem 1.

   (b) Calculate the 92% distribution-free prediction interval for the time it takes to finish a future winning game. Compare this answer to the interval we obtained in Practice Problem 1.

   (c) Compare your answers to parts (a) and (b). Comment.

   (d) The sorted times of victories 25–34 are:

899	935	958	1019	1045	1094	1142	1182	1286	1466

   Use these data to evaluate your answers in (a) and (b). Comment.

Chapter 19

Comparing Two Numerical Response Populations: Independent Samples

Chapter 19 is very much like Chapter 15. The major—and obvious—difference is that in the earlier chapter the response was a dichotomy, but in this chapter the response is a number. If you revisit the material on the *four types of studies* in Section 15.2, you can see that the fact that the response was a dichotomy is irrelevant. In other words, everything you learned earlier about how the interpretation of an analysis depends on the type of study remains true in this chapter. In particular, for an observational study, if you conclude that numerical populations differ, then you don't know—based on the statistical analysis—why they differ. On the other hand, for an experimental study, if you conclude that numerical populations differ, then you may conclude that there is a causal link between the treatment and response.

In addition, the meaning of *independent random samples* for the different types of studies remains the same in the current chapter. There is even an extension of Simpson's Paradox for a numerical response, but time limitations will prevent me from covering this topic.

It is also true that Chapter 19 builds on the work of Chapters 17 and 18. In particular, recall that in Chapter 17 you learned that the population for a numerical response is a picture and the kind of picture depends on whether the response is a count or a measurement.

19.1 Notation and Assumptions

The researcher has two populations of interest. The methods of Chapters 17 and 18 may be used to study the populations separately. In this chapter, you will learn how to compare the populations.

- Population 1 has mean μ_1, variance σ_1^2 and standard deviation σ_1.
- Population 2 has mean μ_2, variance σ_2^2 and standard deviation σ_2.

I realize that specifying both the variance and standard deviation is redundant, but it will prove useful to have both for some of the formulas we develop.

We will consider procedures that compare the populations by comparing their means.

- We assume that we will observe n_1 i.i.d. random variables from population 1, denoted by:

$$X_1, X_2, X_3, \ldots, X_{n_1}.$$

  These will be summarized by their mean $\bar{X}$, variance S_1^2 and standard deviation S_1. The observed values of these various random variables are denoted by:

$$x_1, x_2, x_3, \ldots, x_{n_1}, \bar{x}, s_1^2 \text{ and } s_1, \text{ respectively.}$$

- We assume that we will observe n_2 i.i.d. random variables from population 2, denoted by:

$$Y_1, Y_2, Y_3, \ldots, Y_{n_2}.$$

  These will be summarized by their mean $\bar{Y}$, variance S_2^2 and standard deviation S_2. The observed values of these various random variables are denoted by:

$$y_1, y_2, y_3, \ldots, y_{n_2}, \bar{y}, s_2^2 \text{ and } s_2, \text{ respectively.}$$

- We assume that the two samples are independent.

I apologize for the cumbersome and confusing notation. In particular, in my μ's, σ^2's, n's S^2's, and so on, I use a subscript to denote the population, either 1 or 2; this is very user-friendly. You need to remember, however, that the random variables, data and some summaries from population 1 are denoted by *x's* and the corresponding notions from population 2 are denoted by *y's*. There is a long tradition of doing things this way in introductory Statistics. While it is confusing, its one virtue is that it allows you to avoid double subscripts until you take a more advanced Statistics class.

(Enrichment: Here is the problem with double subscripts—well, other than the obvious problem that they sound, and are, complicated. If I write x_{123} does it mean:

- Observation number 123 from one source of data?
- Observation 23 from population 1? or
- Observation 3 from population 12?

This could be made clear with commas; use $x_{1,23}$ for the second answer above and $x_{12,3}$ for the third answer. The only problem is: In my experience, statisticians and mathematicians don't want to be bothered with commas!**)**

The methods introduced in this chapter involve comparing the populations by comparing their means. For tests of hypotheses, this translates to the null hypothesis being $\mu_1 = \mu_2$, or, equivalently, $\mu_1 - \mu_2 = 0$. For estimation, $\mu_1 - \mu_2$ is the feature that will be estimated with confidence.

Our point estimator of $\mu_1 - \mu_2$ is $\bar{X} - \bar{Y}$. There is a Central Limit Theorem for this problem, just as there was in Chapter 17. First, it shows us how to standardize our estimator:

$$W = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}}. \tag{19.1}$$

Second, it states that we can approximate probabilities for W by using the N(0,1) curve and that in the limit, as both sample sizes become larger and larger, the approximations are accurate.

In order to obtain formulas for estimation and testing, we need to eliminate the unknown parameters in the denominator of W, σ_1^2 and σ_2^2. We also will need to decide what to use for our reference curve: the N(0,1) curve of the Central Limit Theorem and Slutsky or one of the t-curves of Gosset.

Statisticians suggest three methods for handling these two issues, which I refer to as Cases 1, 2 and 3. I won't actually show you Case 3 because I believe that it nearly worthless to a scientist; I will explain why I feel this way.

We will begin with Case 1; I will follow the popular terminology and call this the large sample approximate method.

19.2 Case 1: The Slutsky (Large Sample) Approximate Method

This method comes from Slutsky's Theorem. In Equation 19.1 for W, replace each population variance by its sample variance. The resultant random variable is:

$$W_1 = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{(S_1^2/n_1) + (S_2^2/n_2)}}. \tag{19.2}$$

(Note that I have placed the subscript of '1' on W to remind you that this is for Case 1.) It can be shown that in the limit, as both sample sizes grow without bound, the N(0,1) pdf provides accurate probabilities for W_1. Thus, for finite values of n_1 and n_2, the N(0,1) pdf will be used to obtain approximate probabilities for W_1. As a general guideline, I recommend using Case 1 only if $n_1 \geq 30$ and $n_2 \geq 30$.

The usual algebraic manipulation of the ratio that is W_1 yields the following result.

Result 19.1 (Slutsky's approximate confidence interval estimate of $(\mu_1 - \mu_2)$.) *With the notation and assumptions given in Section 19.1, Slutsky's approximate confidence interval estimate of $(\mu_1 - \mu_2)$ is:*

$$(\bar{x} - \bar{y}) \pm z^* \sqrt{(s_1^2/n_1) + (s_2^2/n_2)}. \tag{19.3}$$

As always in these intervals, the value of z^* is determined by the desired confidence level and can be found in Table 12.1 on page 296.

Before I give you an example of the use of Formula 19.3, I will tell you about the test of hypotheses for this section.

As I stated earlier in this chapter, the null hypothesis is $\mu_1 = \mu_2$ or, equivalently, $\mu_1 - \mu_2 = 0$. There are three options for the alternative:

$$H_1\text{: } \mu_1 > \mu_2; \quad H_1\text{: } \mu_1 < \mu_2; \quad \text{or} \quad H_1\text{: } \mu_1 \neq \mu_2.$$

I will abbreviate these as $>$, $<$ and $\neq$; no confusion should result provided you remember that μ_1 is to the left of the math symbol and μ_2 is to its right.

In order to obtain the formula for the test statistic, I look at Equation 19.2. I want to know how this random variable behaves if the null hypothesis is true. Well, if the null hypothesis is true, then $\mu_1 - \mu_2 = 0$. Make this substitution into Equation 19.2 and we get our test statistic Z, given below.

$$Z = \frac{(\bar{X} - \bar{Y})}{\sqrt{(S_1^2/n_1) + (S_2^2/n_2)}}. \tag{19.4}$$

Given the assumptions of this section, on the additional assumption that the null hypothesis is true, the sampling distribution of Z is approximated by the N(0,1) curve. The observed value of the test statistic Z is given by

$$z = \frac{(\bar{x} - \bar{y})}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}}. \tag{19.5}$$

The rules for finding the P-value are given in the following result.

Result 19.2 *In the formulas below, z is given in Equation 19.5 and areas are computed under the N(0,1) curve.*

1. *For the alternative $\mu_1 > \mu_2$, the approximate P-value equals the area to the right of z. Equivalently, the approximate P-value equals the area to the left of $-z$.*

2. *For the alternative $\mu_1 < \mu_2$, the approximate P-value equals the area to the left of z. Equivalently, the approximate P-value equals the area to the right of $-z$.*

3. *For the alternative $\mu_1 \neq \mu_2$, the approximate P-value equals twice the area to the right of $|z|$. Equivalently, the approximate P-value equals twice the area to the left of $-|z|$.*

In a later section of this chapter I will discuss the quality of the approximations behind Slutsky's confidence interval and Result 19.2.

I will end this section with illustrations of the estimation and testing methods of this section with a real data set from a student project. Other examples are given in the Practice and Homework Problems.

Luke performed a completely randomized design with a numerical response. A trial consisted of Luke hitting a pitched baseball. In treatment 1, he used an aluminum bat and and in treatment 2 he used a wooden bat. The response is the distance, in feet, that the ball traveled. Luke assigned 40 hits to each treatment, by randomization.

Trials that resulted in Luke missing the ball or hitting a foul-tip were 'done over.' This made the randomization a bit trickier (details not given), but I believe that Luke made the correct decision in doing this, for the following reason: The purpose of the study is to compare distances the ball travels to see whether wood or aluminum was superior. I can see no reason to *blame* the bat's material for a poor swing by Luke. (As I recall, Luke stated in his report that he had very few of these do-overs.)

In order to analyze Luke's data, we will assume that the data from each treatment are i.i.d. trials from a population and that the two sets of trials are independent. Luke's data yielded the following summary statistics:

$$\bar{x} = 179.6, s_1 = 62.1, n_1 = 40, \bar{y} = 166.2, s_2 = 54.2 \text{ and } n_2 = 40.$$

Luke's two sample standard deviations are similar in value; note that $62.1/54.2 = 1.146$; i.e., the standard deviation with the aluminum bat is about 15% larger than the standard deviation with the wooden bat. I will return to this issue later in this chapter.

Slutsky's 95% confidence interval estimate of $(\mu_1 - \mu_2)$ (Formula 19.3) is:

$$(179.6 - 166.2) \pm 1.96\sqrt{\frac{(62.1)^2}{40} + \frac{(54.2)^2}{40}} = 13.4 \pm 1.96(13.03) = 13.4 \pm 25.5 = [-12.1, 38.9].$$

Note that I have explicitly written the value of the radical, 13.03, because we will need it soon. In words, based on the confidence interval estimate, Luke's data are inconclusive. The mean with the aluminum bat is between 12.1 feet smaller and 38.9 feet larger than the mean with the wooden bat.

For his test of hypotheses, Luke chose the alternative $>$ because the conventional wisdom in baseball is that a ball hit with an aluminum bat travels farther than a ball hit with a wooden bat. Luke's observed value of the test statistic, Equation 19.5, is

$$z = \frac{(\bar{x} - \bar{y})}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}} = \frac{13.4}{13.03} = 1.028.$$

With the help of

`http://stattrek.com/online-calculator/normal.aspx`

I find that the area under the N(0,1) curve to the right of $z = 1.028$ equals 0.1520. There is evidence in the data in support of Luke's one-sided alternative, but Luke's approximate P-value does not meet the accepted standard for being convincing. Note, as an aside, that the approximate P-value for $<$ is 0.8480 and for $\neq$ is $2(0.1520) = 0.3040$.

19.3 Case 2: Congruent Normal Populations

In the previous section I gave you Slutsky's (large sample) method for comparing population means via estimation and testing. The natural follow-up is for me to show you how to compare means for the situation in which either or both of the sample sizes is *small*. I will do this in the current section.

This section is very *mathematical*, but not in the sense of having lots of algebra. It is mathematical in the sense that I will be presenting methods for a very specific set of mathematical assumptions. In this section, we will assume that the two populations being compared are congruent Normal pdfs.

According to *dictionary.com*, congruent means:

> Coinciding at all points when superimposed.

This implies that the two populations have identical spreads. For example, the N(μ_1, σ) and N(μ_2, σ) curves are congruent for all real numbers μ_1 and μ_2 and all positive real numbers σ. Thus, there are many pairs of Normal pdfs that satisfy the condition of being congruent; and, of course, many that do not.

Recall the definition of a **constant treatment effect**, given in Definition 5.1 on page 91:

In a clone-enhanced study, suppose that the response on treatment 1 minus the response on treatment 2 equals the nonzero number c **for every unit**. In this situation we say that the treatment has a **constant treatment effect** equal to c.

I argued in Part I that the constant treatment effect, if present, greatly simplifies the interpretation of statistical analyses. In short, I would say that assuming a constant treatment effect is both helpful and elegant.

Suppose we have an experimental comparative study. This means, recall, that there is one superpopulation of units and the two populations we compare represent what would happen if all of the units were assigned to the same treatment. Suppose that population 2 is the N(20,5) pdf. If the constant treatment effect is $c = 4$, then population 1 is the N(24,5) pdf. In general, if there is a constant treatment effect, then the two populations are congruent. In addition, if one population is a Normal curve, then so is the other.

To reiterate: In this section we will assume that both populations are Normal pdfs, with the added condition that they have the same variance. In our earlier notation, we assume that

$$\sigma_1^2 = \sigma_2^2.$$

Because these variances are assumed to be the same, for convenience I will drop the subscripts and write σ^2 for the common value of the population variance and σ for the common value of the population standard deviation.

We begin with the random variable W in Equation 19.1 on page 496. In this equation, replace the two population variances with their common value σ^2, yielding:

$$W = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{(\sigma^2/n_1) + (\sigma^2/n_2)}} = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sigma\sqrt{(1/n_1) + (1/n_2)}}. \tag{19.6}$$

In the last expression, I have moved σ^2 outside the square root sign, which mathemagically makes the exponent disappear! It can be shown that on the assumption of this subsection—Normal pdfs with common variances—the distribution of W is given exactly by the N(0,1) pdf. (I mention this in passing; we won't have any use for this fact.)

The remaining issue is the removal of the unknown σ from the formula for W. The proof of the best way to estimate σ is complicated, so I won't give it. In addition, I am unable to show you a brief motivation of the formula; thus, I will just give you the result.

Definition 19.1 (The pooled variance.) *With the notation of this chapter, our point estimate of σ^2 is:*

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}. \tag{19.7}$$

We call s_p^2 the **pooled estimate** *of the variance σ^2. The idea is that we combine, or pool, two estimates of σ^2 (s_1^2 and s_2^2) to obtain a better estimate.*

Note the following about this formula for s_p^2:

1. Each sample variance appears in the numerator.

2. The coefficient of each sample variance is equal to its degrees of freedom.

3. The sum of the coefficients in the numerator equals the number in the denominator. Thus, s_p^2 is referred to as a weighted average (mean) of the two sample variances, with weights given by degrees of freedom.

4. If $n_1 = n_2$, then s_p^2 reduces to $(s_1^2 + s_2^2)/2$, a natural combination, which is the unweighted average (mean) of of the two sample variances.

Below is the main result.

Result 19.3 *Define the random variable W_2 as follows:*

$$W_2 = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_p\sqrt{(1/n_1) + (1/n_2)}}, \tag{19.8}$$

where S_p^2 is the random variable that has observed value s_p^2 given in Equation 19.7.

Given the assumptions of this subsection, the exact distribution of W is given by the t-curve with $df = n_1 + n_2 - 2$.

The usual algebraic expansion of the ratio in Equation 19.8 yields a confidence interval estimate of $(\mu_1 - \mu_2)$, given below.

Result 19.4 (The Gosset confidence interval for $(\mu_1 - \mu_2)$.) *The Gosset confidence interval for $(\mu_1 - \mu_2)$ is given by:*

$$(\bar{x} - \bar{y}) \pm t^* s_p\sqrt{(1/n_1) + (1/n_2)}. \tag{19.9}$$

The value of t^* depends on the sample sizes and the desired confidence level, as described below.

1. Select the desired confidence level and write it as a decimal; e.g., 0.95 or 0.99.

2. Subtract the desired confidence level from one and call it the error rate. Divide the error rate by two and subtract the result from one; call the final answer c; e.g., 0.95 gives $c = 1 - 0.05/2 = 0.975$ and 0.99 gives $c = 1 - 0.01/2 = 0.995$.

3. Go the website

   `http://stattrek.com/online-calculator/t-distribution.aspx`.

   Enter the value of $n_1 + n_2 - 2$ in the *Degrees of freedom* box; enter c in the *Cumulative probability* box; and click on *Calculate*. The value t^* will appear in the *t score* box.

For testing, we use the same hypotheses that we used in Section 19.2, reproduced below for convenience. The null hypothesis is:

$$H_0\text{: } \mu_1 = \mu_2.$$

There are three options for the alternative:

$$H_1\text{: } \mu_1 > \mu_2; \quad H_1\text{: } \mu_1 < \mu_2; \quad \text{or} \quad H_1\text{: } \mu_1 \neq \mu_2.$$

In order to obtain the formula for the test statistic, look at Equation 19.8. We want to know how this random variable behaves if the null hypothesis is true. Well, if the null hypothesis is true, then $\mu_1 - \mu_2 = 0$. Make this substitution into Equation 19.8 and we get our test statistic T, given below.

$$T = \frac{\bar{X} - \bar{Y}}{S_p\sqrt{(1/n_1) + (1/n_2)}}, \tag{19.10}$$

Given the assumptions of this section (congruent normal populations) and the assumption that the null hypothesis is true, the exact distribution of T is given by the t-curve with $df = n_1 + n_2 - 2$. The observed value of the test statistic T is given by

$$t = \frac{\bar{x} - \bar{y}}{s_p\sqrt{(1/n_1) + (1/n_2)}}. \tag{19.11}$$

The rules for finding the P-value are given in the following result.

Result 19.5 *In the rules below, t is given in Equation 19.11 and areas are computed under the t-curve with $df = n_1 + n_2 - 2$.*

1. *For the alternative $\mu_1 > \mu_2$, the P-value equals the area to the right of t. Equivalently, the P-value equals the area to the left of $-t$.*

2. *For the alternative $\mu_1 < \mu_2$, the P-value equals the area to the left of t. Equivalently, the P-value equals the area to the right of $-t$.*

3. *For the alternative $\mu_1 \neq \mu_2$, the P-value equals twice the area to the right of $|t|$. Equivalently, the P-value equals twice the area to the left of $-|t|$.*

I will illustrate the use of these rules with a student project performed by Sheryl. A trial consisted of Sheryl performing a 1.5 mile sprint on her bicycle. In treatment 1, Sheryl loaded her pannier with 20 pounds and in treatment 2 she removed her pannier from her bike. The response is the time, in seconds, Sheryl required to complete the sprint. Sheryl assigned 5 trials to each treatment by randomization.

In order to analyze Sheryl's data, we will assume that we have independent i.i.d. trials from two normal populations with a common variance. Sheryl's data yielded the following summary statistics:

$$\bar{x} = 383.2, s_1 = 4.38, n_1 = 5, \bar{y} = 355.2, s_2 = 4.87, \text{ and } n_2 = 5.$$

Note that the ratio

$$s_2/s_1 = 4.87/4.38 = 1.11,$$

lends some support to the assumption of equal population variances.

We begin our analysis by computing s_p^2.

$$s_p^2 = \frac{4(4.38)^2 + 4(4.87)^2}{5 + 5 - 2} = \frac{4(19.18) + 4(23.72)}{8} = 21.45.$$

Because $n_1 = n_2$ we could have computed:

$$s_p^2 = \frac{(4.38)^2 + (4.87)^2}{2} = 21.45.$$

In any event, $s_p = \sqrt{21.45} = 4.63$.

You may verify that for $df = 5 + 5 - 2 = 8$ and 95%, $t^* = 2.306$. Thus, the 95% confidence interval estimate of $(\mu_1 - \mu_2)$ is

$$(383.20 - 355.20) \pm 2.306(4.63)\sqrt{1/5 + 1/5} = 28.00 \pm 2.306(2.928) =$$

$$28.00 \pm 6.75 = [21.25, 34.75].$$

In words, I conclude that Sheryl's mean time for completing her sprint increased by between 21.25 and 34.75 seconds when the weighted pannier is added to her bike.

I will also perform a test of hypotheses for Sheryl's data. The obvious choice for the alternative is $>$ because all would agree that adding weight will slow the bicycle. (Sheryl was **not** biking down a steep hill!)

The observed value of the test statistic is

$$t = 28.00/2.928 = 9.563.$$

With the help of the t-curve website, you can verify that the area under the t-curve with $df = 8$ to the right of $t = 9.563$ is equal to 0.0000, rounded to the nearest ten-thousandth. Minitab gives a more precise answer: 0.0000059, or slightly more than one in two-hundred thousand. In any event, this is a really small P-value!

I believe that performing a test for Sheryl's data is a bit, well, dumb. It is **obvious** that adding weight will slow Sheryl's biking. If you like to prove the obvious—which admittedly statisticians do quite often—then you will often get really small P-values. By contrast, I do think it is very interesting to estimate *how much* the weight increases her mean time to complete the sprint.

19.4 Case 3: Normal Populations with Different Spread

Case 2, presented above, is a nice piece of mathematics. Mathematically, it is a pretty general result: the populations need to be normal pdfs, but that's a big family; and the two population variances need to be the same number. With these restrictions the probabilities—confidence levels and P-values—are exact. Let me digress and explain the thought process of a mathematical statistician. If this sounds too arrogant, let me explain **my view** of the thought process of a mathematical statistician.

Mathematical results have *conditions* that are necessary for their proof. In Case 2, the conditions are normal populations and congruence (equal variances). Mathematical statisticians **relax** a condition and then figure out what will happen mathematically. Case 3 is the result of mathematical statisticians weakening the Case 2 assumptions by dropping the assumption of congruence. In other words, in Case 3, the first population is the N(μ_1, σ_1) pdf and the second population is the N(μ_2, σ_2) pdf.

Below are the main features of the Case 3 solution.

1. I won't use Case 3 in these *Course Notes*. My reason is given below this list.

2. The Case 3 solution **does not give exact probabilities**. I don't see this as a major problem, but I believe that it should be mentioned.

3. Like the Case 2 method, the Case 3 method uses a Gosset's t-curve as its reference, in this case the Gosset's t-curve is an approximation.

4. The main computational difficulty with Case 3 is that the formula for the degrees of freedom for the approximating t-curve is very complicated. As a result, if one is restricted to using a hand-held calculator, Case 3 is quite a mess. If, however, one takes advantage of living in the information age, the formula for the degrees of freedom is not an issue. Both Minitab and the multi-purpose *vassarstats* website allow one to obtain a Case 3 answer without calculating the degrees of freedom by hand.

5. This is a key point. Everybody agrees that, in practice, we don't need the variances to be **exactly equal** in order for Case 2 to give useful and approximately exact answers. There is some disagreement on how much they can differ before Case 2 answers become seriously deficient. In my opinion they need to differ a great deal—which, for space limitations, I will leave undefined—before I would discard Case 2 in favor of Case 3.

I will share with you two arguments for why I don't like Case 3. Bear with me please, because these arguments take some time to explain.

First, after you have finished this chapter, including the Practice Problems and Homework, look again at all the real data examples that I have given you. In every case I report the values of s_1 and s_2 and note that these values are reasonably close to each other. **This has been my experience with real data.** Almost always with real data that I have seen, the values of s_1 and s_2 have been similar. Each time this happens, it suggests that for the phenomenon being studied, there is, at most, weak evidence of a major difference between σ_1 and σ_2. Of course, one can imagine or manufacture a situation in which σ_1 and σ_2 clearly differ by a great deal, but in my experience these situations often—though not always—are examples of really stupid science! For example, let population 1 be the heights of male college students and let population 2 be the lengths of newborn male humans. (They can't stand yet, so we use length, but it's the same measure as height!) I have no doubt that σ_1 is much larger than σ_2; but, really, who is dumb enough to compare these two populations? Does one really need Statistics to know that college men are taller than newborn males?

This leads me to my second reason. When a researcher decides to compare populations by comparing means, then it is almost always the case that one is trying to find the population with the larger [smaller] mean because, if larger [smaller] responses are preferred, that population will be the better population, Let me introduce you to a hypothetical—and quite fanciful—example of what I mean.

Let's assume that we all agree that, *Life is good*: to die at age 50 is better than to die at age 40, and so on. Thus, suppose that, as Nature, you can decide between two possible distributions for the length of all persons' lives. Your two options are both Normal curves; the first population has

mean $\mu_1 = 70$ years and the second population has mean $\mu_2 = 68$ years. As Nature, you determine which population is better and decide that it will be the distribution for all people. What should you decide? Think about it.

Well, shame on you if you said, "Population 1 because it has a larger mean." You are not fit for the job of Nature! Your decision is too rash. Why do I say this? Because I have not told you the standard deviations of the two populations!

Now, suppose I told you that population 1 is the N(70,30) pdf and that population 2 is the N(68,1) pdf. Now, as Nature, which would you choose? I will now argue that the only sensible choice is population 2.

Indeed, I believe that population 1 would be horrible. It might even have a catastrophic impact on American society! With population 1, 16% of the people would die before age 40 and 16% would live past 100! (You think Social Security has financial problems now; 2.5% of population 1 would live past the age of 130!) By contrast, with population 2, 95% of the people would die between the ages of 66 and 70. (You have no doubt determined that I am old—64 at the time of this typing. Shouldn't my selfishness kick in and have me opt for population 1? No, for two reasons. First, Nature must be immortal and I am taking the role of Nature. Second, even though I enjoy being 64 much more than I imagined I would four decades ago, I really can't imagine that 115 will be loads of fun! Ideally, we all become like the Rutger Hauer character at the end of *Blade Runner*;

`http://www.youtube.com/watch?v=a_saUN4j7Gw`

and not like the character he played in *The Hitcher*—sorry, there is no appropriate link for this movie!)

The moral above is not restricted to Normal populations. If two populations have wildly different variances, then it might be the case that comparing means is not a good idea! Thus, in my mind, Case 3 solves a problem that is mathematically interesting, but that is not important, and indeed might be misleading, to a scientist. Note that this changes the way I want you to view Case 1. As you recall, Case 1 does not require symmetric congruent populations, only large sample sizes. But if—based on data or theory—you suspect that the two populations have very different spreads, think hard about whether you want to compare populations by comparing means.

19.5 Miscellaneous Results

This last section before the *Computing* section briefly introduces some useful ideas and methods.

19.5.1 Accuracy of Case 2 Confidence Levels

In this subsection I will address my decision not to show you Case 3 for Normal pdfs.

I performed three simulation experiments; scan the results in Table 19.1 and then read my description below of the experiments. For each rep, I had Minitab generate independent random samples of sizes $n_1 = n_2 = 20$. The first population is a N(μ, σ) pdf and the second population is a N($\mu, k\sigma$) pdf. In the first simulation, $k = 2$; in the second simulation, $k = 4$; and in the third

Table 19.1: Results from three simulation experiments. Each simulation had 10,000 reps, with a rep consisting independent samples of size $n_1 = n_2 = 20$ from two sequences of i.i.d. trials from a **Normal pdf**. For each sample, the 95% confidence interval estimate of $\mu_1 - \mu_2$ for Case 2 is computed and Nature classifies it as too small, too large or correct.

σ_2/σ_1	Number of Too Small Intervals	Number of Too Large Intervals	Number of Incorrect Intervals
2	230	241	471
4	278	272	550
8	277	274	551

simulation, $k = 8$. In words, the two Normal pdfs being compared are **not congruent**. In fact, the three simulations consider the situation in which the second population's standard deviation is two, four or eight times larger than the first population's standard deviation. These simulations are valid for any value of μ and any positive σ.

The simulation study shows that the actual confidence levels for Case 2 are equal or close to the nominal confidence levels, even though the assumption of congruence is violated. Indeed, for $k = 8$ the two populations are strongly not congruent, yet the Case 2 intervals perform as advertised. In fairness, I must state that if the study is unbalanced, $n_1 \neq n_2$, Case 2 might not perform as well. Moral: Try for a balanced study if possible.

19.5.2 Slutsky; Skewness

Recall that my guide is to use Slutsky's method, Case 1, if both sample sizes equal or exceed 30: $n_1 \geq 30$ and $n_2 \geq 30$. If you wonder why there is no mention of population skewness, keep reading.

I want to show you an important connection between Case 1 and Case 2. In Case 1, we replace the denominator of W in Equation 19.1 by:

$$\sqrt{(S_1^2/n_1) + (S_2^2/n_2)}.$$

In Case 2, we replace the denominator of W by:

$$S_p\sqrt{(1/n_1) + (1/n_2)}.$$

If the study is balanced, $n_1 = n_2$, then

$$S_p^2 = (S_1^2 + S_2^2)/2.$$

Some simple algebra (well, simple if one enjoys algebra; otherwise, it is tedious) shows that in the case of balance (and, hence, replacing n_2 by its equal, n_1):

$$\sqrt{(S_1^2/n_1) + (S_2^2/n_1)} = S_p\sqrt{(1/n_1) + (1/n_1)}.$$

Thus, in the case of balanced studies, the only difference between Cases 1 and 2 is that the former uses the N(0,1) pdf as its reference curve and the latter uses the t-curve with $df = n_1 + n_2 - 2$. Given my restriction on the use of Case 1, then we are comparing a N(0,1) pdf to a t-curve with at least 58 degrees of freedom; these curves are not very different.

Next, I want to look briefly at the issue of skewed populations.

Suppose that both populations 1 and 2 are the log-normal pdf with parameters 5 and 1, pictured in Figure 17.8 on page 449. In Table 17.10 you saw that Gosset's 95% confidence interval estimate of the mean performed very poorly by having too many incorrect intervals for this pdf and $n \leq 320$.

I will give you the results of **one** simulation experiment to explore how skewness effects our Case 2 inference. Note that while I don't want to mislead you, this is only one simulation experiment! We simply do not have time for a more in-depth study of this issue.

Each rep of my simulation experiment generated independent random samples of sizes $n_1 = n_2 = 20$ from two populations, both of which are the log-normal pdf with parameters 5 and 1. Because the populations are identical, $\mu_1 = \mu_2$ and a confidence interval for $\mu_1 - \mu_2$ will be correct if, and only if, it includes zero. Here are my results: 189 of the simulated 95% confidence intervals were too large; and 193 of the simulated 95% confidence intervals were too small. Thus, a total of $189 + 193 = 382$ intervals were incorrect; many fewer than the target of 500. Why did this happen? By taking a difference, $\bar{x} - \bar{y}$, the effect of skewness on a balanced study largely disappears. The *too few* incorrect intervals is the result of the intervals often being too wide, because the skewness effects the individual standard deviations (remember Figure 17.10 on page 455).

My general recommendation is that for a balanced study, Case 2 gives pretty good answers for populations that are not congruent Normal pdfs. The situation for unbalanced studies is much more complicated and I don't have time to present it to you. (Sorry.)

19.6 Computing

The *vassarstats* website that we have used previously is very helpful for this chapter.

19.6.1 Comparison of Means

Please go to:

`http://vassarstats.net`.

The left-side of the page lists a number of options; click on *t-Tests & Procedures*. This takes you to a new set of options; click on the top one, *Two-Sample t-Test for Independent or Correlated Samples*. This takes you to a new page. In the *Setup* section, click on *Independent Samples*. (If you forget to do this, it's no problem; *Independent Samples* is the default.) Next, enter the data, by typing or pasting, and click on *Calculate*.

The above is getting pretty abstract, so let's try this out with some real data. I will use Dawn's data on her cat Bob, which you learned about in Chapter 1. I entered:

1 3 4 5 5 6 6 6 7 8

(chicken responses) for Sample A and I entered

0 1 1 2 3 3 3 4 5 7

(tuna responses) for Sample B and clicked on *Calculate*. I will explain the output presented by *vassarstats*:

1. Under *Data Summary*, we find the sample sizes—both 10—and the means, $\bar{x} = 5.1$ and $\bar{y} = 2.9$. Sadly, the site gives us neither sample standard deviation, but we could obtain them from the entries for *SS*. (If you don't remember how, don't worry.)

2. Under *Results*, we find the value of $\bar{x} - \bar{y} = 2.2$; the observed value of the test statistic, $t = 2.4$, for Case 2; and the P-values for the alternative $>$ and $\neq$. I know that the one-tailed P-value is for the alternative that agrees with the data. Note that the P-value for $<$ is one minus the P-value for $>$.

3. You may safely ignore the information under *F-Test for* . . . , because we are not covering this topic.

4. You may safely ignore the information under *t-Test Assuming Unequal* . . . , because we are not covering this topic. If you can't resist looking, I will note that this is the Case 3 analysis. Note that the Case 3 analysis is nearly identical to the Case 2 analysis.

5. Finally, the bottom section presents the 95% and 99% confidence intervals for the separate means (you learned about this topic in Chapter 17) as well as the Cases 2 and 3 95% and 99% confidence intervals for $\mu_1 - \mu_2$.

19.7 Summary

In this chapter, we consider the problem of comparing two populations with numerical responses. We assume that we have i.i.d. random variables from each population and we assume that the two samples are independent. First, I will consider the problem of estimating the difference of population means, $\mu_1 - \mu_2$.

Case 1 is Slutsky's (large sample) approximate method. The confidence interval estimate of $\mu_1 - \mu_2$ is given in Formula 19.3, which is reproduced below:

$$(\bar{x} - \bar{y}) \pm z^* \sqrt{(s_1^2/n_1) + (s_2^2/n_2)}.$$

My advice is that this formula works well provided $n_1 \geq 30$ and $n_2 \geq 30$. In theory, Slutsky's confidence interval makes no assumptions about the two populations being compared, but note my remarks in this chapter on the issue of populations with extremely different spreads.

Case 2 assumes that the two populations are congruent Normal pdfs. This case yields Gosset's confidence interval estimate of $\mu_1 - \mu_2$, given in Formula 19.9 and reproduced below:

$$(\bar{x} - \bar{y}) \pm t^* s_p \sqrt{(1/n_1) + (1/n_2)}.$$

Recall that s_p^2 is defined in Equation 19.7 on page 500. If the Case 2 assumptions are true, then the confidence level of this interval is exact. Otherwise, this formula works well for populations that are not Normal curves unless the population variances are very different.

Case 3 is Case 2 without the assumption that the Normal curves are congruent. You are not responsible for this case; indeed, I don't even show it to you! I argue why this case is rarely useful at best, and potentially misleading at worst.

Next, I will talk about tests of hypotheses for comparing the population means. The null hypothesis is $\mu_1 = \mu_2$, and there are three options for the alternative:

$$H_1\colon \mu_1 > \mu_2; \quad H_1\colon \mu_1 < \mu_2; \quad \text{or} \quad H_1\colon \mu_1 \neq \mu_2.$$

For Case 1, the observed value of the test statistic is given in Equation 19.5, and is reproduced below:

$$z = \frac{(\bar{x} - \bar{y})}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}}.$$

The rules for using z to find the approximate P-value is given in Result 19.2.

For Case 2, the observed value of the test statistic is given in Equation 19.11, and is reproduced below:

$$t = \frac{\bar{x} - \bar{y}}{s_p\sqrt{(1/n_1) + (1/n_2)}}.$$

The rules for using t to find the P-value is given in Result 19.5.

19.8 Practice Problems

1. Earlier in these notes I told you about my friends Bert and Walt playing mahjong. I mentioned that the version Bert plays is easier than the version Walt plays. In particular, with Bert's version the beginning arrangement of tiles is selected at random from arrangements for which it is *possible to win*. By contrast, Walt's version begins with a random arrangement of tiles. (It is indisputable that for many arrangements, winning is impossible.) Thus, it is no surprise that Bert has a higher probability of winning than Walt. I want, however, to explore a different question: When they both lose, who does better, Bert or Walt?

   Let population 1 denote Bert's score and let population 2 denote Walt's score. In Chapter 17, I gave you the following summary statistics:

   $$\bar{x} = 23.93, s_1 = 11.33, n_1 = 71, \bar{y} = 23.05, s_2 = 10.74 \text{ and } n_2 = 216.$$

   Making our usual assumption of independent samples from two sequences of i.i.d. trials, perform the following analyses.

   (a) Compare the values—by taking a ratio—of the two sample standard deviations. Comment.

   (b) Calculate Slutsky's approximate 95% confidence interval estimate of $\mu_1 - \mu_2$. Comment.

(c) Obtain Slutsky's approximate P-value for the alternative $\neq$. Comment.

2. I presented Reggie's study of darts in the Chapter 1 Homework on page 25. Summary statistics for Reggie's data are below:

$$\bar{x} = 201.53, s_1 = 11.199, \bar{y} = 188.00, s_2 = 15.104 \text{ and } n_1 = n_2 = 15.$$

Make the usual assumptions of this chapter to analyze Reggie's data.

(a) Compare the values—by taking a ratio—of the two sample standard deviations. Comment.

(b) Calculate the values of s_p^2 and s_p.

(c) Calculate the Case 2 (Gosset's) 95% confidence interval estimate of $\mu_1 - \mu_2$. Comment.

(d) Obtain the Case 2 P-value for the alternative $>$. Comment.

3. In this chapter, I showed you that if two the populations are identical and log-normal with parameters 5 and 1, then Gosset's confidence interval works reasonably well for $n_1 = n_2 = 20$. This example shows that if the populations are skewed and **different**, then Gosset might not work so well. In particular, I let population 1 be the exponential pdf with rate equal to 0.1 (mean equal to 10) and I let population 2 be the exponential pdf with rate equal to 0.2 (mean equal to 5). Thus, the true value of $\mu_1 - \mu_2$ is $10 - 5 = 5$. Also, in addition to the two populations being strongly skewed, they have different variances: $\sigma_1^2 = 100$ and $\sigma_2^2 = 25$.

I performed a simulation experiment with 10,000 reps. Each rep consisted of:

- Selecting a random sample of size $n_1 = 20$ from population 1.
- Selecting a random sample of size $n_2 = 20$ from population 2.
- The two samples are independent.
- Gosset's Case 2 95% confidence interval estimate of $\mu_1 - \mu_2$ is obtained.
- Nature (well, me) determines whether the interval estimate is too large, too small or correct.

I obtained the following results: 107 intervals were too large; and 528 intervals were too small. Comment on these results.

19.9 Solutions to Practice Problems

1. (a) The ratio of the larger to the smaller is

$$11.33/10.74 = 1.055.$$

The sample standard deviations are nearly identical.

(b) The confidence interval is:

$$(23.93 - 23.05) \pm 1.96\sqrt{(11.33)^2/71 + (10.74)^2/216} = 0.88 \pm 1.96(1.5304) =$$
$$0.88 \pm 3.00 = [-2.12, 3.88].$$

The interval is inconclusive and quite wide compared to the point estimate of the difference of means.

(c) The observed value of the test statistic is

$$z = \frac{0.88}{1.5304} = 0.575.$$

The area under the N(0,1) pdf to the right of 0.575 is 0.2826. Thus, the approximate P-value for the alternative $\neq$ is $2(0.2826) = 0.5652$. The evidence in support of the alternative is weak.

2. (a) The ratio of the larger to the smaller is

$$15.104/11.199 = 1.349.$$

This is a the largest ratio we have seen, but it is still quite small.

(b) Because the study is balanced,

$$s_p^2 = \frac{(11.199)^2 + (15.104)^2}{2} = \frac{353.548}{2} = 176.774.$$

Thus, $s_p = \sqrt{176.774} = 13.296$.

(c) You can verify that with $df = 15 + 15 - 2 = 28$, $t^* = 2.048$. Thus, Gosset's 95% confidence interval estimate is

$$(201.53 - 188.00) \pm 2.048(13.296)\sqrt{2/15} = 13.53 \pm 2.048(4.855) =$$
$$13.53 \pm 9.94 = [4.59, 23.47].$$

This interval indicates that Reggie's population mean score from 10 feet is between 4.59 and 23.47 points larger than his population mean score from 12 feet.

(d) The observed value of the test statistic is

$$t = \frac{13.53}{4.855} = 2.787.$$

The area under the t-curve with $df = 28$ to the right of 2.787 is 0.0047; this is the P-value for the alternative $>$. It is exact if the populations are Normal pdfs; otherwise, it is approximate.

The evidence in support of the alternative is strong.

3. There are two disappointing results. First, the number of incorrect intervals is $107 + 528 = 635$ is clearly larger than the target value of 500. Not horribly larger, but quite a bit. Second, I am always disappointed when one type of incorrect interval greatly outnumbers the other type. Following the ideas from one population inference, I would not use Case 2 for a one-sided alternative for this situation.

19.10 Homework Problems

1. Recall Sara's study of golf, introduced in Chapter 2. Let population 1 denote the distance, in yards, Sara hit the ball with the 3-Wood and let population 2 denote the distance, in yards, Sara hit the ball with the 3-Iron. I presented the following summary statistics in Chapter 2:

   $$\bar{x} = 106.875, s_1 = 29.87, n_1 = 40, \bar{y} = 98.175, s_2 = 28.33 \text{ and } n_2 = 40.$$

   Making our usual assumption of independent samples from two sequences of i.i.d. trials, perform the following analyses.

   (a) Compare the values—by taking the ratio—of the two sample standard deviations. Comment.

   (b) Calculate Gosset's approximate 95% confidence interval estimate of $\mu_1 - \mu_2$. Comment.

   (c) Obtain Gosset's approximate P-value for the alternative $\neq$. Comment.

2. I introduced you to Dawn's study of her cat Bob in Chapter 1. Below are summary statistics for Dawn's data:

   $$\bar{x} = 5.1, s_1 = 2.025, \bar{y} = 2.9, s_2 = 2.079 \text{ and } n_1 = n_2 = 10.$$

   Make the usual assumptions of this chapter to analyze Dawn's data.

   (a) Compare the values—by taking a ratio—of the two sample standard deviations. Comment.

   (b) Calculate the values of s_p^2 and s_p.

   (c) Calculate the Case 2 (Gosset's) 95% confidence interval estimate of $\mu_1 - \mu_2$. Comment.

   (d) Obtain the Case 2 P-value for the alternative $>$; for the alternative $\neq$. Comment.

3. Please refer to Practice Problem 3. I did the same simulation experiment, but this time both populations were exponential with mean equal to 5. Thus, a correct confidence interval estimate will include $5 - 5 = 0$.

   I obtained the following results: 225 intervals were too large; and 268 intervals were too small. Comment on these results.

Chapter 20

Comparing Two Numerical Response Populations: Paired Data

This chapter is an extension of Chapter 16. In Chapter 16 we considered populations in which each population member or trial yields two dichotomous responses. In the current chapter each population member or trial yields two numbers. In other ways, however, this chapter also extends the work we did in Chapters 17–19.

20.1 Subject Reuse

I will introduce you to the idea of *subject reuse* with an artificial study of drug therapy for tension headaches. We will compare two different ways to design a study. I *cannot* use a real scientific study to make my comparisons because, to my knowledge, medical researchers select a design and use it. They do not investigate a medical issue twice, with two different designs, just to make me happy!

I am interested in studying drug therapies for a fairly mild health ailment, tension headaches. As you will see shortly, it is important that I have chosen an ailment that is both nonlethal and recurrent. I want to compare two drug therapies for the treatment of a tension headache. I will refer to the two therapies as drug A (treatment 1 and population 1) and drug B (treatment 2 and population 2).

We need a response that is a number. Each subject is given the following instructions:

> The next time you experience a tension headache, take the drug we have given to you. Wait 20 minutes. Write down your assessment of your pain on a scale from 0 (no pain) to 10 (worst pain ever).

How can I study this? Going all the way back to Chapter 1, I can use a completely randomized design. Following Chapter 19, I can perform population-based inference on the data I obtain from my completely randomized design. In particular, I can compare the mean of population 1 (drug A), μ_1, to the mean of population 2 (drug B), μ_2. I can estimate $\mu_1 - \mu_2$ with confidence and test the

Table 20.1: Artificial data from a CRD on headache pain, sorted within each treatment.

Position:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Drug A:	2	2	3	3	4	4	5	5	6	6	7	7	8	8	9	9
Drug B:	0	1	2	2	3	3	3	4	4	4	5	5	6	7	7	8

null hypothesis that $\mu_1 = \mu_2$. In order to choose an alternative, we need more information about the drugs. Three scenarios come to mind, listed below:

- Drug A is a placebo and drug B supposedly is beneficial. In this situation, remembering that smaller responses are preferred to larger responses, my alternative would be $>$.
- Drug A is the extra-strength version of drug B. In this situation, my alternative would be $<$.
- Drugs A and B are different active drugs. In this situation, my alternative would be $\neq$.

Suppose now that I have 32 subjects available for study and I am willing to pretend that they are a random sample from my superpopulation of interest. I decide to use a balanced design. Thus, I will use the online randomizer to assign 16 subjects to each treatment.

The artificial data for my CRD on the 32 subjects is given in Table 20.1. The data have been separated by treatments and sorted within each treatment. You can verify the following values of summary statistics (or trust me if you don't need additional practice on these computations):

$$\bar{x} = 5.500, s_1 = 2.366, \bar{y} = 4.000, s_2 = 2.251 \text{ and } n_1 = n_2 = 16.$$

Next, I calculate

$$s_p^2 = \frac{(2.366)^2 + (2.251)^2}{2} = 5.3325 \text{ and } s_p = \sqrt{5.3325} = 2.309.$$

The 95% confidence interval estimate of $\mu_1 - \mu_2$ is (see Formula 19.9 on page 501):

$$(5.50 - 4.00) \pm 2.042(2.309)\sqrt{2/16} = 1.50 \pm 2.042(0.8164) = 1.50 \pm 1.67 = [-0.17, 3.17].$$

This interval is inconclusive because it contains both positive and negative numbers. For future reference, note that the half-width of this interval is 1.67.

For a test of hypotheses, from Equation 19.11 on page 502, the observed value of the test statistic is

$$t = 1.50/0.8164 = 1.837.$$

With the help of our website calculator,

`http://stattrek.com/online-calculator/t-distribution.aspx`,

we find that the area under the t-curve with $df = 16 + 16 - 2 = 30$ to the right of 1.837 is equal to 0.0381. Thus, the approximate P-value for the alternative $>$ is 0.0381 and the approximate P-value for the alternative $\neq$ is $2(0.0381) = 0.0762$.

Let's look at the data in Table 20.1 again. In the drug A row, two subjects gave a response of 2—not much pain—and two gave a response of 9—a great deal of pain. In words, for drug A there is a large amount of subject-to-subject variation. The same is true for drug B. The idea behind the **randomized pairs design** (RPD) is to attempt to reduce this subject-to-subject variation.

I mentioned above that it is important that tension headaches are nonlethal and recurrent. Recurrence is important because if each subject has a headache (which is necessary in the CRD for us to obtain a response from each subject) then the subject will have a second headache. The RPD we learn about below will use responses from two headaches per subject, compared to the CRD which looked at one headache per subject. Nonlethal is important because—and I don't mean to be insensitive—in order to have a second headache the subject must survive the first one.

Admittedly, I am ignoring studies that would involve looking at 3, 4, 5 or more headaches per subject. I must draw the line somewhere!

You can now see the reason for the term *subject reuse*. We *reuse* each subject and, thus, obtain two responses per subject. And, somewhat obviously, because our goal is to *compare* the two treatments, for each subject we obtain a response from both treatments. Thus, for example, subject Sally gives us two numbers: her pain with drug A and her pain with drug B.

My next step is to provide you with artificial headache pain data from an RPD. My goal is to compare my RPD to my CRD for the artificial headache pain study. What is a fair way to do this? Well, my CRD had 32 subjects, with one response per subject, yielding a total of 32 observations. I could have 32 subjects in my RPD, but that would yield $32 \times 2 = 64$ observations. This strikes me as an unfair comparison. Thus, instead, my RPD below has only 16 subjects; with each subject giving two responses, I will have a total of $16 \times 2 = 32$ observations, the same as I had in my CRD. In fact, my RPD has exactly the same 32 observations as my CRD did. The data for my RPD is given in Table 20.2. Let's take a moment to make sure we can read this table correctly.

I have 16 subjects in my RPD and they have been labeled, for ease of reference, in the first row. If you compare the Drug A row of Table 20.2 with the Drug A row of Table 20.1, you can easily verify (because of the sorting in both tables) that the 16 responses to drug A are the same for the two data sets. You can also verify that the 16 responses to drug B are the same for the two data sets, but it's a lot easier to trust me on this! Let's look at subjects labeled 1 and 16. Subject 1 gives small responses (2 and 3) for both drugs, and subject 16 gives large responses (9 and 7) for both drugs. In words, subjects 1 and 16 vary a great deal in their responses to drug A and they vary a great deal in their responses to drug B.

Table 20.2 contains a row of numbers unlike any we have seen previously. For each subject I have calculated the difference in the subject's responses: response to A minus response to B. In symbols, the difference d is equal to $x - y$ for each subject. Let's look at subjects 1 and 16 again, but now let's look at their values of d. For subject 1, $d = -1$, and for subject 16, $d = 2$. Remembering that smaller values of x and y are better, a negative value of d indicates that the subject responded better to A than to B, whereas a positive value of d indicates that the subject responded better to B than to A. The d's for subjects 1 and 16 are much closer to each other (a

Table 20.2: First set of artificial data from an RPD on headache pain.

	Subject															
Drug	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
A(x)	2	2	3	3	4	4	5	5	6	6	7	7	8	8	9	9
B(y)	3	0	4	3	5	1	2	2	4	3	4	7	5	6	8	7
Difference(d)	-1	2	-1	0	-1	3	3	3	2	3	3	0	3	2	1	2

difference of 3) than either their values on A (a difference of 7) or B (a difference of 4). Thus, at least for these two subjects, there is less subject-to-subject variation on d than on both x and y. In fact, I reported earlier that the standard deviations of the x's and y's are, respectively, 2.366 and 2.251. By comparison, you may verify that the standard deviation of the d's is 1.592. Thus, by examining the three standard deviations we arrive at the same conclusion we have from looking at subjects 1 and 16: the subject-to-subject variation is smaller for the differences than it is for both drugs.

Let's gather together our various summary statistics:

Source	Symbol	Mean	Standard deviation
Treatment 1	x	5.50	2.366
Treatment 2	y	4.00	2.251
Difference	d	1.50	1.592

Notice that

$$\bar{x} - \bar{y} = 5.50 - 4.00 = 1.50 = \bar{d}.$$

On reflection, we realize that this is true for any set of data; calculating two means and subtracting gives the same answer as first subtracting then finding the mean of the differences. It is obvious that this argument can be extended to an entire finite population. Let μ_d denote the mean of the population of differences. Then:

$$\mu_d = \mu_1 - \mu_2. \tag{20.1}$$

Equation 20.1 is also true for populations for trials. (The argument is a bit trickier; I recommend that you simply believe me.) This leads to the following very important realization:

- Inference for $\mu_1 - \mu_2$—estimation or testing—is equivalent to inference for μ_d.

In particular, if we are willing to assume that our X's are a random sample from a population, then our Y's are also a random sample from the same population, although the two samples (the X's and the Y's) are not independent samples. If we let

$$D_1, D_2, D_3 \ldots D_m$$

denote the random variables that yield the observed values

$$d_1, d_2, d_3 \ldots d_m,$$

then it also follows that the D's are a random sample from the population of differences. Because the D's are a random sample from a single population, the methods of Chapters 17 and 18 may be used to analyze them.

In particular, Gosset's confidence interval estimate of μ in Chapter 17, Formula 17.6, yields—after we change the symbols—the following result.

Result 20.1 *Gosset's confidence interval estimate of $\mu_d = \mu_1 - \mu_2$ is:*

$$\bar{d} \pm t^*(s_d/\sqrt{m}). \tag{20.2}$$

In this formula, $\bar{d}$ and s_d are the sample mean and standard deviation, respectively, of the differences. The number of pairs is denoted by m and the degrees of freedom for t^* is $(m-1)$.

I will illustrate the use of Formula 20.2 for our first set of artificial data from an RPD on headache pain, given in Table 20.2. For $df = m - 1 = 16 - 1 = 15$, you can verify that t^* for the 95% confidence level is 2.131. Thus, the 95% confidence interval estimate of $\mu_d = \mu_1 - \mu_2$ is

$$1.50 \pm 2.131(1.592/\sqrt{16}) = 1.50 \pm 2.131(0.398) = 1.50 \pm 0.85 = [0.65, 2.35].$$

This interval **is conclusive**; the mean pain on drug A is between 0.65 and 2.35 units larger the mean pain on drug B.

Recall that when we had **exactly the same data** from a CRD, the confidence interval estimate was:

$$1.50 \pm 1.67.$$

Thus, the confidence interval is—approximately—one-half as wide for the RPD as it is for the CRD. Subject reuse is effective! More accurately, I have given you artificial data that made subject reuse effective.

Recall also that, as a very rough guide, we must quadruple the number of subjects to reduce the half-width of a confidence interval by a factor of two. (This is rough because with more data the value of t^* will definitely be reduced and the various sample standard deviations will likely change.) Thus, very roughly, I would need $4 \times 32 = 128$ subjects on a CRD to obtain the same precision that I get from an RPD with 16 subjects! As the expression goes, "Work smarter, not harder!"

We can also perform a test of hypotheses for data from an RPD. For the null hypothesis that $\mu_d = 0$, we rewrite the observed value of the test statistic—using the notation of this chapter—given in Equation 18.2 on page 470:

$$t = \frac{\bar{d}}{s_d/\sqrt{m}} = \sqrt{m}(\bar{d}/s_d). \tag{20.3}$$

The three rules for finding the P-value are summarized in the following result.

Result 20.2 *For the null hypothesis that $\mu_d = 0$, and t given in Equation 20.3, the rules for finding the P-value are below. In these rules, areas are computed under the t-curve with $df = m - 1$.*

1. *For the alternative $\mu_d > 0$, the P-value equals the area to the right of t. Equivalently, the P-value equals the area to the left of $-t$.*

2. *For the alternative $\mu_d < 0$, the P-value equals the area to the left of t. Equivalently, the P-value equals the area to the right of $-t$.*

3. *For the alternative $\mu_d \neq 0$, the P-value equals twice the area to the right of $|t|$. Equivalently, the P-value equals twice the area to the left of $-|t|$.*

In the above result, if the population of differences is a Normal pdf, then the P-values are exact; otherwise, they are approximations and the comments from Chapters 17 and 18 regarding their accuracy are relevant.

For the data in Table 20.2, the observed value of the test statistic is

$$t = \sqrt{16}(1.50/1.592) = 3.769.$$

Using the website,

`http://stattrek.com/online-calculator/t-distribution.aspx`,

we find that the area under the t-curve with $df = m - 1 = 16 - 1 = 15$ to the right of 3.769 equals 0.0009. Thus, the approximate P-value for $>$ is 0.0009 and the approximate P-value for $\neq$ is $2(0.0009) = 0.0018$. For comparison, the P-value for $>$ for a CRD with the same responses was shown earlier to equal 0.0381. Thus, for alternative $>$ or $\neq$, the approximate P-value from the RPD value is more than 38 times smaller than the approximate P-value from the CRD!

Think about the question: Have I convinced you that an RPD is better than a CRD for a study of headache pain? I hope not; all of my data are artificial. What I **have shown** you is that it is **possible** that an RPD can be better than a CRD. In the name of basic fairness, I should show you that the opposite also can be true.

Table 20.3 provides a second set of artificial data for an RPD on headache pain. As with Table 20.2, the data in Table 20.3 are the same responses for both drugs as given in the original CRD, Table 20.1. For these new data, it can be shown that $\bar{d} = 1.50$ and $s_d = 3.162$. The first of these summaries is no surprise; because the x's and y's have not changed, $\bar{x} = 5.50$, $\bar{y} = 4.00$ and, perforce, $\bar{d} = \bar{x} - \bar{y} = 1.50$. Note, however, that for these new data, s_d is much larger than both s_1 and s_2 and, hence, s_p too.

I will evaluate Formula 20.2 with these new data. Gosset's 95% confidence interval estimate of μ_d is:

$$1.50 \pm 2.131(3.162/\sqrt{16}) = 1.50 \pm 2.131(0.7905) = 1.50 \pm 1.68 = [-0.18, 3.18].$$

Notice that the half-width of this new interval, 1.68, is larger, but only slightly larger, than the half-width of the interval for the CRD data, 1.67.

We have seen that for the first set of artificial data, the RPD gives a much more sensitive analysis than the CRD. For the second set of artificial data, however, the analysis is virtually the same for the RPD and CRD. What is it about these data sets that is causing this difference? Well,

Table 20.3: Second set of artificial data from an RPD on headache pain.

	Subject															
Drug	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
A(x)	2	2	3	3	4	4	5	5	6	6	7	7	8	8	9	9
B(y)	8	3	2	4	2	4	7	0	5	5	3	1	3	7	6	4
Difference(d)	-6	-1	1	-1	2	0	-2	5	1	1	4	6	5	1	3	5

the simple answer is that the second set has a much larger value of s_d than the first data set. In particular, the half-width of the confidence interval for μ_d is

$$t^*(s_d/\sqrt{m});$$

clearly, as s_d increases, the half-width increases and the analysis becomes less sensitive.

I am not really satisfied with the above answer. The value of s_d is somewhat of a mystery; what makes it larger or smaller? It turns out that it is possible to understand better what is happening if we draw a picture of the data.

20.2 The Scatterplot

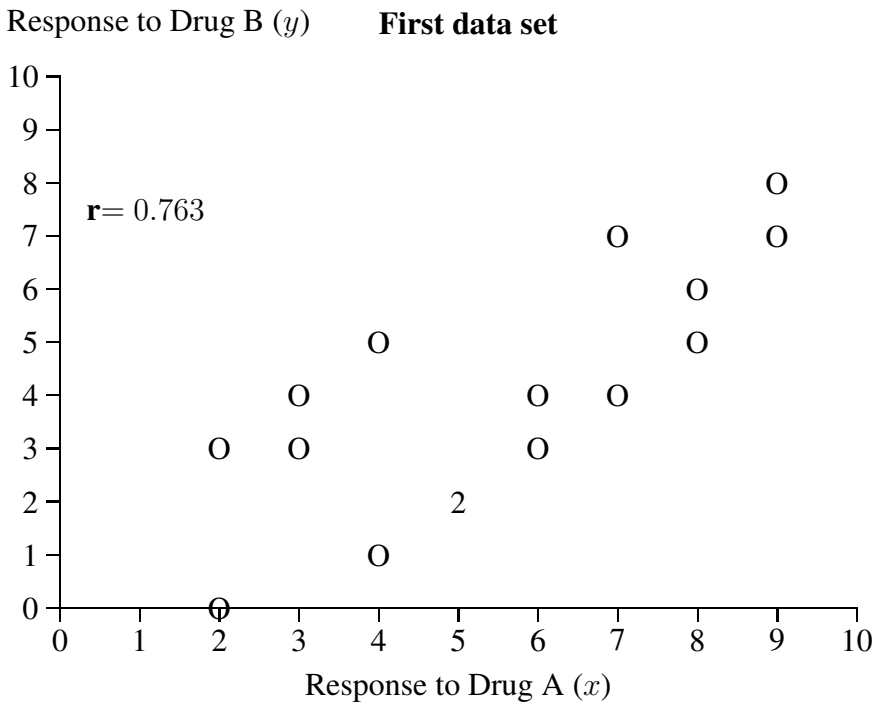
I want to introduce you to a very important picture in Statistics, one I actually used—without much explanation—in Figure 17.9 on page 454 and Figure 17.10 on page 455. Figure 20.1 presents two scatterplots, one for each our sets of artificial data from an RPD. Let's take a few minutes to examine these pictures. Let's look at the top picture, the scatterplot of y versus x for the first set of artificial data from an RPD on headache pain.

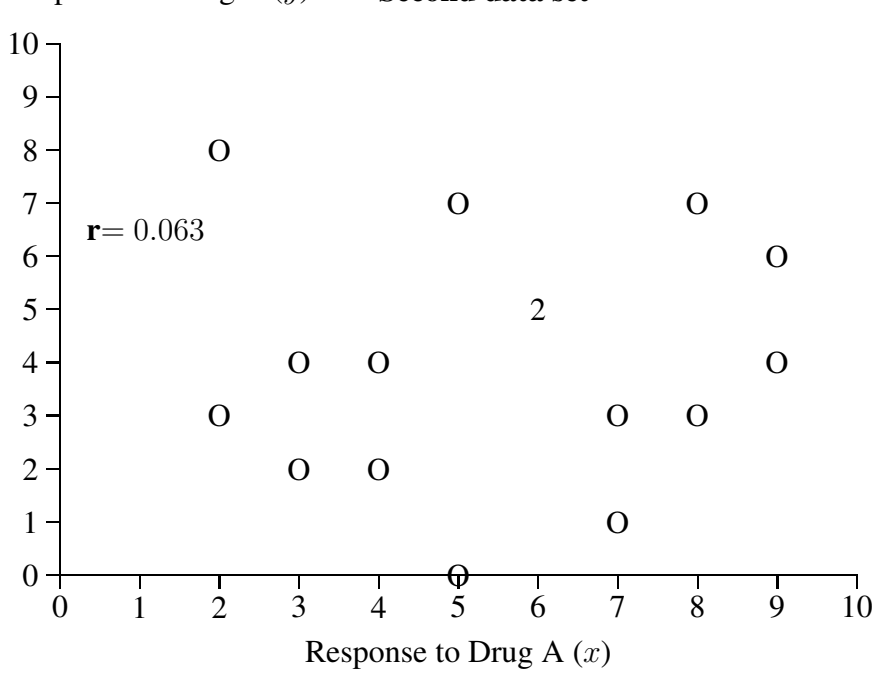
The scatterplot begins with the familiar coordinate system from childhood, with the vertical axis corresponding to y (response to Drug B in this plot) and the horizontal axis corresponding to x (response to Drug A in this plot). We will refer to this as a plot of y versus x. Next, both axes are given scales that are sufficient to include the entire plot, as described below. I have given each axis the values 0 through 10.

At this point you may have noticed something untoward about my scatterplot: even though both axes measure the same *thing*—subjective assessment of pain—I have used different scales on the two axes. In particular, in my picture the x values are stretched out a bit compared to the y values. Why did I do this? I give you two reasons.

1. Scatterplots are used extensively in Statistics, most notably in **regression analysis**, which we will study in the next two chapters of these notes. Overwhelmingly the norm in these applications, especially regression, is that the y and x features are like *apples and oranges*; i.e. there is no natural relationship between the features. For example, if a unit is an adult male human, then x could be his height in inches and y could be his weight in pounds. There is neither a natural nor obvious way to *choose the same scale* for y and x. As a result, when statisticians choose scales we mostly are concerned with the next item.

Figure 20.1: Scatterplots of the response to drug B versus the response to drug A for the 16 Subjects in the two artificial data sets from RPDs on headache pain.





2. Statisticians generally prefer scatterplots for which the *width* is greater than its *height*. Such a picture is deemed to be more aesthetically pleasing than a square. (If you are interested in this topic, see the Wikipedian entry for the *golden rectangle*:

   `http://en.wikipedia.org/wiki/Golden_rectangle`

   or read *The Da Vinci Code*!)

My scatterplot displays the pair of values x and y for each of the 16 subjects in my RPD. For example, consider subject 1; its values are $x = 2$ and $y = 3$. Subject 1 appears in the scatterplot as an 'O' at the location (ordered pair) $(x, y) = (2, 3)$. (Make sure that you can locate this 'O.') Thus, because there are 16 subjects in my RPD, there are 16 O's in Figure 20.1. Except that there aren't; there are actually only 14 O's in the scatterplot and a numeral '2.' The numeral 2 is located at $(x, y) = (5, 2)$; it indicates that there should be two O's at this point because subjects 7 and 8 both had $x = 5$ and $y = 2$.

Now, look at the scatterplot. What do you see? We will look at many scatterplots when we study regression, so I am going to keep this brief. First, the relationship between x and y is **not deterministic**; it is **not a mathematical function**—the value of x does **not determine** the value of y (nor does the value of y determine the value of x). I mention this because, I conjecture, you have had a great deal of experience with deterministic relationships in your various mathematics classes. In Statistics, however, we almost always study relationships that are not deterministic.

As a statistician I look at the scatterplot in Figure 20.1 and I see two main features: the relationship between x and y is *increasing* and it looks *linear*. *Increasing* is self-explanatory. But if it's not: as we move from unit to unit in a way in which the values of x are increasing—in every day language, scan the scatterplot from left to right—the values of y tend to become larger. *Linear* is my subjective assessment that the pattern in the scatterplot can be described reasonably well by a straight line and **does not require** a curve to describe it. Again, let me remind you that we will consider these issues in greater detail when we study regression later in these notes.

Whenever the relationship between x and y looks linear it is reasonable to summarize the relationship by calculating the **correlation coefficient**, denoted by r. Figure 20.1 tells us that the correlation coefficient of its first scatterplot is $r = 0.763$. We will learn a great deal about the correlation coefficient when we study regression. Let me just say for now that the possible values of the correlation coefficient fall between -1 and 1 inclusive: $-1 \leq r \leq 1$. Also, an increasing [decreasing] relationship between x and y makes r positive [negative]. For the purpose of an RPD, the correlation coefficient plays a role in a mathematical relationship that exists between our three standard deviations (for the x's, for the y's and for the d's), The relationship is

$$s_d^2 = s_1^2 + s_2^2 - 2rs_1s_2. \tag{20.4}$$

This equation can be illustrated with the values of the three standard deviations and $r = 0.763$:

$$s_d^2 = (1.592)^2 = 2.534464, \text{ and}$$

$$s_1^2 + s_2^2 - 2rs_1s_2 = (2.366)^2 + (2.251)^2 - 2(0.763)(2.366)(2.251) =$$

$$5.597956 + 5.067001 - 8.127272 = 2.537685,$$

which are the same, except for round-off error.

We saw earlier that as the value of s_d increases the half-width of the confidence interval for μ_d also increases. As a result, as s_d increases, subject reuse becomes less useful. We can see from Equation 20.4 that as the value of r increases, the value of s_d^2 and, hence s_d, decreases. **Thus, the effectiveness of subject reuse grows with the value of r.** This is important because as you learn more about how r relates to a scatterplot, you will be better at deciding whether subject reuse is effective.

For example, the second scatterplot in Figure 20.1 is for our second set of artificial data from an RPD. In this picture I see only a very weak increasing relationship between x and y. My visual assessment agrees with the value of $r = 0.063$ which is barely larger than 0. (Again, we will learn more about this in the next chapter.)

Let's now go back in time to before we collected our data. Imagine that I am a researcher who knows a great deal about headache pain. I know that if my scatterplot of values of x and y looks like the second scatterplot in Figure 20.1, then pairing won't be any better than a CRD. If, indeed, my scatterplot provides a smaller value of r—including negative values—then pairing is less effective than a CRD. If, however, my scatterplot yields an r substantially larger than 0.063, then pairing would be effective, possibly extremely effective. Based on my expertise as a headache pain researcher, I am convinced that there will be an increasing relationship between x and y and that the relationship will be substantially stronger than one that yields $r = 0.063$. (Does this make sense to you **medically**? Why or why not?) Thus, given my expert opinion, I would definitely opt for pairing over independent samples.

20.3 Putting the 'R' in RPD

I have talked (well, keyboarded) a great deal about the 'P' in an RPD, but have said nothing about the 'R;' I will do so now.

Randomization occurs at each pair in an RPD. In general, let m denote the number of pairs in an RPD. This means that the data will consist of m values each of x's, y's and d's. For my two headache RPDs, $m = 16$. At each pair there are two choices for the assignment of treatments to members of the pair; they are:

- Assign the first member of the pair to treatment 1 and the second member of the pair to treatment 2. We denote this possibility as 1.
- Assign the first member of the pair to treatment 2 and the second member of the pair to treatment 1. We denote this possibility as 2.

Thus, at each pair our *randomizer* must give us either a 1 or a 2, with these options being equally likely to occur. Also, the decisions at different pairs must be statistically independent.

There are a number of physical devises that will allow us to randomize for an RPD. Instead, I will focus on an electronic method using the randomizer we learned about in Chapter 2. We begin by going to the website

`http://www.randomizer.org/form.htm`.

This site asks you to provide input information. I will walk you through the choices.

- The first question is: **How many sets of numbers do you want to generate?**
  We want an assignment for one RPD; thus, leave it at the default value of 1.
- The second question is: **How many numbers per set?**
  Enter m which equals 16 for our headache pain study.
- Next, you need to specify: **Number range.**
  For an RPD, this will always be from 1 to 2.
- Next, we have another question: **Do you wish each number in a set to remain unique?**
  Answer: No.
- Next, we have another question: **Do you wish to sort the numbers that are generated?**
  Answer: No.
- You may ignore the final question; i.e. we are happy with the default response.
- You are now ready to click on the box: **Randomize Now!**

I operated our randomizer with the choices above and obtained:

Pairs			
1–4	5–8	9–12	13–16
2,1,1,1	2,1,1,2	2,2,1,1	2,1,2,1

(I have added some headings and spacings above to make the string of 1's and 2's easier to read.) In particular, we see that

- In pairs (subjects) 2, 3, 4, 6, 7, 11, 12, 14 and 16 treatment 1 (Drug A) is assigned to the first headache and treatment 2 (Drug B) is assigned to the second headache.
- In pairs (subjects) 1, 5, 8, 9, 10, 13 and 15 treatment 2 (Drug B) is assigned to the first headache and treatment 1 (Drug A) is assigned to the second headache.

Let me say a bit about why we randomize the order of the treatments within each pair. There are two main reasons:

1. If we performed randomization-based inference—we won't because of time limitations—the process of randomization becomes the basis for our inference; in particular, the P-value is obtained by looking at all possible assignments in an RPD.
2. For scientific *validity*.

Regarding this second reason: As an honorable scientist you strive to learn things that are, indeed, true. But you also want the *scientific community* to take your work seriously.

For example, you might decide that randomizing is silly and a waste of effort. Instead you decide to have every subject take treatment 1 first and then treatment 2. At a personal level this might lead you to conclusions that are false. As a global matter I would be amazed if the *scientific community* paid much attention to your conclusions. The issue is that there *might* be an **order effect** in your study. What do I mean by this?

Imagine a situation in which all, or nearly all, subjects would give a lower response to their first headache than to their second headache, **even if the pain levels were, indeed, identical**. Or imagine the opposite pattern, where responses are systematically lower on the second headache compared to the first. In either of these situations, a decision to *always give treatment 1 first* would bias the study. The possibility of such an order effect causes the *scientific community* to discount, or even ignore, your findings.

Let's look at the randomization I obtained above for the headache RPD. In nine pairs drug A is taken before drug B, and in only seven pairs drug B is taken before drug A. Thus, if there is indeed an order effect, one of the drugs (I can't tell which one without knowing the direction of the order effect) has a slight advantage over the other.

There is available to a researcher a design that is a bit more complicated than an RPD. It is called the **crossover design** and it has two features that are not present in an RPD:

1. A crossover design forces balance between what I earlier called '1' and '2.' More precisely, the number of pairs that have treatment 1 first (what I called '1') is exactly equal to the number of pairs that have treatment 2 first (what I called '2'). Thus, unlike my RPD above which had nine 1's and seven 2's, the crossover design would, perforce, have eight of each. This makes obvious a modest limitation on a crossover design: the number of pairs must be an even number.

2. For a crossover design, the analysis of the data explicitly incorporates—and estimates—the order effect, as compared to an RPD—the population-based method is given above—that ignores a possible order effect in the analysis. As a result, the analysis of a crossover design is more complicated than the analysis of an RPD. If, indeed, there is a large order effect (admittedly, large is vague here) then a crossover design can be more powerful than the corresponding RPD. Sadly, because these notes cannot cover every topic in Statistics, I will not show you how to analyze data from a crossover design.

20.4 Other Ways to Form Pairs

Thus far, I have discussed subject reuse as the only way to obtain paired data. Other methods are possible. I am going to be very cautious in my presentation of this material.

1. I will show you a situation other than subject reuse for which pairing is valid.

2. I will show you a situation other than subject reuse for which pairing gives wildly invalid results.

I will give you a rule that helps distinguish between these situations, but there will be holes in my rule; i.e., my rule does not necessarily cover every situation that could arise in science. Why? My standard reason: we cannot cover everything in a one semester course.

I begin with a situation in which pairing is valid.

20.4.1 Forming Pairs from Adjacent Trials

Let's return to the game of Tetris. I want to focus (again) on an entire game as a trial. Many years ago, I enjoyed playing Tetris My game had a feature that allowed the player to see or not see the next shape while manipulating the current shape. (Seeing was the default.) It seemed to me that selecting the default, preview, option would lead to much higher scores. So, I decided to collect data to investigate this matter.

A game is a trial and the response is the number of lines I completed before the game ended. I decided to perform 20 trials, with 10 on each setting. I was very worried that fatigue or boredom would affect my later scores, so I formed pairs out of consecutive trials: 1 and 2; 3 and 4; and so on. I will slow down and present these ideas carefully. Please refer to Table 20.4.

Find the rows that begin with *Trial*. The first such row lists trials 1–10 and the second such row lists trials 11–20. I would prefer it if these 20 trials were physically all in the same row, but our *paper* isn't wide enough.

Find trials 1 and 2; in the row immediately above, these trials are identified as the trials that form pair 1. Next, you see that trials 3 and 4 form pair 2; trials 5 and 6 form pair 3; and so on; and trials 19 and 20 form pair 10.

Next, I went to the randomizer—details not shown—and it gave me the following assignment:

Pair:	1	2	3	4	5	6	7	8	9	10
Randomizer gives:	2	1	1	1	1	2	1	2	1	2

The randomizer gave '1' to pairs 2–5, 7 and 9. This means that within these pairs, the first game was played on treatment 1 (preview) and the second game was played on treatment 2 (no preview). The randomizer gave '2' to pairs 1, 6, 8 and 10. This means that within these pairs, the first game was played on treatment 2 (no preview) and the second game was played on treatment 1 (preview). This explanation I have just given can be seen in the *Treatment* rows of Table 20.4.

After all of this work, it was time for me to have fun! I finally was able to play my 20 games of Tetris. In the first game, I set the machine to *no preview*—treatment 2—and obtained a score of 84. In the second game, I set the machine to *preview*—treatment 1—and obtained a score of 106. And so on, as displayed in the *Response* rows of Table 20.4.

Table 20.4 provides an accurate description of how my data were collected, but it needs to be rewritten to facilitate a statistical analysis. Table 20.5 rewrites my data in a form that is ready for analysis. (You should check to make sure you *understand* how I used Table 20.4 to create Table 20.5. You don't need to *check every entry*, just make sure that you understand the process.)

Not surprisingly, and obviously from even a quick glance at the data, I was a much better player with the preview option. It is not so clear that pairing was beneficial; we shall explore this issue below.

Table 20.4: The RPD to compare the preview and no preview options in Tetris.

Pair:	1		2		3		4		5	
Trial:	1	2	3	4	5	6	7	8	9	10
Treatment:	2	1	1	2	1	2	1	2	1	2
Response:	84	106	112	93	118	86	102	86	112	94

Pair:	6		7		8		9		10	
Trial:	11	12	13	14	15	16	17	18	19	20
Treatment:	2	1	1	2	2	1	1	2	2	1
Response:	88	110	130	108	91	110	127	79	91	138

Table 20.5: Paired data to compare the preview and no preview options in Tetris.

	Pair									
Treatment	1	2	3	4	5	6	7	8	9	10
1: Preview (x)	106	112	118	102	112	110	130	110	127	138
2: No preview (y)	84	93	86	86	94	88	108	91	79	91
Difference ($d = x - y$)	12	19	32	16	18	22	22	19	48	47

I calculated the following summary statistics:

$$\bar{x} = 116.5, s_1 = 11.56, \bar{y} = 90.0, s_2 = 7.77, \bar{d} = 26.5, s_d = 11.87 \text{ and } m = 10.$$

With $df = 9$, the value needed for the 95% confidence interval is $t^* = 2.262$. Thus, the 95% confidence interval for μ_d is

$$26.50 \pm 2.262(11.87/\sqrt{10}) = 26.50 \pm 2.262(3.754) = 26.50 \pm 8.49 = [18.01, 34.99].$$

At the 95% confidence level, my population mean score with the preview feature is between 18 and 35 lines larger than my population mean score without the preview feature.

I can also perform a test of hypotheses on my Tetris data. Using the *Inconceivable Paradigm*, I select $\mu_d > 0$ as my alternative. The observed value of the test statistic is

$$t = 26.50/3.754 = 7.059.$$

The area under the t-curve with $df = m - 1 = 9$ to the right of 7.059 is (using Minitab) 0.0000296, just smaller than 3 in one-hundred-thousand. This is the approximate P-value for $>$. Note that if the alternative had been $\neq$, then the approximate P-value would be twice as large, just smaller than 6 in one-hundred-thousand.

For comparison, we will now pretend that the data come from independent random samples. First,

$$s_p^2 = [(11.56)^2 + (7.77)^2]/2 = 97.00.$$

Thus, $s_p = \sqrt{97} = 9.85$. Moreover, for $df = 16 + 16 - 2 = 30$, we get $t^* = 2.101$. Thus, the 95% confidence interval for $\mu_1 - \mu_2$ is

$$(116.5 - 90.0) \pm 2.101(9.85)\sqrt{1/10 + 1/10} = 26.50 \pm 9.25 = [17.25, 35.75].$$

The half-width for the RPD interval, 8.49, is 8.2% smaller than the half-width, 9.25, for the pretend CRD. Thus, pairing seems to have been effective, but not as dramatically as it was in my first set of artificial data on headache pain.

20.4.2 When is it Valid to do a Paired Data Analysis?

The methods given above—confidence intervals and tests of hypotheses using the differences as data—are valid in the following situations. Of course, implicit is the notion that we are willing to assume we have i.i.d. random variables.

1. It is valid if the units (subjects or trials) are reused. Scientifically, it is much better if one is able to use randomization, but it's not necessary for statistical validity. In my experience, the most common type of unit reuse in an observational study is a before/after study.

2. If one forms pairs of units (trials or subjects) by matching different units based on some feature—often prognosis in medicine, then the analysis is valid in two situations:

   (a) Within each pair, units are assigned to treatments by randomization.

   (b) For an observational study on two finite populations of subjects, pairs are formed at the *population level*. If pairs are formed at the *sample level*, regardless of how, then a paired data analysis is invalid and, indeed, can be grossly misleading. The notions of *population level* and *sample level* are discussed below.

I will now provide some examples of the ideas listed above.

First, let's look at an example of a before and after study. Suppose we have $m = 50$ subjects who are interested in losing weight. A study might proceed as follows. Each person is weighed at the beginning of the study. Each person then follows a rigorous program of diet and exercise for, say, three months. at which time each person is weighed again. If x [y] is a subject's weight at the beginning [end] of the study, then $d = x - y$ is the amount of weight the subject lost during the study. (Keep in mind that $d < 0$ means that the subject's weight increased.)

Is this weight loss study really paired data?

- Yes, because we get two numbers from each subject and it is meaningful to calculate their difference.

- No, because we can view the data as one response, the difference in weights.

In my opinion, it does not matter which of these viewpoints you adopt, the data are analyzed the same way and the scientific interpretation is unchanged.

Too often in a before and after study, researchers forget the need for a control group, as illustrated in the following example. **Full disclosure:** I found this example years ago in a textbook on designing experiments in the social sciences. I don't have a reference; and I can't swear that the authors were being honest!

Anyways, in August, 1939, students at an American university were given a pre-test to measure their attitudes towards the government of Nazi Germany. Then they took a four-week course that presented that government in a positive light. At the end of the course, the students were given a post-test to determine the extent to which the course influenced the students' attitudes. There was, however, an unforeseen difficulty: On September 1, 1939, while the course was still in session, Germany invaded Poland, starting World War 2. As a result, I sincerely doubt that the differences between pre-test and post-test scores were due to the course! A control group would have improved this study greatly, but I suspect it was doomed in any event.

Regarding item 2(a) in our earlier list: forming pairs of different units and then randomizing the assignment of unit to treatment within each pair. I advocate this method for trials—as I demonstrate above for my Tetris study—but am not a fan of this method for subjects. For subjects, I believe it is better to form blocks of subjects, as I describe in Chapter 4 of my textbook, *Statistics: Learning in the Presence of Variation*. In addition, if you do form pairs this way, I believe that randomization-based inference is valid, but not population-based inference. Not everybody, however, agrees with me. Sadly, we have time for neither this topic nor a presentation on blocks.

The remainder of this subsection is devoted to item 2(b) in my list: forming pairs at the *population level* and *sample level*.

Are husbands taller than their wives? Are husbands older than their wives? Personally, I have never been interested in either of these questions, but I must admit that during my long life, I have heard many people talk about them. More pragmatically, I can't think of an example of pairing at the population level other than one involving husbands and wives.

First, a disclaimer. At the time of my typing these words, I live in Wisconsin, a state in which a legal marriage consists of exactly two people, one of each sex. The fact that my example is restricted to such pairs should not be interpreted in any way politically, etc.

Let's focus on height. There is a population of husbands in Wisconsin and there is a population of wives in Wisconsin. Let μ_1 denote the mean height of the husbands and μ_2 denote the mean height of the wives. My goal is to estimate $\mu_1 - \mu_2$. Here are two statistically valid ways for me to sample these populations:

1. I could select a random sample of m men from the population of husbands. I could select a random sample of m women from the population of wives. I would have my samples be independent of each other. I would determine the height of each of the $2m$ persons in my study.

2. I could select a random sample of m women from the population of wives. I would determine the height of each of the m women in my study as well as the heights of their husbands.

With the first method, I would analyze the data with the methods of Chapter 19, independent samples. For the second method, I would analyze the data with the methods of this chapter, paired data. Based on my many years of observation of married couples in Wisconsin, I conjecture that

there is a *pretty strong* positive correlation between the heights of husbands and wives; thus, I would use the second method of sampling. If my conjecture is correct, my paired data analysis will be more efficient than independent samples would have been. If my conjecture is wrong, my paired data will still be valid, but it won't be as efficient as independent samples would have been.

To summarize, I formed pairs of all members of the two populations—which, necessarily, needed to have exactly the same number of members. This is what I mean by forming pairs at the *population level*.

I end this material with an cautionary tale that, I hope, will convince you to never form pairs at the sample level, but, first, a story from my career.

For part of my career at the University of Wisconsin–Madison, part of my job was to provide statistical advice to graduate students from other departments. One day a student came to me with her data. She was willing to assume that she had independent random samples of size $n_1 = n_2 = 40$ from two populations. (Her advisor said that) She **needed to show** that the two populations had different means. She knew the methods of Chapter 19 and had applied them to her data. Sadly, following standard statistical reasoning, she could **not** conclude that the population means were different.

After she explained all of the above to me, we had the following conversation:

> BW: So, why are you here?
>
> Student: Somebody told me that if I had paired data I would get a smaller P-value and, thus, be able to make my advisor happy.
>
> BW: That might be true. Do you want to do a new study, one with paired data?
>
> Student: No, I want you to pair my data.
>
> BW: Huh? I don't understand.
>
> Student: (With exasperation) I want you to take my data, manipulate them into pairs to give me the answer I need.
>
> BW: Oh.

I have cited the above exchange many times in my teaching. I then point out that unlike physics, chemistry and mathematics, there is no *demon* in Statistics. What do I mean by this? Well, if you don't believe in physics, gravity might kill you. If you don't believe in chemistry, a mixture of ammonia and bleach might kill you. If you don't understand fractions, somebody might take all of your money by continually forcing you to *make change*. You can, however, perform any number of ridiculous statistical analyses and nothing bad will happen to you!

Suppose that you want to determine which university has taller men: UW–Madison or the University of Minnesota–Twin Cities. (I realize that this is silly; bear with me please.) You select independent random samples of sizes $n_1 = n_2 = 40$ from both populations. Denote your observed data from Wisconsin by:

$$x_1, x_2, x_3 \ldots x_{40}.$$

Similarly, denote your observed data from Minnesota by:

$$y_1, y_2, y_3 \ldots y_{40}.$$

Sort each set of data, yielding

$$x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq \ldots \leq x_{(40)} \text{ and}$$

$$y_{(1)} \leq y_{(2)} \leq y_{(3)} \leq \ldots \leq y_{(40)}.$$

Thus, for example, $x_{(1)}$ [$y_{(1)}$] is the height of the shortest of the 40 men in the Wisconsin [Minnesota] sample; $x_{(40)}$ [$y_{(40)}$] is the height of the tallest of the 40 men the Wisconsin [Minnesota] sample; and so on.

Next, we form pairs at the sample level: we match $x_{(1)}$ with $y_{(1)}$; $x_{(2)}$ with $y_{(2)}$; and so on; and we match $x_{(40)}$ with $y_{(40)}$. In other words, we form pairs **based on the value of the response**. Finally, I create 40 d's for my data set:

$$d_1 = x_{(1)} - y_{(1)}; d_2 = x_{(2)} - y_{(2)}; \ldots; d_{40} = x_{(40)} - y_{(40)}.$$

I summarize my 40 differences with $\bar{d}$ and s_d. Finally, I calculate Gosset's 95% confidence interval for $\mu_d = \mu_1 - \mu_2$:

$$\bar{d} \pm 2.023(s_d/\sqrt{40}).$$

What happens if we do this? To answer this question, I need to involve Nature and computer simulations.

Suppose, for example, Nature knows that the two populations are identical and both are the Normal curve with $\mu = 69$ inches and $\sigma = 3$ inches. Thus, a confidence interval for $\mu_d = \mu_1 - \mu_2$ will be correct if, and only if, it contains zero. I performed a simulation experiment with 10,000 reps to investigate the actual performance of this confidence interval for paired data. The results were:

- A total of 3,523 confidence intervals were too large;
- A total of 3,527 confidence intervals were too small; thus,
- A total of 7,050 confidence intervals were incorrect.

Note that there should be approximately 500 incorrect confidence intervals. Seven thousand fifty is quite a bit larger than 500. This simulation study shows convincingly that forming pairs based on the response is invalid! By the way, the above simulation applies to all pairs of Normal curves that are congruent; i.e., the two population means don't need to be the same number. Similar results will be obtained for noncongruent Normal curves, but they will require a different simulation experiment to discover just how horrible the method performs! Similar results are true for populations that are not Normal curves. In short, this method is always bad!

I have never found a textbook that was shameless enough to propose the above method—sort the data, form pairs, subtract, etc. Alarmingly, however, I did find several textbooks that advocated the following form of experimental design. I will state their suggestion in terms of the above height study.

They do not say, "Form pairs based on the response, height;" instead, they advocate forming pairs based on another feature that is correlated with height, perhaps weight. **This is also invalid!** If you do this the actual confidence level of your nominal 95% confidence interval will be much smaller than 95%.

20.5 An Extended Example

Pairing is a very exciting topic. (I know, exciting is like funny; if it's really funny, do I need to tell you?) It is exciting because it allows a researcher to use scientific knowledge to improve a study; i.e., it's not about math or algebra.

When I decide to investigate a topic statistically, after I have a general notion of the response, I always ask myself the following two-part question:

1. What factor(s) do I *suspect* will cause variation in the responses from unit to unit?

2. Of the factors listed, can I *deal with* one or more of them by forming pairs?

Are you a fan of major league baseball? Well, sadly, this example will be more interesting if you are. In any event, I will proceed.

One of the charms of major league baseball is that the dimensions of the 30 major league ballparks **are not constant**. The most famous ballpark (some residents of New York and Chicago might disagree) is Fenway Park in Boston. The distance from home plate to its left field fence is the shortest in the major leagues, partly offset by the fact that said fench is the tallest, measuring 37 feet, two inches high. The second most famous ballpark (again, according to me) is Wrigley Field in Chicago. Wrigley Field is the topic of this example.

Wrigley Field has a reputation for being *hitter friendly*; in particular, the *conventional wisdom* is that it is easier to hit a home run in Wrigley Field than in an *average ballpark*. I will investigate this issue. How might one investigate this issue?

Here is my first attempt. Take for my response the total number of home runs hit in a stadium in a year. Think of this as a 30-population problem, with each year giving us another observation for each of the 30 populations (ballparks).

For example, in the 2013 National League season, a total of 2,145 home runs were hit in its 15 ballparks, for a mean of $214/15 = 143$ home runs per ballpark. The four largest responses are: Milwaukee, 185; Cincinnati, 184; Philadelphia, 176; and Chicago, 175. The three smallest responses are: Miami, 84; Pittsburgh 106; and St. Louis, 108. The Wrigley Field data support the *conventional wisdom*; the number of home runs hit there was well above the mean.

Do you see a weakness in the above discussion? Here is one. To paraphrase the NRA, "Ballparks don't hit home runs, players do." It is inarguable that the management of a baseball team considers its ballpark while building its roster of players. Thus, for example, part of the reason there were more home runs hit in Milwaukee than in Miami is that the former's roster contained more *power hitters*.

Many years ago somebody—sorry, I don't know who gets credit—had a clever idea. Let me show you some data and then explain the idea. Look at Table 20.6. Let's look at 1967. We see the response value 160 for the Cubs' (Chicago's team) home games. This is the total number of home runs hit in Wrigley Field in 1967. Thus, it is the same idea as the response values I gave you earlier for the 2013 season. Here is the twist. We compare this value, 160, to the total number of home runs hit by both teams in all of the Cubs' away games, 110. By comparing all of the Cubs' home games with all of their away games, we have—for the most part—removed the effect of rosters.

Table 20.6: Number of home runs, for both teams, in Cubs games, 1967–1987.

	Season									
Location	1967	1968	1969	1970	1971	1972	1973	1974	1975	1976
Home	160	166	148	201	144	146	138	139	125	155
Away	110	102	112	121	116	99	107	93	100	73
Home−Away	50	64	36	80	28	47	31	46	25	82

	Season									
Location	1977	1978	1979	1980	1982	1983	1984	1985	1986	1987
Home	151	117	151	116	115	140	156	202	168	204
Away	88	80	111	100	112	117	79	104	130	164
Home−Away	63	57	40	16	3	23	77	98	38	40

By the way, here are two summary statistics for the data in Table 20.6:

$$\bar{d} = 46.20 \text{ and } s_d = 24.42.$$

If I view the 20 seasons of data as the result of observing 20 i.i.d. trials, then I can obtain a 95% confidence interval estimate of μ_d:

$$46.20 \pm 2.093(24.42/\sqrt{20}) = 46.20 \pm 11.43.$$

In words, on average, Wrigley Field increases the mean number of home runs by at least 34.77 and at most 57.63 per season.

20.6 Computing

The extremely versatile and useful *vassarstats* website can be used to analyze paired data. I will illustrate the method for my Tetris data in Table 20.5. Go to the website:

`http://vassarstats.net`.

The left-side of the page lists a number of options; click on *t-Tests & Procedures*. This takes you to a new set of options; click on the top one, *Two-Sample t-Test for Independent or Correlated Samples*. This takes you to a new page. In the *Setup* section, click on *Correlated Samples*. (If you forget to do this, it's a big problem because *Independent Samples* is the default.) Next, enter the data, by typing or pasting, and click on *Calculate*.

The website gives me the following information:

$$m = n_A = n_B = 10; \bar{x} = 116.5; \bar{y} = 90.0; \text{ and } \bar{d} = \text{ Mean}_a - \text{ Mean}_b = 26.5.$$

It also reports that the observed value of the test statistic is $t = 7.06$ with approximate P-value < 0.0001 for both of the alternatives $>$ and $\neq$. The *vassarstats* testing output is all consistent, though a bit less precise, than what I obtained earlier by hand. Finally, *vassarstats* reports a variety of confidence intervals, including 26.5 ± 8.4846 as the 95% confidence interval estimate of μ_d. This is the same answer I obtained, except for round-off error.

I have found a website that will create a scatterplot and compute the correlation coefficient:

`http://www.wessa.net/rwasp_Pregnancy%20and%20cognition.wasp#output`.

You are not responsible for using this site; I am very grateful that it exists, but it's a bit tedious to use. (For example, it requires a fair amount of time to delete the site's default data before you can enter your own data.) If you want to try it out; I suggest that you use my Tetris data. As a partial check, you should obtain $r = 0.2955$ for the correlation coefficient. Or you could use either one of my RPDs for headache pain to see whether your output matches the scatterplot in Figure 20.1.

20.7 Summary

This chapter continues the theme of Chapter 19; namely, comparing the means of two populations. Instead of independent samples, in this chapter we have paired data. Each pair gives: a response from population 1, x; and a response from population 2, y. These two numbers can be used to compute $d = x - y$, which can be viewed as a response from the population of differences.

Here's another way to view this structure: We have random samples from both populations 1 and 2, but the random samples are not independent of each other. The result is that we have a random sample from the population of differences.

In Chapter 19, when considering population means, we focused on estimating $\mu_1 - \mu_2$ with confidence and testing the null hypothesis that $\mu_1 = \mu_2$. In the current chapter, these inference problems become estimating μ_d with confidence and testing the null hypothesis that $\mu_d = 0$.

Mathematically, Chapter 20 reduces to the problem of inference for a single population mean (of the population of differences). This problem was studied in Chapters 17 and 18 and I recommend using Gosset's procedures, subject to the caveats mentioned in these earlier chapters.

A researcher needs to be careful to avoid misusing the formulas for paired data. In particular, we found that the methods for paired data are appropriate for:

- Unit reuse, with or without randomization;
- Forming pairs of adjacent trials, using randomization to assign one trial of each pair to each treatment; and
- Matching subjects at the population level, as described earlier.

I gave a simple and dramatic example illustrating that paired data methods should never be used for pairing performed at the sample level, again, as described earlier.

Finally, you learned how to create and interpret a scatterplot of pairs of responses. Thinking about the likely pattern in such a scatterplot can help a researcher decide whether to have a design with independent samples or paired data.

Table 20.7: Data for the RPDs described in Practice Problems 1 and 2.

	Pair									
Treatment	1	2	3	4	5	6	7	8	9	10
Smoking (x)	173	175	169	175	180	184	182	186	190	188
Not smoking (y)	163	160	157	165	167	159	170	155	153	164
Difference ($d = x - y$)	10	15	12	10	13	25	12	31	37	24

20.8 Practice Problems

1. Bascom Hill is a long, steep hill (by Wisconsin standards) in the center of the university campus in Madison. A student in my class, Damion, wondered whether smoking a cigarette affected his climbing of Bascom Hill. He performed an RPD with response equal to the time, measured to the nearest second, he needed to walk from the bottom to the top of the hill. The first treatment consisted of walking while smoking a cigarette and the second consisted of walking while not smoking a cigarette. Damion formed pairs from his trials, exactly as I did for the Tetris study described in this chapter.

   Damion's data are in Table 20.7. Below are various summary statistics for these data.

   $$\bar{x} = 180.2, \bar{y} = 161.3, s_1 = 6.99, s_2 = 5.44 \text{ and } s_d = 9.67.$$

   (a) Calculate Gosset's 95% confidence interval estimate of μ_d. Write one sentence that interprets your confidence interval.

   (b) Find the approximate P-value for the alternative $\mu_d > 0$.

   (c) Pretend that the data came from a CRD instead of an RPD. Calculate Gosset's 95% confidence interval estimate of $\mu_1 - \mu_2$.

   (d) Compare your answers to (a) and (c). In your opinion, which would have been a better way to conduct the study; an RPD or a CRD? Explain your answer.

   (e) Use Equation 20.4 to determine the value of the correlation coefficient for x and y.

2. Now suppose that Damion had ended his RPD after the first five pairs were completed. Use the *vassarstats* website to redo problem 1. For part (e), to save time you may use the following summary statistics, which I obtained from *vassarstats*:

   $$s_1 = s_2 = 3.9749 \text{ and } s_d = 2.1213.$$

3. Alisa performed an RPD to compare bowling with one hand (the usual method) and bowling two handed ('granny style;' her words, not mine). A trial consisted of a game of bowling and the response was Alisa's score. Below are the results of the study:

Game	Hands	Score	Game	Hands	Score
1	One	97	6	One	110
2	Two	85	7	One	123
3	Two	91	8	Two	96
4	One	108	9	One	125
5	Two	95	10	Two	94

(a) Present these data in a format similar to what I used in Table 20.7. Put one-handed bowling in the first row; i.e., the x's. Don't analyze these data; I simply want you to make sure you can transform one table into another.

(b) Assuming Alisa used our website randomizer, what output did it give her?

20.9 Solutions to Practice Problems

1. (a) First,

$$\bar{d} = \bar{x} - \bar{y} = 180.2 - 161.3 = 18.9.$$

Next, t^* for $df = 10 - 1 = 9$ is 2.262. Thus, Gosset's 95% confidence interval estimate of μ_d is:

$$18.90 \pm 2.262(9.67/\sqrt{10}) = 18.90 \pm 2.262(3.058) = 18.90 \pm 6.92 = [11.98, 25.82].$$

The mean time to walk up the hill while smoking is between 11.98 and 25.82 seconds larger than the mean time to walk up the hill while not smoking.

(b) The observed value of the test statistic is

$$t = 18.90/3.058 = 6.1805.$$

The area under the t-curve with $df = 9$ to the right of 6.1805 equals (using Minitab) 0.00008. This is the approximate P-value.

(c) First,

$$s_p^2 = \frac{(6.99)^2 + (5.44)^2}{2} = 39.227 \text{ and } s_p = \sqrt{39.227} = 6.263.$$

Next, t^* for $df = 10 + 10 - 2 = 18$ is 2.101. Thus, Gosset's 95% confidence interval estimate of $\mu_1 - \mu_2$ is:

$$18.90 \pm 2.101(6.263)\sqrt{2/10} = 18.90 \pm 5.88 = [13.02, 24.78].$$

(d) The half-width from the pretend CRD, 5.88, is 15.0% narrower than the half-width from the actual RPD, for the same data. This supports the notion that a CRD would have been better, **but** we don't really know what would have happened with a CRD.

(e) First,

$$s_d^2 = (9.67)^2 = 93.5089.$$

Next,

$$s_1^2 + s_2^2 - 2rs_1s_2 = (6.99)^2 + (5.44)^2 - 2r(6.99)(5.44) = 78.4537 - 76.0512r.$$

Setting these equal to each other, we get:

$$76.0512r = 78.4537 - 93.5089 = -15.0552 \text{ or } r = -0.198.$$

2. I entered the data into *vassarstats*, being careful to specify *Correlated Samples* and obtained the following relevant summaries:

$$\bar{x} = 174.4, \bar{y} = 162.4 \text{ and } \bar{d} = 12.0.$$

(a) The website tells me that the 95% confidence interval estimate of μ_d is 12.00 ± 2.64. Notice that this interval is much narrower than the interval from all ten pairs!

(b) The website tells me that the observed value of the test statistic is $t = 12.65$ with $df = 4$ and that the approximate P-value for $>$ is 0.0001125.

(c) I enter the same data into the website, being careful to specify *Independent Samples*. The site tells me that the 95% confidence interval estimate of $\mu_1 - \mu_2$ is 12.00 ± 5.81.

(d) The half-width of the confidence interval for the actual RPD, 2.64, is 54.6% narrower than the half-width for the pretend CRD with the same data. This is a huge difference! We don't know for sure, however, what would have happened with a CRD, but the RPD does look better.

(e) From Equation 20.4,

$$(2.1213)^2 = (3.9749)^2 + (3.9749)^2 - 2r(3.9749)^2.$$

This becomes:

$$4.499914 = 2(15.79983) - 2r(15.79983); \text{ or } 2r(15.79983) = 27.099746; \text{ or } r = 0.858.$$

3. (a) Alisa's table is below.

	Pair				
Treatment	1	2	3	4	5
One-handed (x)	97	108	110	123	125
Two-handed (y)	85	91	95	96	94
Difference ($d = x - y$)	12	17	15	27	31

(b) The site gave her: 1, 2, 2, 1, 1.

20.10 Homework Problems

1. Martha and Lisa performed an RPD to investigate Martha's juggling skills. The first treatment was Martha juggling three tennis balls; the second treatment was Martha juggling three large apples. The response is the length of time, measured to the nearest second, that the three items where in what they called *a regular cycle of juggling*. Below are selected summary statistics:

$$\bar{x} = 6.100, \bar{y} = 5.200, s_1 = 3.888, s_2 = 3.517, s_d = 5.826 \text{ and } m = 40.$$

   (a) Calculate Gosset's 95% confidence interval estimate of μ_d. Write one sentence that interprets your confidence interval.

   (b) Find the approximate P-value for the alternative $\mu_d > 0$.

   (c) Pretend that the data came from a CRD instead of an RPD. Calculate Gosset's 95% confidence interval estimate of $\mu_1 - \mu_2$.

   (d) Compare your answers to (a) and (c). In your opinion, which would have been a better way to conduct the study; an RPD or a CRD? Explain your answer.

   (e) Use Equation 20.4 to determine the value of the correlation coefficient for x and y.

2. Deborah's son Scotty is convinced that his Snowbie sled is slower than his friend Sam's Sno-Racer. An RPD was conducted to investigate this issue. A trial consisted of a slide down a local hill. The response is the time, measured to the nearest tenth of a second, that Scotty required to complete a slide. The first treatment consists of Scotty riding his Snowbie and the second treatment is Scotty riding Sam's Sno-Racer. Below are the results of the study.

Trial	Treat.	Time	Trial	Treat.	Time	Trial	Treat.	Time
1	2	11.3	7	2	9.0	13	2	10.1
2	1	12.0	8	1	12.1	14	1	8.8
3	2	11.3	9	1	8.9	15	2	9.9
4	1	11.1	10	2	10.7	16	1	10.5
5	2	10.1	11	2	10.6	17	1	12.2
6	1	8.4	12	1	9.8	18	2	12.1

   (a) Present these data in a format similar to what I used in Table 20.7.

   (b) Use the *vassarstats* website to obtain Gosset's 95% confidence interval estimate of μ_d.

   (c) Use the *vassarstats* website to obtain Gosset's approximate P-value for the alternative $\mu_d > 0$.

   (d) Assuming Deborah used our website randomizer, what output did it give her?

Chapter 21

Simple Linear Regression

Linear regression analysis is one of most popular methodologies in all of Statistics. The Statistics Department at UW–Madison offers a one-semester course, Statistics 333, devoted to it. In this chapter and the next, I will introduce you to the subset of regression methods that fall under the name **simple linear regression**.

In the current chapter I will present **descriptive** methods of regression and in Chapter 22, I will present methods of **inference**. As you will see, we will be looking at scientific problems in which each unit (subject or trial) yields two numbers; one denoted by x and the other by y. To distinguish between data from different cases— by the way, units are called cases in regression—we will use subscripts. Thus, for example, our first case gives the pair of numbers (x_1, y_1); the second case gives (x_2, y_2); and so on. Our general notation is that case 'i' gives (x_i, y_i). When I am being less formal, I will sometimes refer to the x's and the y's, without subscripts.

In our first example below, the cases are individual spiders and the two variables determined for each spider are its heart rate and body weight. One of the spiders, for example, has body weight $x = 0.045$ grams and heart rate $y = 60$ beats per minute. Thus, for this spider the pair (x, y) is the pair of numbers $(0.045, 60)$. Obviously, we need to be careful to keep the order of these numbers straight; the pair $(60, 0.045)$ would denote a spider that weighs 60 grams and has a heart beat about once every $1/0.045 = 22.2$ minutes! A really big spider with, presumably, a very short life span.

Remember that we use lower case letters, x and y, to denote the numbers obtained for any particular case. We will user upper case letters, X and Y, to denote the variables being determined. For example, in our spiders example, X denotes body weight in grams and Y denotes heart rate in beats per minute.

Also, please remember the following: In this chapter I make no probability assumptions about the data

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n),$$

where n is the number of cases in the data set. In other words, in this chapter I will **not** assume that these cases are a random sample—smart or dumb—from a finite population of cases **nor** will I assume that they are the result of observing i.i.d. trials. Such assumptions will be considered in Chapter 22.

Table 21.1: Body weight, in grams, and heart rate, in beats per minute, for five categories of 48 spiders.

Small Hunters		**Tarantulas**		**Large Hunters**		**Web Weavers**		**Primitive Hunters and Weavers**	
Weight	Rate	Weight	Rate	Weight	Rate	Weight	Rate	Weight	Rate
0.045	60	10.75	11	0.980	27	0.422	27	0.050	13
0.031	61	11.10	13	0.623	43	0.387	44	0.090	15
0.105	90	8.01	14	0.483	15	0.324	48	0.104	19
0.093	125	13.80	10	0.431	19	0.234	55	0.108	9
0.139	100	12.60	11	0.324	22	0.439	36	0.132	10
0.050	108	11.40	12	0.289	27	0.357	42	0.117	12
0.161	82			1.135	19	0.325	68	0.095	17
0.146	98			0.906	23	0.106	54	0.127	22
0.140	105			0.591	23	0.325	63		
				0.570	25	0.287	75		
				1.152	34	0.404	45		
				1.363	36	0.540	63		
						0.506	68		

21.1 The Scatterplot and Correlation Coefficient

You received a brief exposure to the scatterplot and correlation coefficient in Chapter 17 and a more extended introduction in Chapter 20.

When I was writing a textbook, approximately 20 years ago, I found some interesting data on spiders [1] that I present in Table 21.1. I am not an arachnologist—indeed, I can't even spell it without help; thus, I can't really speak to why these data are important. Therefore, I will follow the approach in the journal article. In addition, these spider data illustrate some interesting statistical issues.

In this chapter, our focus will be on examining the association between two numbers; in the current case, body weight and heart rate of spiders. First, however, it is useful to examine these variables separately for our five categories of spiders. Various descriptive statistics are presented in Table 21.2. Let me make a few brief comments about the means in this table.

1. Tarantulas are much heavier than the other types of spiders. At the other extreme, small hunters and primitive hunters and weavers are quite tiny. It is reassuring to note that small hunters are, indeed, smaller than large hunters!

2. The mean heart rate for small hunters is much larger than the other means. At the other extreme are tarantulas and primitive hunters and weavers.

3. It is striking how the similarly sized small hunters and primitive hunters and weavers have

Table 21.2: Summary statistics for body weight, in grams, and heart rate, in beats per minute, for five categories of 48 spiders.

		Body Weight		Heart Rate	
Category	n	Mean	St. Dev.	Mean	St. Dev.
Tarantulas	6	11.3	1.96	11.8	1.47
Primitive hunters and weavers	8	0.103	0.026	14.6	4.50
Large hunters	12	0.737	0.357	26.1	8.03
Web weavers	13	0.358	0.114	52.9	14.1
Small hunters	9	0.101	0.049	92.1	21.5

such different heart rates. Also, tarantulas and primitive hunters and weavers, despite their vastly different sizes, have similar mean heart rates.

For each spider, I have two numbers: heart rate and weight. It will be useful to view these two variables in an asymmetrical fashion. In particular, I ponder the following question:

Which of the following perspectives makes more sense (scientifically)?

- A spider's heart rate influences it body weight.
- A spider's body weight influences it heart rate.

I choose the second of these perspectives. As you will see below, if you disagree with my choice, some—but not all—of your analyses will differ from mine.

Anyways, given my chosen perspective, the language we use in Statistics is to refer to the heart rate as the **response** and the body weight as the **predictor**. In the old days, heart rate was referred to as the **dependent variable** and body rate was referred to as the **independent variable**; the idea being that the former depended on the latter and the latter, well, didn't depend on anything! Fortunately, this older terminology is dying out; I say fortunately because this use of independent is confusing because it does not match our earlier use of the term. Some social scientists' appear to prefer the words exogenous (for predictor) and endogenous (for response).

In any event, all agree to refer to the predictor with the symbol X and the response with the symbol Y. (Thus, one way to keep the ex/end—ogenous terms straight: **ex**ogenous is X and **end**ogenous is nearer the **end** of the alphabet.)

Note the following. The implicit perspective in all regression analyses in these *Course Notes* is:

A case's value of X influences its value of Y.

In the current situation, we are interested in two numerical variables from each spider; its body weight X and its heart rate Y. For any particular spider, these two variables take on numerical values, denoted by lower case letters, x and y. Let's look at the data in Table 21.1 for the small

hunters. First, I note that there are data for nine small hunters; thus, I set the sample size at $n = 9$. The data set for small hunters consists of $n = 9$ pairs of numbers, first symbolically as:

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_9, y_9),$$

and then, numerically as (reading down the table):

$$(0.045, 60), (0.031, 61), (0.105, 90), \ldots, (0.140, 105).$$

We will now draw a picture of these nine pairs of numbers, called the scatterplot. (I told you a bit about scatterplots in Chapters 17 and 20; forgive me the redundancies below.) I like to think of a scatterplot as a dotplot in two dimensions, viewed from above. The scatterplot of heart rate versus body weight for the $n = 9$ small hunter spiders is presented in the upper left picture in Figure 21.1. First note that there are nine circles in this scatterplot, one for each pair of values (x, y). Recall that our first observation from a small hunter is the pair $x = 0.045$ and $y = 60$. Can you find this spider's circle in the scatterplot? (Answer: Find the two circles in the southwest corner of the scatterplot; the circle to the right in this twosome is the one we seek.) Locate a few more of the (x, y) pairs in the scatterplot; you don't necessarily need to find all nine pairs, just enough to convince yourself that you understand the process.

After constructing a scatterplot we look for isolated cases. This brings me to one of the main reasons I selected these spider data for an introduction to this material. I believe that with a small value of n is is extremely difficult to decide whether cases are isolated. Moreover, any such decision tends to have big implications for how we interpret the data. I would label the two cases in the southwest corner of the scatterplot as being isolated from the other seven cases. I will now argue why this is important.

After considering the possibility of isolated cases, we look for a pattern in the scatterplot. The first pattern we look for is the following.

> As the value of x increases (i.e., moving our eyes left-to-right across the scatterplot) what happens to y? Below are three possibilities:
>
> - As x increases, y also increases—called an increasing relationship; or
> - As x increases, y decreases—called a decreasing relationship; or
> - Neither of the above (stay tuned for more details on this).

For the $n = 9$ small hunters, the relationship between x and y is increasing. If, however, we cover up my two candidates for isolated cases and look at the remaining seven cases, I would say that there is a clear and pretty strong decreasing relationship between x and y. The difference between having an increasing relationship and having a decreasing relationship usually is **huge** in science.

Let me make one more comment about isolated cases for the small hunters. I am **not** advocating that you discard the two isolated cases before analyzing the data. But neither am I advocating that you keep them before analyzing the data. This decision should be made by a scientist, not a statistician. I hope, of course, that the scientist will use solid knowledge—and not wishful thinking—in making such a choice. My main goal is to encourage you to realize that with a

Figure 21.1: Scatterplot and correlation coefficient, r, of heart rate (beats per minute) versus body weight (grams) for each of four categories of spiders.

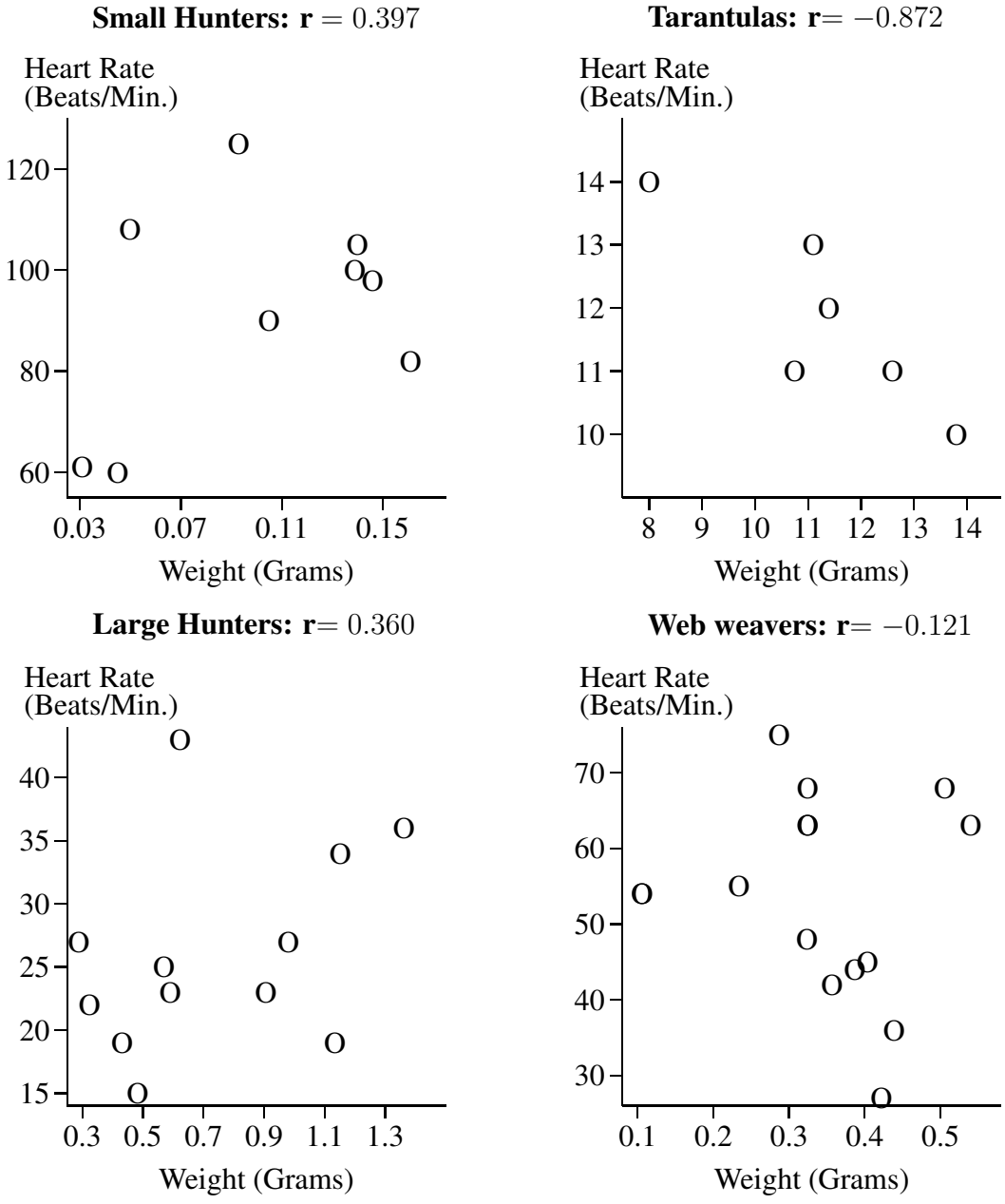
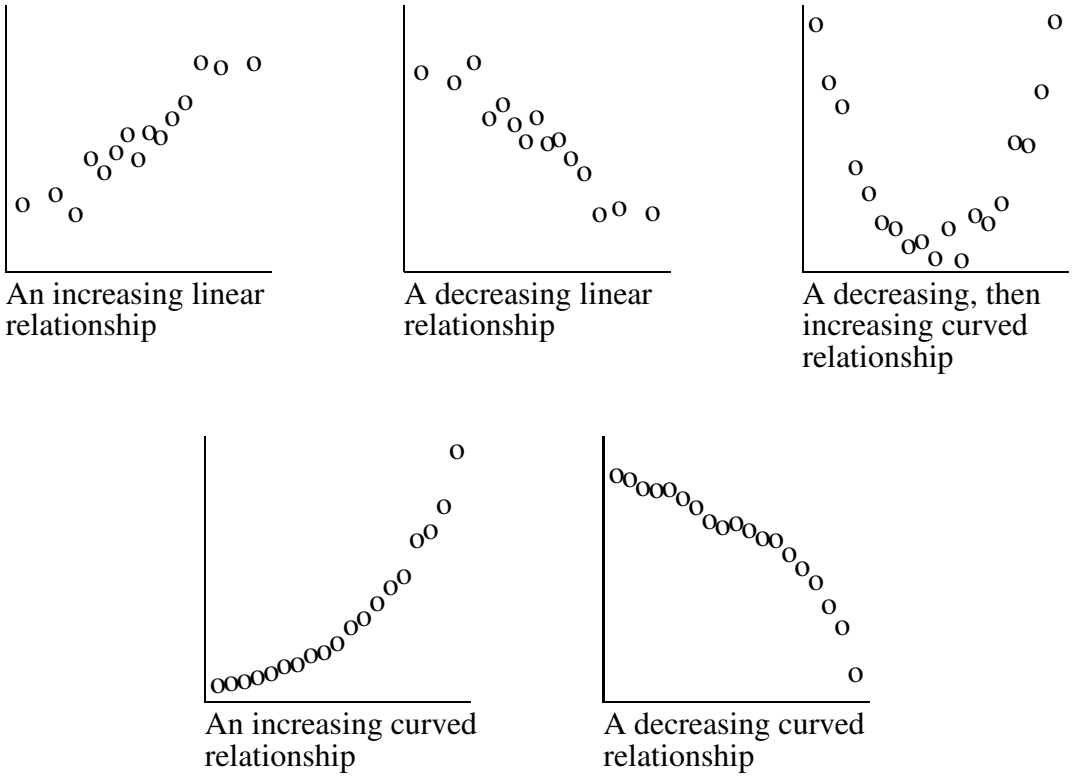


Figure 21.2: Five scatterplots.



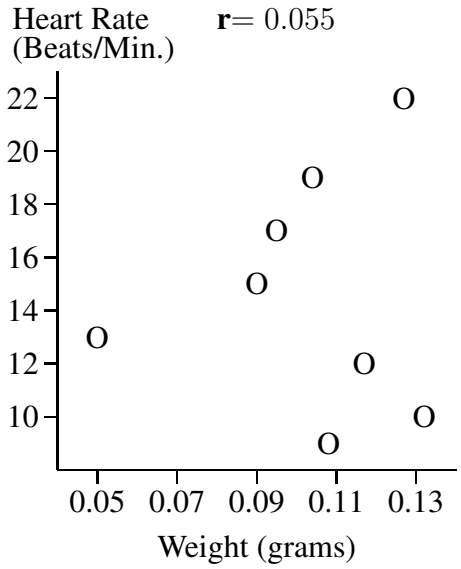
slight change in the data, the analysis could change drastically. An obvious way to have a change in the data is by having the researcher deliberately discard one or more cases. It is important, however, to also remember that there is some chance involved in the data we have. I am not ready to argue whether or not I am willing to pretend that these nine spiders are a random sample from the population of all small hunters. (Try to imagine a way to obtain an actual random sample of spiders!) It is, however, worth realizing that it is conceivable that our two *isolated spiders* might not have ended up in our data set. (This is reminiscent of Kenny's data on speeds of cars in which we had to acknowledge the chance aspect of the large outlier even being in the data set.)

Briefly examine the other three scatterplots in Figure 21.1; what do you see? Regarding isolated cases, I see two possibilities: the large hunter with the largest heart rate and the tarantula with the smallest body weight. You, of course, may reasonably disagree with the possibilities I see.

Looking at the patterns in these three scatterplots I see: increasing for the larger hunters; strongly decreasing for the tarantulas; and weakly decreasing for the web weavers. I see the same patterns whether or not I exclude my two candidates for isolated cases.

Let's briefly leave our study of spiders; I want to be a bit more general. Look at Figure 21.2. This figure presents five scatterplots: two present increasing relationships; two present decreasing

Figure 21.3: Scatterplot of heart rate (beats per minute) versus body weight (grams) for eight spiders classified as primitive hunters and weavers.



relationships; and the remaining scatterplot shows neither. Or both. Depends on how you look at it. What I most want you to note is that two of the scatterplots reveal a linear relationship between x and y and three of the scatterplots reveal a curved relationship between x and y. **With the exception of Section 22.4, in the remainder of these *Course Notes* we will restrict attention to relationships that are linear.** Regression analysis is very useful for studying curved relationships, but we won't have time to explore this topic.

I don't want to mislead you with Figure 21.2. In science, it is not always so easy to decide whether a relationship is curved or linear. In my opinion, the four scatterplots in Figure 21.1 all reveal linear relationships. If you don't agree, remember the following. Statisticians *hope* to find a linear relationship because assuming a linear relationship provides some advantages—in ease of the work and, especially, interpretation—over assuming a curved relationship. Thus, I am aware that I might be *too eager to see a linear relationship.*

Figure 21.3 presents the scatterplot for the fifth category of spiders, the primitive hunters and weavers. In this picture, I see: no isolated cases; and a linear relationship that is neither increasing nor decreasing: As I move my eyes from left-to-right, the values of y jump around, but trend neither up nor down.

Each of my five spider scatterplots includes a number r, which is called the correlation coefficient. Many of you may be familiar with the correlation coefficient. I will assume that you are not familiar with it and will now explain it.

Following our notation from Chapter 19, denote the mean and standard deviation of the x's by $\bar{x}$ and s_1, respectively. Also, denote the mean and standard deviation of the y's by $\bar{y}$ and s_2,

respectively. Our only restriction on these values is:

$$s_1 > 0 \text{ and } s_2 > 0. \tag{21.1}$$

In words, the x's [y's] are not all the same number. Remember that our goal is to determine whether the value of x influences the value of y. If all of the x_i's are the same number, how can we look for influence? Or, if all of the y_i's are the same number, there can be no evidence of them being influenced by anything!

Definition 21.1 (The correlation coefficient.) (Pearson's product moment) correlation coefficient *is denoted by r and given by the following equation:*

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_1 s_2} \tag{21.2}$$

By the way, you can see why I restrict our attention to data sets that satisfy Condition 21.1; otherwise, both the numerator and denominator in Equation 21.2 would equal zero. Note that I will **never** ask you to compute r by hand. Our *scatterplot* website from Chapter 20 reports the value of r, but I won't make you use it. In this course, I will give you r **or** provide you with enough information so that obtaining r is a matter of simple (according to me!) algebra.

There are several important properties of the correlation coefficient. For convenience, I list six of them below under the heading of a result. When you read through these you will see that the first property is not really a mathematical result; it is simply terminology. Also, the second property is a bit imprecise. I trust that you will forgive these transgressions of mine. In any event, read through these properties quickly; the list is followed by explanations. Also, the 12 scatterplots in Figure 21.4 will illustrate my explanations.

Result 21.1 (Six properties of the correlation coefficient.) *The following six properties will help you develop some intuition for the value of the correlation coefficient, given in Equation 21.2.*

1. If the correlation coefficient is greater than zero, the variables Y and X are said to have a **positive linear relationship**; if it is less than zero, the variables are said to have a **negative linear relationship**; if it equals zero, the variables are said to have **no linear relationship**, or to be **uncorrelated**.

2. The correlation coefficient is **not appropriate** for summarizing a curved relationship between Y and X. Therefore, it is *always* necessary to examine a scatterplot of the data to determine whether computation of the correlation coefficient is appropriate.

3. The value of the correlation coefficient is always between -1 and $+1$. It equals $+1$ if, and only if, all data points lie on a straight line with positive slope; it equals -1 if, and only if, all data points lie on a straight line with negative slope.

4. The farther the value of the correlation coefficient from zero, in either direction, the 'stronger' the linear relationship.

5. The value of the correlation coefficient does not depend on the units of measurement chosen by the experimenter. More precisely, if X is replaced by $aX + b$ and/or Y is replaced by $cY + d$, where a, b, c, and d are any numbers with a and c bigger than zero, then the correlation coefficient of the new variables is equal to the correlation coefficient of X and Y. The numbers a and c are required to be positive to avoid reversing the direction of the relationship; a related result can be obtained if a and/or c are negative, but it will not be needed in these *Course Notes*. Among many examples, this result shows that changing from miles to inches, pounds to kilograms, degrees Celsius to degrees Fahrenheit, or seconds to hours will not change the correlation coefficient.

6. The correlation coefficient is symmetric in X and Y. In other words, if the researcher changes perspective and relabels the predictor and response, the correlation coefficient will not change. In particular, if there is no natural assignment of the labels predictor and response to the two numerical variables, the value of the correlation coefficient is not affected by which assignment is chosen.

21.1.1 Explanations of the Six Properties of the Correlation Coefficient

In the explanations below, imagine that the 12 scatterplots in Figure 21.4 are labeled $A, B, \ldots, L$ according to the following correspondence:

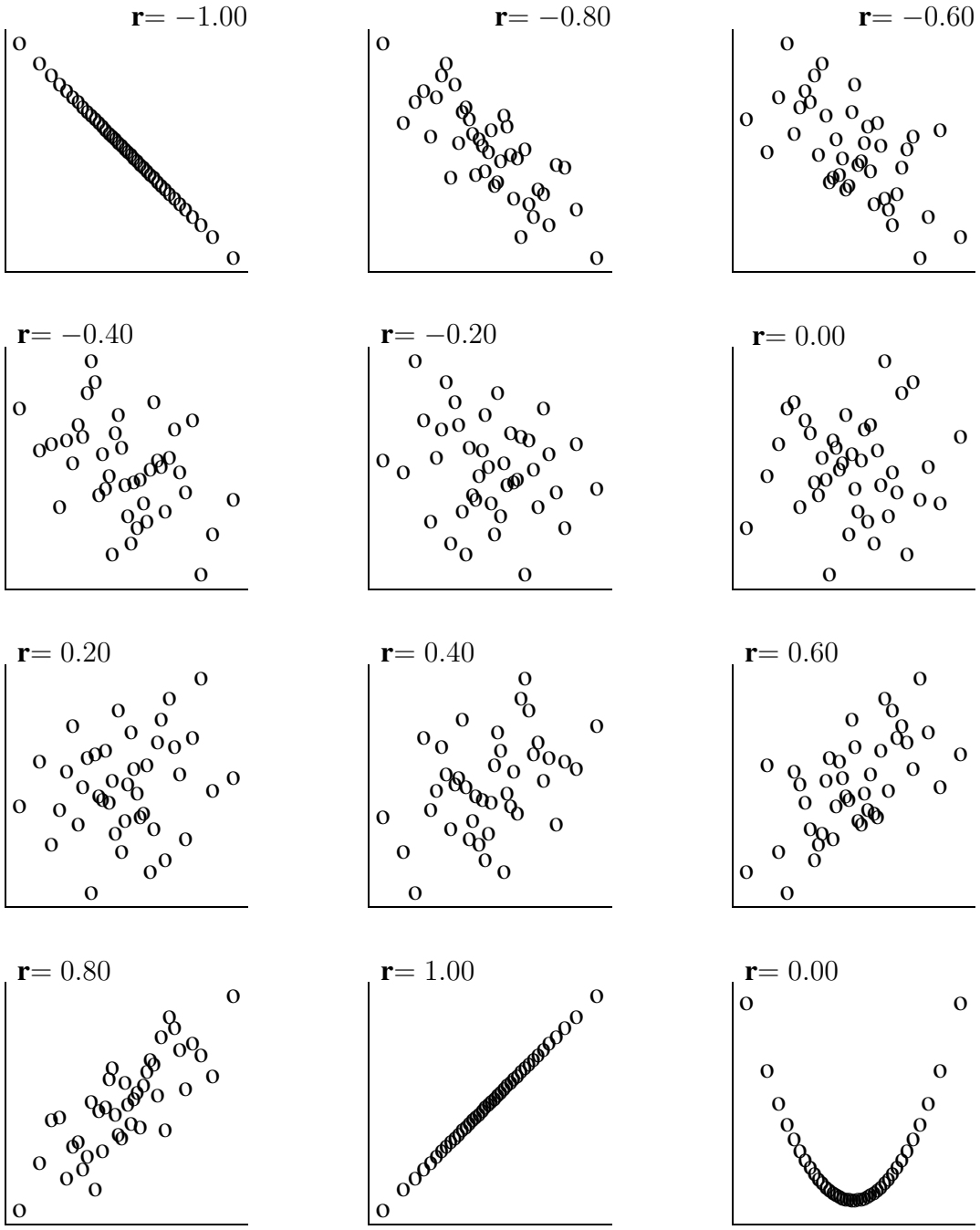
A	B	C
D	E	F
G	H	I
J	K	L

Below I explain—or at the least, expand upon—the six properties of the correlation coefficient that are listed above. I suggest that before you read each explanation below, reread the statement of the property above.

1. This first item is about terminology. Actually, it's a bit more than terminology, but it's easy to miss the extra. Look at our earlier scatterplots for small hunters and large hunters. Visually, we (well, me anyways, I am the one who controls the keyboard) agreed that both of these scatterplots revealed increasing linear relationships. Sure enough; both correlation coefficients are positive numbers: 0.397 and 0.360. Also, visually, the scatterplots for tarantulas and web weavers revealed decreasing linear relationships. Sure enough; both correlation coefficients are negative numbers: -0.872 and -0.121. Have you spotted what's extra?

   Well, literally, the first item makes no mention of what we see visually. (Formulas, after all, involve math and math doesn't care very much about what we *see*. Math rarely asks our opinion!) The first item tells us: If $r > 0$, then there is an increasing linear relationship; i.e., whether something is increasing or decreasing is no longer a matter of visual assessment, it is the result of a computation. If you can't see the increasing linear relationship, then that is **your problem**; the correlation coefficient is not going to change to make you happy!

Figure 21.4: Twelve scatterplots and their correlation coefficients.



This issue comes into play with the scatterplot for the primitive hunters and weavers. The correlation coefficient is $r = 0.055$ which is positive; thus, whether you see it or not, there is an increasing linear relationship. As we will see soon and then more precisely later, it is a very weak increasing linear relationship.

The twelve scatterplots are encouraging: plots G–K look increasing and each has a positive correlation correlation coefficient; plots A–E look decreasing and each has a negative correlation correlation coefficient; and plot F appears to have no linear trend and its r equals 0. Plot L is an anomaly that I will consider in the next item.

2. I have always loved plots like our plot L. I think it's because I am red-green colorblind. If you need a break, go to

   `http://www.toledo-bend.com/colorblind/Ishihara.asp`

   to see how the world looks to those of us who are red-green colorblind. (I can definitely see the 25; sort of can see the 56; and when I am told it's a 29, I can almost see it; but the other three numbers don't seem to be there at all!)

   Anyways, as plot L shows, the correlation coefficient is *colorblind* when it comes to seeing curved patterns.

3. Property 3 is a simple consequence of some things we learn a bit later. I do, however, want to say a few things about it. You will never obtain a correlation coefficient that is larger than $+1$ or smaller than -1. In addition, these extremes are obtained only when all data points fall exactly on a straight line, as mentioned in the property. The two extremes have led to confusion among students, so I do want to comment on them.

   (a) What is the value of r if all of the points lie on a line with slope equal to 0?

   Answer: If all points lie on a horizontal line, then all cases have the same value for y, making $s_2 = 0$, which I do not allow for reasons stated earlier.

   (b) What is the value of r if all of the points lie on a vertical line?

   Answer: If all points lie on a vertical line, then all cases have the same value for x, making $s_1 = 0$, which I do not allow for reasons stated earlier.

   (c) Why doesn't the **numerical value of the slope** matter? In particular, shouldn't a slope equal to $+2$ imply a stronger relationship than a slope equal to $+1$?

   Answer: This one is tricky. Think about my spider data with Y equal to heart rate, in beats per minutes, and X equal to body weight, in grams. Suppose that we had a new category of spiders for whom all points lie exactly on a straight line with slope equal to $+2$. Note, however, that the slope being $+2$ is tied to my choice of units. If I changed the units for Y to beats per hour, then:

   - All of the y data values would increase by a factor of 60; and
   - Thus, the slope would increase by a factor of 60 and become $+120$.

Thus, the exact same data give a slope of $+2$ or $+120$ and in both cases the correlation coefficient r would equal $+1$.

4. For positive values of the correlation coefficient, you can see this fact by moving your attention from F to G to H to I to J to K. For negative values of the correlation coefficient, you can see this fact by moving your attention from F to E to D to C to B to A.

   Also, note the symmetry in, say, scatterplots C and I. Scatterplot C gives $r = -0.60$ and I gives $r = +0.60$. While one scatterplot shows a decreasing relationship and the other shows an increasing relationship, if you look carefully you can see that the two patterns have exactly the same **strength**. If you cannot see this, I offer two suggestions:

   - If you hold a mirror up to scatterplot C, it becomes scatterplot I; or
   - Wait until we learn about the **coefficient of determination**, denoted by R^2.

5. In the optional Appendix near the end of this chapter, I will discuss briefly why—algebraically—property 5 is true. For now, note that we can rewrite the correlation coefficient as follows:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_1 s_2} = (\frac{1}{n-1}) \sum \frac{(x_i - \bar{x})}{s_1} \frac{(y_i - \bar{y})}{s_2}.$$

   In this latter form, we are taking the product of the standardized version of x with the standardized version of y and then summing the results. If, for example, relative to some group, your height is one standard deviation above the mean in inches, then it is one standard deviation above the mean in centimeters, kilometers, miles or even light-years. In other words, the correlation coefficient is not influenced by units.

   This property is very important because if you read that for some collection of cases the correlation coefficient for height and weight is some number r, then you know that the units—inches, meters or light-years matched with grams, pounds or tons—don't matter.

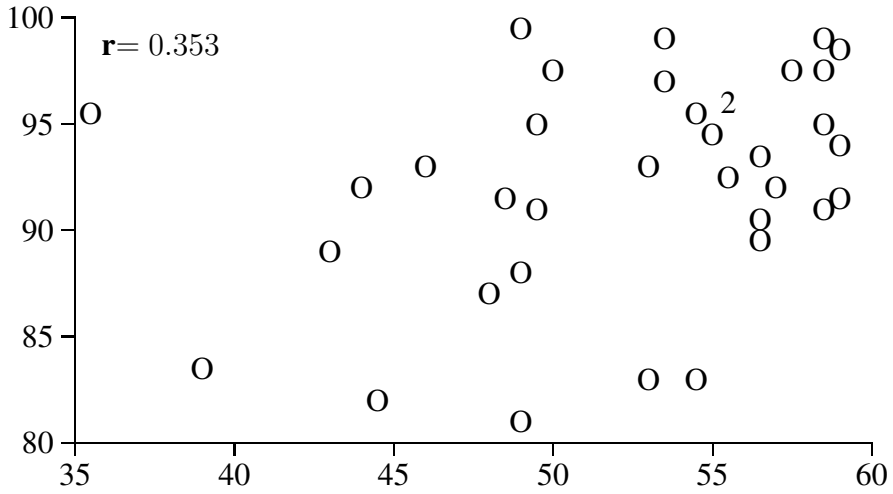
6. Please look at definition of the correlation coefficient in Equation 21.2. You will see that if you change all the x symbols to y symbols and all the y symbols to x symbols, then the formula remains the same; i.e. the correlation coefficient does not depend on which variable is labeled X and which is labeled Y.

21.1.2 Exam Scores in Statistics 371

I end this section with data from $n = 36$ students who took my traditional section of Statistics 371 during a recent summer school term. Each students took two exams: the midterm and the final. The maximum number of points on the exams was 60 for the midterm and 100 for the final. I graded in half-point increments. The scatterplot of the final exam score, Y, versus the midterm exam score, X, is given in Figure 21.5.

Following my own advice, I look for isolated cases. I see one: the student with the lowest score on the midterm had a very high score on the final. Either including or excluding my one isolated case, I see an increasing linear relationship. Including all 36 students, $r = 0.353$; excluding the one isolated case, $r = 0.464$.

Figure 21.5: Final exam score versus midterm exam score for 36 students. There is a '2' in the scatterplot because two subjects had $(x, y) = (55.5, 96.0)$.



21.2 The Least Squares Regression Line

Look again at the scatterplot of final exam score versus midterm exam score in Figure 21.5, but delete the one isolated case. If you are not happy—or, at least, are confused—about this deletion, I will mimic the work below for all 36 students in the Practice Problems. Thus, you will see precisely the effect on my analysis of deleting the one isolated case.

My goal is to find the equation of the line that best describes this scatterplot. This is a big task! It will take some time simply to explain what I mean. The line that best describes the scatterplot is called the **least squares regression line**, or the **best line** for short.

I am going to spoil the story for you; a bit like first reading the last 10 pages of a mystery novel. I am going to give you the best line for these data—or this scatterplot; whichever way you prefer to say it is fine.

The best line for my $n = 35$ pairs of exam scores has intercept $b_0 = 68.42$ and slope $b_1 = 0.4516$. I want to be able to write this as an equation and I do so as:

$$\hat{y} = b_0 + b_1 x, \text{ which becomes } \hat{y} = 68.42 + 0.4516x \text{ for my } n = 35 \text{ exam scores.}$$

I suspect that this equation looks a bit strange to you. To explore my hunch, I googled *equation of a line* and the first item on the list gave the equation

$$y = mx + b.$$

This is the form I learned as a child and that I taught during my brief career as a teaching assistant in math. It is strangely comforting that some math equations are, if not timeless, long-lived. Our

current equation,

$$\hat{y} = b_0 + b_1 x$$

is notably different than $y = mx + b$ and it will be useful for me to take a few minutes to explore the differences.

First, let's look at the left sides of these equations: $\hat{y}$ versus y. In Statistics, we may not write our line as $y = \ldots$ because our collection of (x, y) values do **not**, in general, fall on a line. Indeed, they fail to fall on a line for all of our real data examples, past, present and future. For convenience, **we need** to have a symbol for the values of $b_0 + b_1 x$ and we choose to use the symbol $\hat{y}$. You are familiar with statisticians' use of a hat to denote a point estimate, as in $\hat{p}$ in Chapter 12. In other settings, statisticians use a hat to denote prediction. Indeed, if I had taken the time to show you point predictions in Chapter 14—recall we did prediction intervals only—I would have used $\hat{y}$ as my point prediction of the number of successes y. For the current data set, we will view the quantity $68.42 + 0.4516x$ as the predicted final exam score given that the midterm score is x; thus, the use of a hat is natural.

Next, let's look at the right sides of these equations: $b_0 + b_1 x$ versus $mx + b$. I don't presume to speak for mathematicians, but I conjecture that they put the mx before the b because the slope is the more important component of the equation of a line: it denotes the change in y for *any unit change* in x, whereas the intercept is literally the value of y when x equals 0. Statisticians agree that the slope is far more important than the intercept, yet we put the intercept, b_0, before the slope, b_1, in our presentation of the equation. Why?

We are studying **simple** linear regression. **Simple** implies that there is exactly one predictor variable. As you may already know, in many scientific problems one predictor is not enough to obtain useful answers. (For example, models for climate change are not restricted to the single predictor: concentration of carbon dioxide in the atmosphere.) When we use more than one predictor, the methodology is called **multiple linear regression**. Suppose, for example, that we have two predictors. The first issue with two predictors is that we need to be able to distinguish between them. We do this by calling one predictor X_1 and the other X_2. In this situation, case i would yield three numbers: its response, y_i; its value on the first predictor, $x_{1,i}$; and its value on the second predictor, $x_{2,i}$. **Note that we would need to use the dreaded double subscripts!**

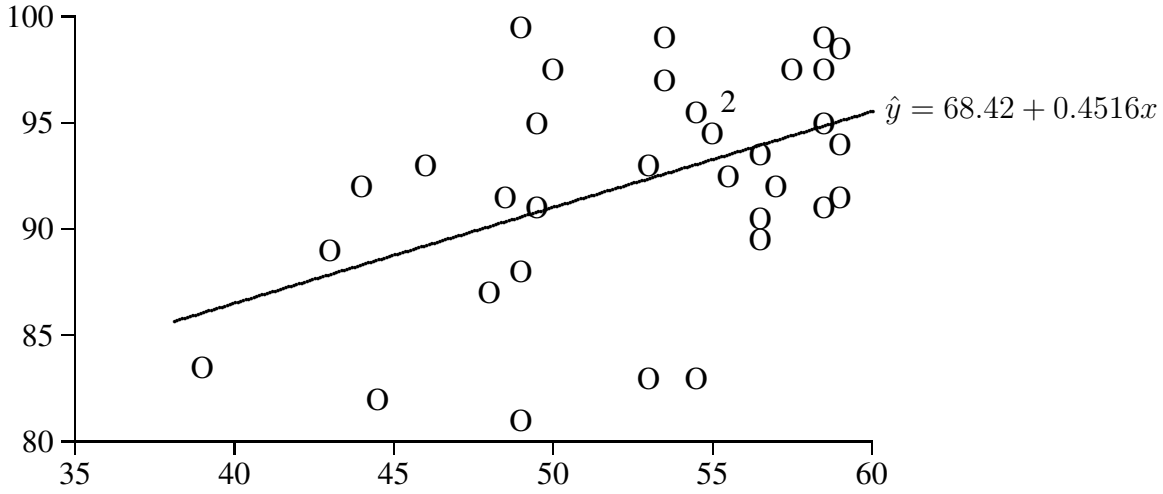
Anyways, with two predictors, we write the regression line as

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2.$$

You can now see why we have replaced b and m from math with b_0 and b_1: If we are allowing for an arbitrary number of predictors, we need to distinguish coefficients via subscripts to avoid running out of letters! Finally, statisticians put the intercept before the slope because if we later add additional predictors to our analysis, we like to place them at the (right) end of the equation rather than inserting them in the middle so that the intercept can maintain its lowly position at the end.

Figure 21.6 presents the scatterplot of the 35 pairs of exam grades with the graph of the regression line. Looking at this picture, I opine that the regression line appears to describe the data **well**, but **best?** In this course we will be happy simply to use the line; if you want to learn why it is the

Figure 21.6: Scatterplot of final exam score versus midterm exam score for $n = 35$ students, with the graph of the regression line $\hat{y} = 68.42 + 0.4516x$.



best line—based on the Principle of Least Squares—then you should read the optional Appendix near the end of this chapter.

We need to look more carefully at how the line describes the data. Go to the scatterplot and locate the case that has $x = 44.5$ and $y = 82.0$. For ease of presentation, I will call the student with these scores Sally—not his/her real name. Next, I substitute (plug-in) Sally's value of $x = 44.5$ into the regression line and obtain her value of $\hat{y}$:

$$\hat{y} = 68.42 + 0.4516(44.5) = 68.42 + 20.10 = 88.52.$$

We now have three numbers for Sally:

> Her midterm score: $x = 44.5$; her actual final exam score: $y = 82.0$; and her predicted final exam score: $\hat{y} = 88.52$.

Thus, her actual final exam score was 6.52 points lower than its prediction based on Sally's midterm exam score and the regression line. Looking at Figure 21.6, we *see* that Sally's 'O' is 6.52 points below the regression line. Sally's 'O' is quite far from the line, which tells us that the line does not describe Sally's data very well; or, if you prefer, this tells us that the prediction of Sally's final exam score is quite different from her actual score.

We now create a fourth number for Sally, to supplement her values of x, y and $\hat{y}$. We denote this new number by e and call it Sally's **residual**; its formula is below.

$$e = y - \hat{y}, \text{ which for Sally is } e = 82.0 - 88.52 = -6.52.$$

Sally's residual compares, via subtraction, her actual final score and her predicted final score.

Persons who are new to regression often wonder why statisticians define the residual as the difference $(y - \hat{y})$ instead of its negation, $(\hat{y} - y)$. One reason can be seen from Figure 21.6. When I view this figure, I naturally look at how the data points (the O's) are placed *relative to the regression line*. Sally's 'O' is below the line; down is the direction of smaller numbers, hence of the negative numbers. Thus, I want Sally's residual to be negative because it is below the regression line. Similarly, for any circle above the line, the residual is positive. If a circle is exactly on the regression line, then its residual is zero.

My extended discussion of Sally's data point can be modified for each of the other 34 students in my data set. With all this additional work, I will call upon my computer to help me. Table 21.3 presents output from Minitab for our current data set. I need to take a few minutes to walk you through this output; it contains a great deal of information!

Minitab begins by telling us:

```
The regression equation is: Final = 68.4 + 0.452 Midterm
```

This is Minitab's way of saying that the regression line is:

$$\hat{y} = 68.4 + 0.452x.$$

Each term in this equation has one fewer significant digit than I gave you earlier, but as we will see soon, Minitab also gives more precise values of the intercept and slope. Minitab's presentation is *quaint*, some might say *anachronistic*; it was created for an age, circa 1970, when a computer printer behaved like a typewriter that does not have a backspace key. Minitab could print a y or it could print a hat, but it could not print both. On the brighter side, Minitab does allow me to name my variables to make the output more user-friendly; I made the natural choices of *Midterm* for X and *Final* for Y.

Farther down the output, Minitab presents:

```
Predictor        Coef     SE Coef          T        P
Constant       68.420       7.963       8.59    0.000
Midterm        0.4516      0.1501       3.01    0.005
S = 4.635       R-Sq = 21.5%
```

We will ignore most of this until Chapter 22, but note that the *Coef* (short for coefficient) for the constant predictor is 68.42 and for the midterm is 0.4516, agreeing with my earlier reported values of the intercept and slope.

The remainder of the output is a listing for all 35 cases in the data set. The headings for all columns should make sense to you, excepting column 4; by *Fit* Minitab means the value of $\hat{y}$. You should note that Obs(ervation) 4 is Sally and Minitab reports the same values we determined earlier by hand. Well, except that Minitab gives one more digit of precision in columns 4 and 5.

Take a couple of minutes to peruse the information in the output. Note that the entries for observations 21 and 22 are identical; these are the two students who both scored $x = 55.5$ and $y = 96.0$. With the same value of x, they necessarily have the same value of $\hat{y}$; and with the same values of both $\hat{y}$ and y, they have the same residual.

Table 21.3: Edited Minitab output for the regression of final exam score on midterm exam score for 35 students.

The regression equation is: Final = 68.4 + 0.452 Midterm

Predictor	Coef	SE Coef	T	P
Constant	68.420	7.963	8.59	0.000
Midterm	0.4516	0.1501	3.01	0.005

S = 4.635 R-Sq = 21.5%

Obs	Midterm	Final	Fit	Residual
1	39.0	83.5	86.032	-2.532
2	43.0	89.0	87.838	1.162
3	44.0	92.0	88.290	3.710
4	44.5	82.0	88.515	-6.515
5	46.0	93.0	89.193	3.807
6	48.0	87.0	90.096	-3.096
7	48.5	91.5	90.322	1.178
8	49.0	99.5	90.548	8.952
9	49.0	88.0	90.548	-2.548
10	49.0	81.0	90.548	-9.548
11	49.5	91.0	90.773	0.227
12	49.5	95.0	90.773	4.227
13	50.0	97.5	90.999	6.501
14	53.0	83.0	92.354	-9.354
15	53.0	93.0	92.354	0.646
16	53.5	97.0	92.580	4.420
17	53.5	99.0	92.580	6.420
18	54.5	95.5	93.031	2.469
19	54.5	83.0	93.031	-10.031
20	55.0	94.5	93.257	1.243
21,22	55.5	96.0	93.483	2.517
23	55.5	92.5	93.483	-0.983
23	55.5	92.5	93.483	-0.983
24	56.5	93.5	93.934	-0.434
25	56.5	89.5	93.934	-4.434
26	56.5	90.5	93.934	-3.434
27	57.0	92.0	94.160	-2.160
28	57.5	97.5	94.386	3.114
29	58.5	99.0	94.838	4.162
30	58.5	95.0	94.838	0.162
31	58.5	91.0	94.838	-3.838
32	58.5	97.5	94.838	2.662
33	59.0	91.5	95.063	-3.563
34	59.0	98.5	95.063	3.437
35	59.0	94.0	95.063	-1.063

I think that it is time that I told you how I obtained the equation of the regression line. Again, if you want to see the algebra behind this, you should read the optional Appendix near the end of this chapter.

Result 21.2 (The equation of the regression line.) *The equation of the regression line is*

$$\hat{y} = b_0 + b_1 x, \tag{21.3}$$

where b_0 and b_1 are given by:

$$b_1 = r(s_2/s_1) \text{ and } b_0 = \bar{y} - b_1\bar{x}, \tag{21.4}$$

where r is the correlation coefficient defined in Equation 21.2.

Notice that we need five summary statistics to obtain the regression line:

$$\bar{x}, s_1, \bar{y}, s_2 \text{ and } r.$$

The first two of these summaries are for the x values—i.e., they ignore the y's—and the next two are for the y values. Only the last one, r, looks at how the x and y values are associated. This provides an example of why the correlation coefficient is important: The regression line is a function of: how the x's behave by themselves; how the y's behave by themselves; and the correlation coefficient. In other words, all we need to know about how the x's and y's influence each other (vary together) is contained in the value of r.

I will never ask you to obtain the regression line from a set of data by hand. In these *Course Notes*, I have shown you a site that will compute

$$\bar{x}, s_1, \bar{y} \text{ and } s_2,$$

and another site that will compute r. Also, I have shown you a site that will compute simply the regression line. I will not ask you to use these sites on the final exam and, hence, am not bothering to list them again.

Another, better, option is to use a statistical software package and a computer to obtain the regression line, as I do above with Minitab. I will show you more about how to interpret output from Minitab in Chapter 22. In the current chapter—and on the final—I will give you the five summary statistics you require to obtain the regression line **by hand**. Let me give you a couple of examples of the method.

Example 21.1 (The regression line for the exam scores data.) *After deleting the isolated case, for the 35 students pictured in Figure 21.5, I computed:*

$$\bar{x} = 52.786, s_1 = 5.295, \bar{y} = 92.257 \text{ and } s_2 = 5.154.$$

Recall that I previously told you that $r = 0.464$. We can now evaluate Equation 21.4:

$$b_1 = 0.464(5.154/5.295) = 0.4516 \text{ and } b_0 = 92.257 - 0.4516(52.786) = 68.4188.$$

Thus, the regression line is

$$\hat{y} = 68.42 + 0.4516x,$$

as stated earlier.

Example 21.2 (The regression lines for the tarantula and small hunter data sets.) *Please refer to the scatterplot of heart rate versus body weight for the $n = 6$ tarantulas in Figure 21.1. Recall that I previously told you:*

$$\bar{x} = 11.277, s_1 = 1.955, \bar{y} = 11.833, s_2 = 1.472 \text{ and } r = -0.872.$$

We can now evaluate Equation 21.4:

$$b_1 = -0.872(1.472/1.955) = -0.6566 \text{ and } b_0 = 11.833 + 0.6566(11.277) = 19.24.$$

Thus, the regression line is

$$\hat{y} = 19.24 - 0.6566x.$$

With a similar argument, for the small hunters we have:

$$\bar{x} = 0.101, s_1 = 0.049, \bar{y} = 92.1, s_2 = 21.5 \text{ and } r = 0.397.$$

We can now evaluate Equation 21.4:

$$b_1 = 0.397(21.5/0.049) = 174.2 \text{ and } b_0 = 92.1 - 174.2(0.101) = 74.51.$$

Thus, the regression line is

$$\hat{y} = 74.51 + 174.2x.$$

Figure 21.7 presents the scatterplot for the tarantulas and for the small hunters with their regression lines. Don't worry about being able to draw lines on a scatterplot; if you ever get a job doing regression analysis, (I hope) you will have computer software to help!

Now that you have three examples of regression lines—exam scores and two categories of spiders—I want to make some general comments about the regression line.

First, take the equation of the regression line,

$$\hat{y} = b_0 + b_1 x,$$

and replace the symbols b_0 and b_1 by the expressions in Equation 21.4; we get:

$$\hat{y} = \bar{y} + r(s_2/s_1)(x - \bar{x}). \tag{21.5}$$

Let's suppose that we have a case for which $x = \bar{x}$; in words, this case is average—one might say *mediocre*—on its value of the predictor. What is its predicted response? Well, substituting $\bar{x}$ for x into Equation 21.5 we obtain:

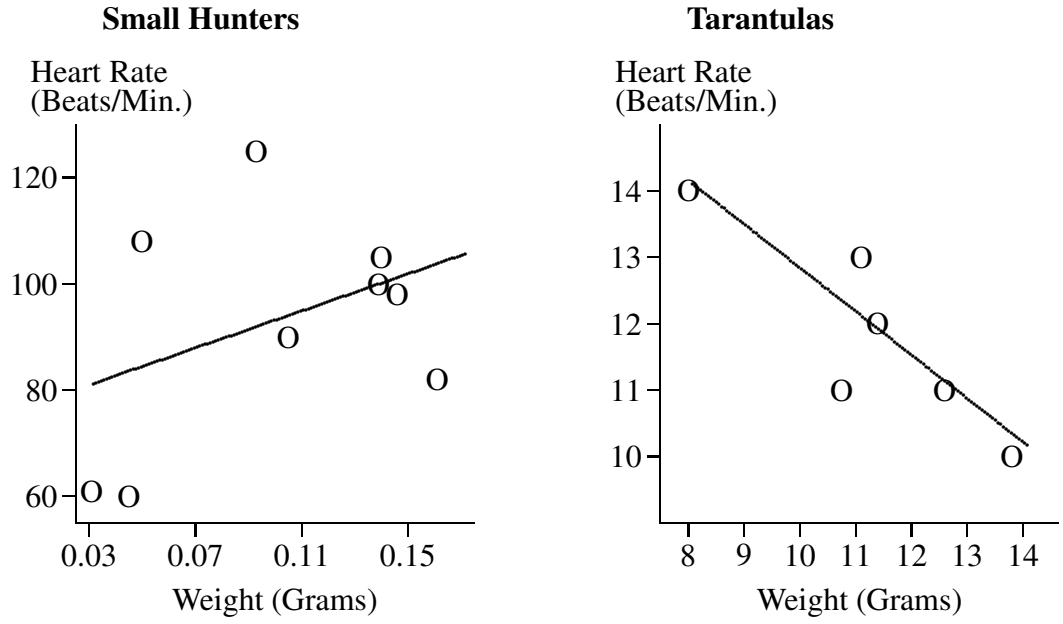
$$\hat{y} = \bar{y} + b_1(\bar{x} - \bar{x}) = \bar{y} + 0 = \bar{y}.$$

Thus, if a case is mediocre on x then the regression line predicts that it will be mediocre (i.e., equal $\bar{y}$) on y. I have suggested that this result be labeled **The Law of the Preservation of Mediocrity**, but, so far, without success.

Visually, the Law tells us that the regression line must pass through the point $(\bar{x}, \bar{y})$.

Actually, my Law of the Preservation of Mediocrity has some merit. Let me explain. As presented in these notes, the regression line is the result of applying the Principle of Least Squares. But this principle is, at least in part, motivated by mathematical convenience. (Stop shouting all you Ph.D.'s in Math!) The obvious practical question is:

Figure 21.7: Scatterplot and the regression line of heart rate (beats per minute) versus body weight (grams) for small hunters and for tarantulas.



Does this mathematical convenience yield *sensible* answers?

Based on the Preservation of Mediocrity, I can state, "Perhaps." Here is why. If the Preservation of Mediocrity were **not** true for the regression line, that would seem stupid! After all, how could it make sense with a linear relationship to predict that a mediocre x yields an above [below] average y!

21.3 The Regression Effect and the Regression Fallacy

I will introduce these ideas with a real data set. This is a large data set; I have pairs of numbers for $n = 124$ cases. Having a large amount of data is useful for this section. Another good feature is that the values of $\bar{x}$ and $\bar{y}$ are almost identical and the standard deviations s_1 and s_2 are similar. As you will see, the material in this section is easier to understand if $\bar{x} = \bar{y}$ and $s_1 = s_2$, but such exact agreement is rare in real data.

Alas, these data have some bad features. First, they are data from Major League Baseball. Thus, it is not classically a biology example, but baseball is played by human animals! Second, if you are a baseball fan, you will probably be annoyed when you learn that the data are more than 25 years old!

Thus, you might wonder: Why am I using these very old sports data? To be honest, any such example is a lot of work, I am running out of time to complete these notes and I performed the analysis of these data many years ago.

I will refer to my data set as the Batting Averages Study. Its cases are the 124 baseball players who had 200 or more official at bats during both the 1985 and 1986 American League seasons. Each player's two variables are his batting averages (number of hits divided by official number of at bats) for the two seasons [2]. I will let Y [X] denote the 1986 [1985] batting average. In short, I want to use a player's 1985 batting average to predict his 1986 batting average.

If you are not a baseball fan, there are three things you should realize:

1. A batting average is a proportion of successes, not a mean. (Note to literal math-types: yes, every proportion is in a sense a mean, but if we were happy with that name, we would never use the word proportion.)

2. When comparing two batting averages, the larger one is better.

3. A batting average of say, 0.325, is never read literally as 325 thousandths; it is read 325, as in, "He batted 325." (What did you expect from people who call a proportion an average?)

Figure 21.8 presents a scatterplot of the 124 cases and the regression line: $\hat{y} = 0.095 + 0.633x$. When I look at the scatterplot I see one definite isolated case and two possibilities. The definite isolated case is Floyd Rayford who batted $x = 0.306$ in 1985 and $y = 0.176$ in 1986; going from the ninth largest batting average in 1985 to the lowest in 1986 is unusual! (In 1987, Mr. Rayford batted 0.220 in only 50 at bats and his Major League career was over. He went on to have a long career as a minor league coach; those who can't ....) The two possible isolated cases are: Wade Boggs, $(0.368, 0.357)$, who had the highest batting average both years and Don Mattingly, $(0.324, 0.352)$, who had the third highest batting average in 1985 and the second highest in 1986. In short, the data points for Boggs and Mattingly are isolated because they were great both years.

I will include all 124 cases in my analysis, although one might argue that Floyd Rayford should be deleted.

The five summary statistics for the data are:

$$\bar{x} = 0.266, s_1 = 0.028, \bar{y} = 0.264, s_2 = 0.032 \text{ and } r = 0.554.$$

As I stated earlier, the means are nearly identical and the standard deviations are similar; the mean of the batting averages went down a bit and the standard deviation of the batting averages went up a bit, both comparisons from 1985 to 1986.

Please forgive me the briefest of digressions. Let me tell you about a group of people I find very annoying. I call them the **naive predictors**. A naive predictor believes that the future **should be exactly the same as the past**. For example, if today I make 58 out of 100 free throws, a naive predictor thinks that tomorrow I **should make exactly 58 out of 100 free throws and if I fail to do so, then *something is wrong***! To a naive predictor, if I make 59 out of 100 free throws tomorrow, then there must be **a reason why**. It must be very frustrating to be a naive predictor!

Batting 300 (or higher) is considered to be quite good in baseball. The 12 players who batted 300 or higher in 1985 are listed in Table 21.4. A naive predictor would expect y to equal x for these 12 players (indeed for all players). Notice, however, that for 10 of the 12 players the 1986 batting average was lower than the 1985 batting average. And not just a little bit lower: 67 points lower for Salas, 53 for Iorg, 51 for Henderson and the massive 130 for Rayford. The mean decrease for

Figure 21.8: Scatterplot of 1986 versus 1985 American League batting average with regression line: $\hat{y} = 0.095 + 0.633x$.

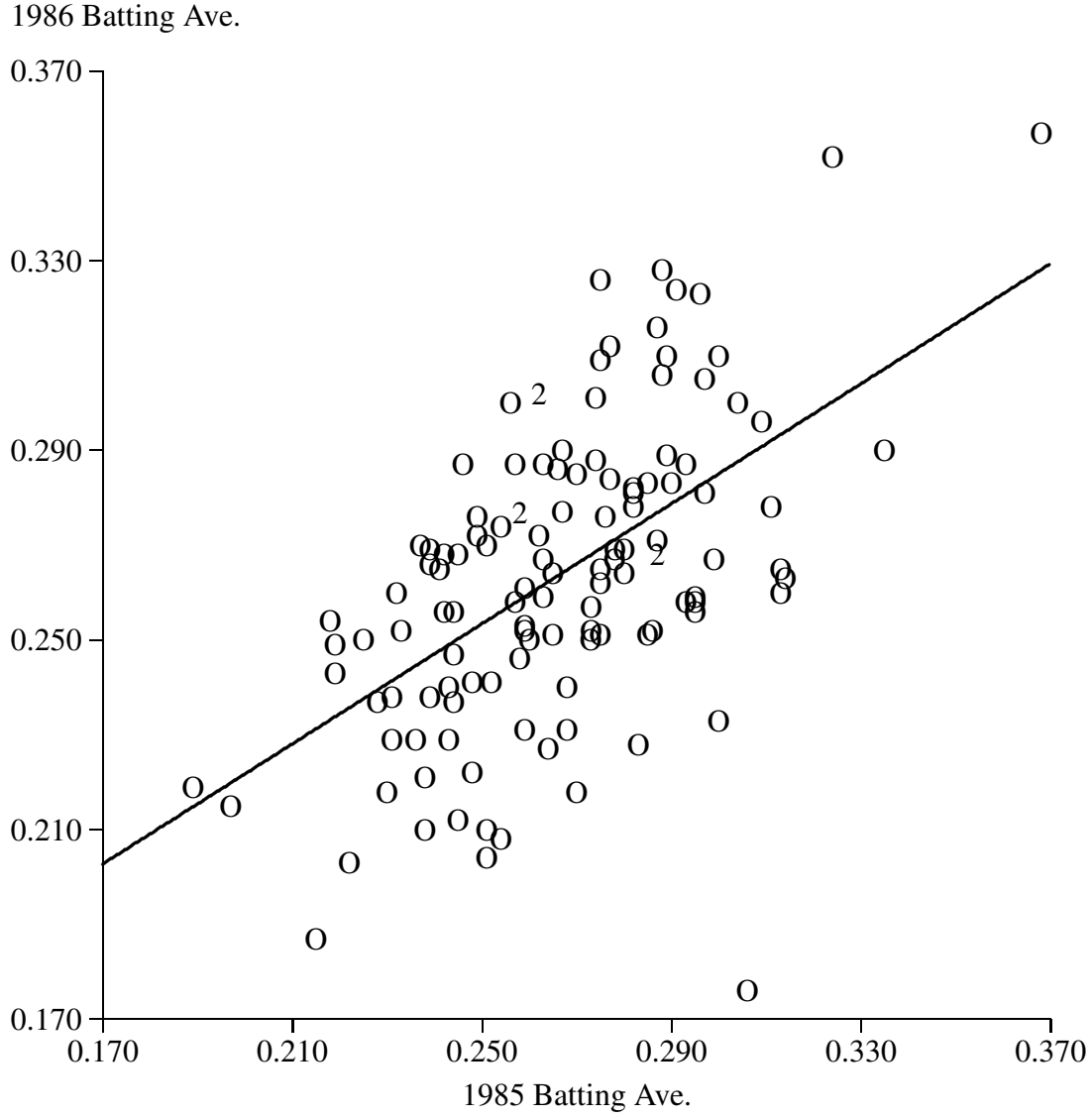


Table 21.4: 1985 and 1986 batting averages for the 12 players who batted 300 or more in 1985.

Name	1985 x	1986 y	Change $y-x$	$\hat{y}$	Residual e
Floyd Rayford	0.306	0.176	-0.130	0.289	-0.113
Mark Salas	0.300	0.233	-0.067	0.285	-0.052
Garth Iorg	0.313	0.260	-0.053	0.293	-0.033
Rickey Henderson	0.314	0.263	-0.051	0.294	-0.031
Wayne Tolleson	0.313	0.265	-0.048	0.293	-0.028
George Brett	0.335	0.290	-0.045	0.307	-0.017
Brett Butler	0.311	0.278	-0.033	0.292	-0.014
Harold Baines	0.309	0.296	-0.013	0.291	0.005
Wade Boggs	0.368	0.357	-0.011	0.328	0.029
Juan Beniquez	0.304	0.300	-0.004	0.287	0.013
Phil Bradley	0.300	0.310	0.010	0.285	0.025
Don Mattingly	0.324	0.352	0.028	0.300	0.052
Mean:			-0.035		-0.014
Mean without Rayford:			-0.026		-0.005

these 12 players is 35 points, a huge amount for a batting average. Even deleting Rayford, the mean decline is 26 points.

Why did this happen? First, let's look at some obvious possibilities.

1. Perhaps batting averages were simply much lower in 1986 than they were in 1985. **No**. As we saw, the mean batting average in 1986, 0.264, is only two points less than the mean batting average in 1985, 0.266. A two point drop overall does not explain a 35 point drop for the top 12 hitters of 1985!

2. Perhaps the spread in the batting averages decreased dramatically from 1985 to 1986. The consequences of this declining spread would include that the most extreme values in 1985 would shrink towards the mean in 1986. **No**. As we saw, the standard deviation of the batting averages in 1986, 0.032, is actually *larger than* the standard deviation of the batting averages in 1985, 0.028.

Before I explain why these changes occurred, let's look at the nine worst hitters—based on batting average—in 1985 and see how they did in 1986. The data are presented in Table 21.5. While the changes in this new table are less dramatic than what we had earlier—there is no anti-Rayford or anti-Salas in this group—there is still a notable pattern: seven of the nine hitters improved in 1986 and the mean change for the nine hitters is an improvement of 14 points.

To summarize, the naive predictors are too optimistic about good hitters and too pessimistic about bad hitters.

How do the regression line predictions perform for these 21 extreme (in 1985) hitters?

Table 21.5: 1985 and 1986 batting averages for the nine players who batted 228 or lower in 1985.

Name	1985 x	1986 y	Change $y-x$	$\hat{y}$	Residual e
Rick Manning	0.218	0.254	0.036	0.233	0.021
Dick Schofield	0.219	0.249	0.030	0.234	0.015
Rob Wilfong	0.189	0.219	0.030	0.215	0.004
Greg Gagne	0.225	0.250	0.025	0.237	0.013
Steve Buechele	0.219	0.243	0.024	0.234	0.009
Julio Cruz	0.197	0.215	0.018	0.220	-0.005
Pat Sheridan	0.228	0.237	0.009	0.239	-0.002
Darryl Motley	0.222	0.203	-0.019	0.236	-0.033
Gorman Thomas	0.215	0.187	-0.028	0.231	-0.044
Mean:			0.014		-0.002

- For the 12 best hitters in 1985, the regression prediction, $\hat{y}$, was too large—the residual is negative—for seven players and too too small—the residual is positive—for five players. The mean of these 12 residuals is -0.014 and, if we exclude Floyd Rayford, the mean of the remaining residuals is -0.005.

  Clearly, the least squares line is much better than the naive predictors for the players in Table 21.4.

- For the nine worst hitters in 1985, the regression prediction, $\hat{y}$, was too large—the residual is negative—for four players and too small—the residual is positive—for five players. The mean of these nine residuals is nearly zero, -0.002.

  Clearly, the least squares line is much better than the naive predictors for the players in Table 21.5.

Here is what generalizes about the above example on baseball players and batting averages. The following is **not a mathematical result**, although, as you will see soon, there is a math result that supports it. The following is an **empirical result**; it is true for real data. Indeed, if you can find data, real or pretend, that satisfies my conditions, but violates my conclusion, let me know right away! Who knows, this could result in an

> **insert-your-name-here Paradox**, similar to the Simpson's Paradox you learned about earlier.

- **Conditions:**
  - There is a linear relationship in the data between X and Y.

- The means and approximately equal, $\bar{x} \approx \bar{y}$, and the standard deviations are approximately equal, $s_1/s_2 \approx 1$. The correlation coefficient is positive, but smaller than 1. (A similar result is true if the correlation coefficient is negative, but larger than -1, but I want to keep this simple.)

- **Conclusion:**

  - For cases with an x larger than the mean $\bar{x}$, the values of y tend to be smaller than x but larger than the mean $\bar{y}$.
  - For cases with an x smaller than the mean $\bar{x}$, the values of y tend to be larger than x but smaller than the mean $\bar{y}$.

As best I can tell, this result first appeared in a 1886 paper by (later Sir) Francis Galton, [3]. Here is what Galton did.

A case for Galton consisted of a pair of men, an adult father and his adult first born son. He took X to be the height of the father and Y to be the height of the son. (In Galton's day what we call height was called stature.) My conditions above are met by Galton's data and my conclusion is true for his data too. Namely, Galton noted that the sons of extremely tall fathers also tended to be tall, but shorter than their fathers. Also, he noted that the sons of extremely short fathers also tended to be short, but taller than their fathers. Galton's conclusion? I think that the title of his paper says it all:

Regression Towards Mediocrity in Hereditary Stature.

Apparently, Galton thought that eventually all Englishmen would be the same height; after all, if tall fathers beget shorter sons and short fathers beget taller sons, what else can one conclude? Galton's error was in failing to note that s_2 was approximately equal to s_1. If, indeed, men were regressing to the same height, then s_2 should be noticeably smaller than s_1. (Also, if the phenomenon of regression of heights was actually occurring from generation to generation, why weren't all men the same height by 1886?) In any event, Galton's use of the word regression has persisted to this day; hence, the name of this chapter.

This leaves the question: If the son's of tall men are becoming shorter, but not too much shorter; and the son's of short men are becoming taller, but not too much taller; how is it that s_2 is **not** smaller than s_1? The answer is quite simple: while the *extremes are collapsing to the mean* this is counteracted by the fact that the son's of average height fathers show more spread; some are much taller than average dad; some are much shorter than average dad; and some are approximately the same height as average dad.

Historical Note: Please do not construe the above as a criticism of Galton or his work. I am unqualified for either task, but do note that he was a giant in the field of quantitative social sciences. Stephen Jay Gould has written eloquently on the difficulty with judging scientists from another era; if you are interested in this topic, read his book *The Mismeasure of Man* [4], which is on my list of the five best books I have ever read. (I have read nearly 2,000 books.) Not only is the above tale not a criticism, I cannot swear to its accuracy. As with FDR and Ronald Reagan, supporters and detractors of Galton have very different views of his errors, if any. In particular, my

statement that Galton believed all Englishmen would eventually be the same height is disputed. I personally would be surprised if he believed this at his death in 1911, but in 1886? Who knows? As Muhammad Ali once said,

> A man who views the world the same at fifty as he did at twenty has wasted thirty years of his life.

Even **if** part of my tale is above is historically inaccurate, I believe it is a good way to introduce you to the topic of the regression effect. I am open to suggestions for a better story.

I will now show you the math behind the earlier result, as promised.

It is insightful to rewrite Equation 21.5 as follows:

$$\hat{y} = \bar{y} + r(s_2/s_1)(x - \bar{x}) \text{ becomes}$$

$$\hat{y} - \bar{y} = r(s_2/s_1)(x - \bar{x}) \text{ becomes}$$

$$\frac{\hat{y} - \bar{y}}{s_2} = r(\frac{x - \bar{x}}{s_1}). \tag{21.6}$$

This last is equation is **not** designed for "Plugging in x to obtain $\hat{y}$." It is designed to help us *understand* the regression line better. This improved understanding requires a fair amount of work.

It will help if we use a specific value of r, say $r = 0.554$ from our batting average data. Let's suppose that we have a hitter whose 1985 batting average is

$$x = \bar{x} + s_1 = 0.266 + 0.028 = 0.294.$$

Actually, none of the 124 players in my data set had $x = 0.294$, but I don't mind because I simply am trying to explore Equation 21.6.

As discussed earlier, a naive predictor would predict 0.294—i.e., no change—for the 1986 batting average of this player. A somewhat more sophisticated naive predictor might reason as follows:

> This player achieved one standard deviation above the mean on x; thus, I predict that he will achieve one standard deviation above the mean on y.

Thus, the more sophisticated naive predictor would obtain

$$\bar{y} + s_2 = 0.264 + 0.032 = 0.296.$$

The regression line disagrees with both versions of the naive predictor. The regression line says that the predicted value of y, i.e., $\hat{y}$, equals only $r = 0.554$—not one—standard deviations more than the mean:

$$\frac{\hat{y} - \bar{y}}{s_2} = r \text{ or } \hat{y} = \bar{y} + rs_2 = 0.264 + 0.554(0.032) = 0.264 + 0.018 = 0.282.$$

Let me share with you my *picturesque* interpretation of Equation 21.6. I will give it in terms of the batting average study and its $r = 0.554$; I trust that you will be able to extend my interpretation to other studies and other values of r that are strictly between 0 and 1.

Consider again the (fictitious) player who had $x = 0.294$. This is a good batting average for 1985; it is one standard deviation larger than the mean 1985 batting average of the 124 players. I anticipate that this man will also be good in 1986; good *but not as good as he was in 1985*. In particular, I predict that only 55.4% (the percentage version of $r = 0.554$) of *whatever* made him better than average in 1985 will **persist** to 1986. My view of the world is: skill persists, luck does not. Thus, my picturesque interpretation is that 55.4% of what made him special in 1985 was skill, the other 44.6% must have been luck. As a statistician, by *luck* I mean what most people mean plus I include all factors that are not included in my analysis. An obvious factor that I left out of my batting average study was the age of the player. For example, George Brett was a great Hall-of-Fame baseball player, but the drop in his batting average from $x = 0.335$ to $y = 0.290$ was—in my opinion—in part due to his turning 33 years-old during the 1986 season. Brett had some good seasons after 1985, but nothing compared to his performance before 1986.

The consequences on predictions of the presence of r in Equation 21.6 is called the **regression effect**. (Remember if r was not there, or was equal to one, then the least squares predictions would be the sophisticated naive predictions.) The **regression fallacy** is the mistake of believing that the regression effect must be due to something other than simply the fact that r is smaller than 1.

21.4 Some Comments on the Regression Line

In this section I will gather together and present several *loose ends* that I failed to mention earlier in this chapter. Sorry, but I could not find a way to mention these earlier without disrupting the flow of ideas.

21.4.1 Don't Round the Predictions!

Do you remember that in Chapter 14, I showed you how to predict the total number of successes in future Bernoulli trials or future observation of a Poisson Process? At that time, I recommended that you round your answers; for example, predicting that the total number of successes would be between 53.7 and 81.2 seemed silly because the number of successes would be, perforce, an integer.

Now, fast forward to the exam scores example of this chapter. Recall that the final exam was graded in one-half point increments; as a result, for example, 92.0 was a possible score on the final and 92.5 was a possible score on the final, but any number between these two was impossible. Thus, if, say, you obtained $\hat{y} = 92.3$ it would seem that my advice would be to round this to 92.5. **Wrong!** In regression we do not round our value of $\hat{y}$, we are happy to leave it equal to an impossible value. Why? There are two reasons:

1. First, a minor reason: We obtain the $\hat{y}$'s by applying the Principle of Least Squares. If we go rounding these off, we are no longer following the principle.

2. Now the more important reason: If we round off the values of $\hat{y}$ to possible values, then the *regression line* is no longer a line; for my exams example, it will look like a staircase.

Of course, anytime we deal with measurements, there will be rounding. Usually, with complicated computations there will be rounding. Either of these is fine; just don't round in order to obtain a *possible response*. Nobody cares about that.

21.4.2 We Call it the Regression *Line*, but ...

Lines in mathematics are infinite, they go on forever. Lines in Statistics don't. More accurately, statisticians should refer to the regression line as the *regression line segment*.

For every set of data we have examined, there is a limited range of x values. For two examples:

- In the exam scores data set, the midterm scores range from a low of 39.0 to a high of 59.0.
- In the batting averages study, the 1985 batting averages range from a low of 0.189 to a high of 0.368.

For the spiders, the exams and the batting averages, I looked at a scatterplot and declared the relationship to be linear. Obviously, I have **no empirical evidence** on whether the relationship is linear outside the range of the x's in my data set. The following example makes this point quite well, but please don't call me sadistic for showing it to you. Paraphrasing President Nixon, "I would never do this experiment; it would be wrong."

Example 21.3 (Fish activity and water temperature.) *These data are from a student's project many years ago. Sadly, I don't remember the student's name; if you are out there and read this, let me know!*

My student was a biology major and she reported that she had read in a textbook,

> *As water temperature increases, fish activity increases.*

My student owned an aquarium with a water temperature control and she decided to collect data on her fish to investigate the textbook's claim. She let Y denote fish activity and X denote water temperature in degrees Fahrenheit. One of the most interesting things she learned in her study was the difficulty in measuring fish activity! She did, however, manage to obtain a good set of valid data, a scatterplot of which is presented in Figure 21.9.

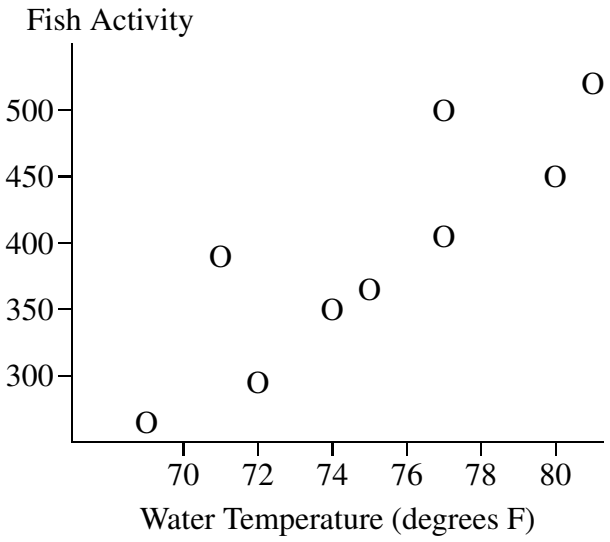
Look at the scatterplot for a moment; what do you see? Well, I see no isolated cases. I see an increasing linear relationship between x and y and I can almost see the graph of the regression line superimposed on the scatterplot. It appears that for water temperature in the range of 69 to 81 degrees, the textbook is correct.

Here is my question for you:

> Do you believe that it is scientifically valid to extend the regression line to temperatures above 81 degrees? To 83 degrees? To 90 degrees? To 212 degrees?

Well, obviously, at 212 degrees there won't be any fish activity! Similarly, there won't be any fish activity at 32 degrees.

Figure 21.9: Scatterplot of fish activity versus water temperature (degrees F).



21.4.3 The Regression of X on Y

In math, suppose you have the equation $y = mx + b$ for $m \neq 0$. You can then solve for x in terms of y and obtain $x = (y - b)/m = (1/m)y - b/m$. Note that the product of the slopes of these two lines is $m(1/m) = 1$.

In Statistics, the situation is a bit more complicated. First, let me note that in many scientific problems, it is *reasonable* to consider using Y to predict X. For example, on exams it is possible that a student will miss a midterm and the teacher and student decides to use the final to predict the midterm score. For the batting averages study, one might want to use the 1986 batting average to predict the 1985 batting average. To this end, let's look at the representation of the regression line given in Equation 21.5:

$$\hat{y} = \bar{y} + r(s_2/s_1)(x - \bar{x}).$$

To obtain the regression line for using y to predict x, we simply interchange the roles of y and x—which includes interchanging the roles of s_1 and s_2—in the above and obtain the following result. (Recall that changing the roles of x and y has no effect on the correlation coefficient r.)

Result 21.3 (The regression line for using y to predict x.) *The regression line for using y to predict x is*

$$\hat{x} = \bar{x} + r(s_1/s_2)(y - \bar{y}). \tag{21.7}$$

This new regression line is **not** just obtained by taking the old regression line (y on x) and solving for x. The easiest way to see this—i.e., that avoids messy algebra—is to note that the product of the slopes of these two regression lines is:

$$r(s_2/s_1) \times r(s_1/s_2) = r^2,$$

which is smaller than 1 unless r equals $+1$ or -1; in other words, unless all data points lie exactly on a straight line. If the lines are mathematically equivalent, then the product of these slopes must equal 1, as shown at the beginning of this subsection.

21.5 Summary

In this chapter, we consider scientific problems for which each unit—called a case now—yields two numbers. Data from n cases are represented by

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n).$$

Usually the variables, X and Y, are viewed asymmetrically by the researcher. In particular, Y is viewed as the response and X as the predictor. These labels convey two features:

- That the researcher has a greater interest in Y than in X; and/or
- The value of X is obtained primarily to help one better understand Y.

If the researcher truly views the variables symmetrically—for example, if the two numbers are the adult IQs of first versus second born identical twins—then the researcher should assign the labels—for my example, say, X for the first born—in any arbitrary manner. In this situation, however, remember that each assignment gives a different regression line. (The value of the correlation coefficient, however, is not affected by the assignment.)

Following what we did in Chapter 1, begin by drawing a picture of the data. In Chapter 1 with one variable, our first picture was the dot plot. The scatterplot of this chapter is an extension of the idea of a dot plot. I will not ask you to draw a scatterplot by hand, but if I give you a particular pair (x, y) from a data set, you need to be able to locate it in the scatterplot.

Given a scatterplot for a set of data, the first thing to do is to look for one or more isolated cases. Next, look at the scatterplot—perhaps the *new* scatterplot after deleting one or more isolated cases—and determine whether the pattern it reveals is linear. In Chapter 21 and most of Chapter 22, we only consider data sets that possess a linear relationship between X and Y.

The correlation coefficient, defined in Equation 21.2, tells us the **direction** and **strength** of the linear relationship in the data. Make sure that you understand the six properties of the correlation coefficient that are given in Result 21.1 on page 546. Note the 12 prototypical scatterplots and their correlation coefficients given in Figure 21.4 on page 548. In particular, if given a scatterplot, you should be able to determine its approximate correlation coefficient.

The correlation coefficient, r, along with the means and standard deviations of the x's and the y's—i.e., five numbers in total—allow us to find the line that best describes the data set. This best line is determined by an application of the Principle of Least Squares. You are not responsible for either understanding or applying the Principle of Least Squares, but if you are interested in these issues, see the Appendix near the end of this chapter.

The best line is called the regression line and it is given in Result 21.2 on page 556. I prefer the following representation of it, given in Equation 21.5:

$$\hat{y} = \bar{y} + r(s_2/s_1)(x - \bar{x}).$$

This expression allows us to see easily the Law of the Preservation of Mediocrity, namely that if a case has $x = \bar{x}$, then the case's $\hat{y} = \bar{y}$. In other words, the regression line passes through the point $(\bar{x}, \bar{y})$.

In the data set, case i has two numbers: x_i and y_i. By substituting x_i into the regression line for x, we get a third number for the case, its predicted value:

$$\hat{y}_i = b_0 + b_1 x_i.$$

Each case also has a residual:

$$e_i = y_i - \hat{y}_i,$$

giving each case a fourth number.

The residual equals 0 if, and only if, the prediction is perfect; i.e., $\hat{y}_i = y_i$. A positive [negative] residual means that the actual y_i is larger [smaller] than the predicted value $\hat{y}_i$. In terms of the scatterplot, a case is exactly on the line if, and only if the prediction is perfect. A case is above [below] the regression line if its residual is positive [negative].

Another way to write the regression line is given in Equation 21.6:

$$\frac{\hat{y} - \bar{y}}{s_2} = r(\frac{x - \bar{x}}{s_1}).$$

For r strictly between 0 and 1, this equation shows the **regression effect**:

> For any $c > 0$ [$c < 0$] and any case with x equal to $\bar{x} + cs_1$—in words, the value of x is c standard deviations larger [smaller] than the mean of the x's—the predicted value of y is only $r \times c$ standard deviations larger [smaller] than mean of the y's.

In 1886, Francis Galton discovered this phenomenon and termed it *regression towards mediocrity* and the term regression stuck. The **regression fallacy** is to believe that the regression effect has a cause other than the fact that r is smaller than one.

Finally, the chapter ends with a few unconnected remarks, with examples:

1. Report the value of $\hat{y}$ without regard to whether it is a possible value of y.
2. The regression line does not extend infinitely.
3. It is mathematically possible to regress X on Y; whether this makes sense scientifically will depend on the problem. An important feature of the two regression lines is that the product of their respective slopes is r^2.

21.6 Practice Problems

1. Refer to the exam data on 36 students, presented in Figure 21.5. In this chapter, I found the regression line for the 35 cases that remained after I deleted the isolated case with $x = 35.5$ and $y = 95.5$. You are given the following summary statistics for the entire set of 36 cases:

$$\bar{x} = 52.31, s_1 = 5.961, \bar{y} = 92.35, s_2 = 5.109 \text{ and } r = 0.353.$$

(a) Obtain the regression line using the midterm exam score to predict the final exam score for the 36 cases.

(b) Use the regression line to obtain the predicted value of y for the following four values of x: 46.0, 50.0, 55.0 and 59.0. Compare your four predictions to the values in Table 21.3. Comment.

2. Figure 21.10 presents a scatterplot of the public's rating of a movie versus the rating given by a panel of movie critics for $n = 20$ movies, circa 1991. The worst possible rating is one star and the best possible rating is four stars. Compare this scatterplot to the 12 scatterplots presented in Figure 21.4. One of the following numbers is the correlation coefficient for these data; which one is it?

$$-0.83, -0.36, 0.00, 0.36, 0.83.$$

3. Figure 21.11 is a scatterplot of the number of victories (in 82 games) in 1991–92 versus the number of victories (again, in 82 games) in 1990–91 for the $n = 27$ NBA (National Basketball Association) teams.

   One of the following numbers is the correlation coefficient for these data; which one is it?

$$0.000, 0.052, 0.302, 0.724, 1.000.$$

4. Sometimes we don't need to use fancy statistics to learn from a scatterplot. We simply need to avoid dumb ideas, as this example shows.

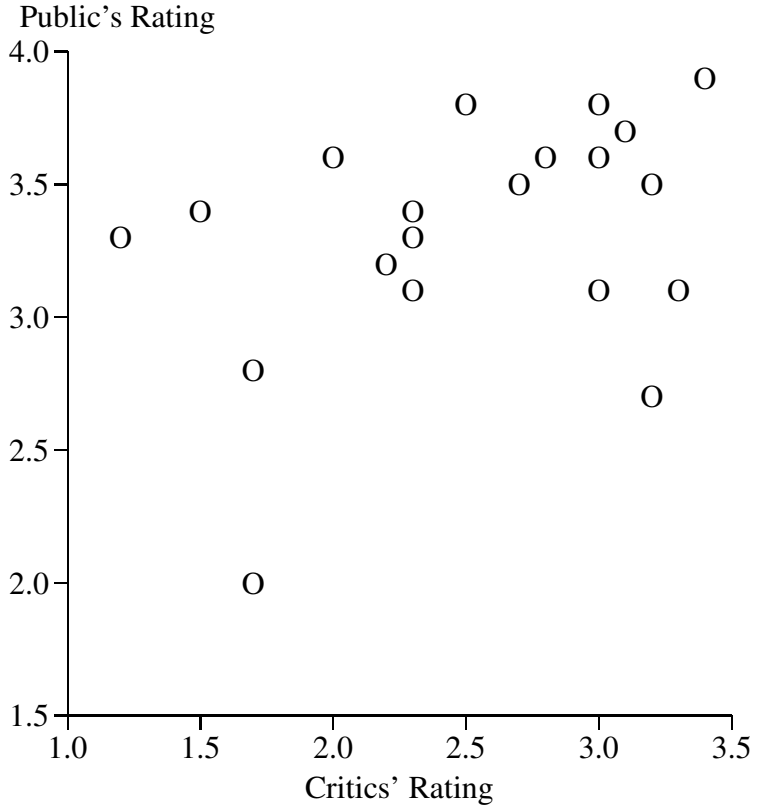
   The data and details of this example are taken from [5] which provides additional information for the interested reader. On January 28, 1986, shortly after lift-off one of the rocket boosters on the space shuttle *Challenger* exploded resulting in the death of the seven crew members. A Presidential Commission concluded that the disaster was caused by the failure of an O-ring in a field joint on the rocket booster. The Commission further concluded that the failure was due to a faulty design which made the O-ring unacceptably sensitive to a number of factors, including temperature. The O-rings had been damaged on several of the 24 previous shuttle program flights. Figure 21.12 is the scatterplot of the number of incidents of thermal distress to field joint O-rings versus the launch temperature for 23 shuttle flights before the *Challenger* disaster. (The hardware from the fourth shuttle flight was lost at sea.)

   (a) Write a few sentences that describe what the scatterplot reveals.

   (b) Based on the scatterplot (and not hindsight) criticize the decision to launch the *Challenger* when the temperature was 31 degrees.

   (c) Unfortunately, the night before the *Challenger* launch when managers discussed the effect of temperature on field joint O-rings, they decided the launches that yielded $y = 0$ were irrelevant. Look at the seven cases in Figure 21.12 that have $y > 0$; is there a convincing relationship between Y and X?

Figure 21.10: Scatterplot of the public's versus critics' rating of 20 movies, circa 1991.



5. (See Practice Problem 3 above.) For the data in Figure 21.11:

$$\bar{x} = \bar{y} = 41.00, s_1 = 12.96, s_2 = 13.08 \text{ and } r = 0.724.$$

(a) Explain why it is no surprise that $\bar{x} = \bar{y} = 41.00$.

(b) Find the equation of the regression line. Use the form

$$\hat{y} = \bar{y} + r(s_2/s_1)(x - \bar{x}).$$

Explain this equation to a basketball fan who has not read this chapter.

(c) In the 1990–91 season: Chicago won 61 games; Seattle won 41 games; and Miami won 24 games. Calculate the value of $\hat{y}$ for each of these teams.

(d) In the 1991–92 season: Chicago won 67 games; Seattle won 47 games; and Miami won 38 games. Calculate the value of the residual e for each of these teams.

Figure 21.11: Scatterplot of the number of victories in the 1991–92 season versus the number of victories in the 1990-91 season for 27 National Basketball Association (NBA) teams.

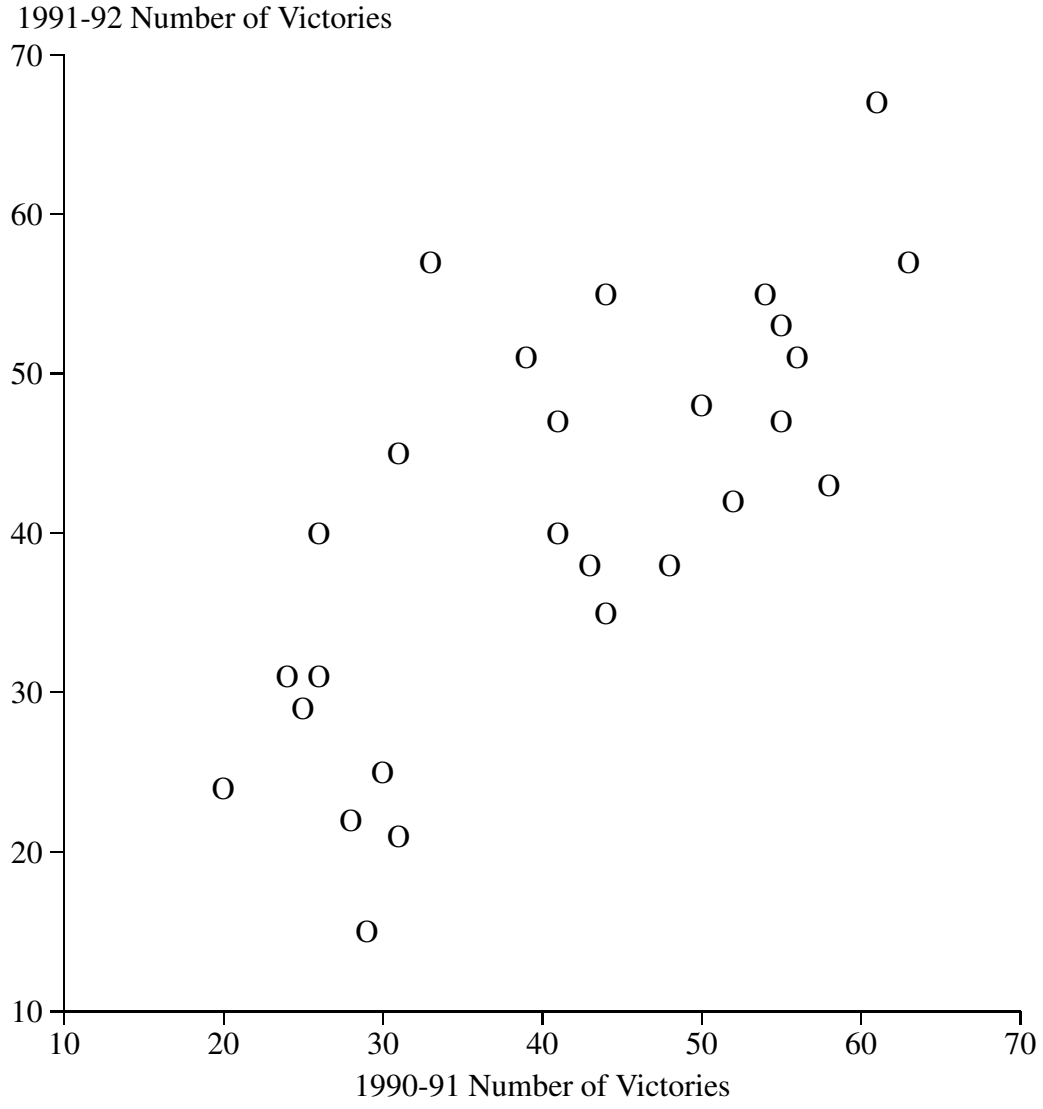
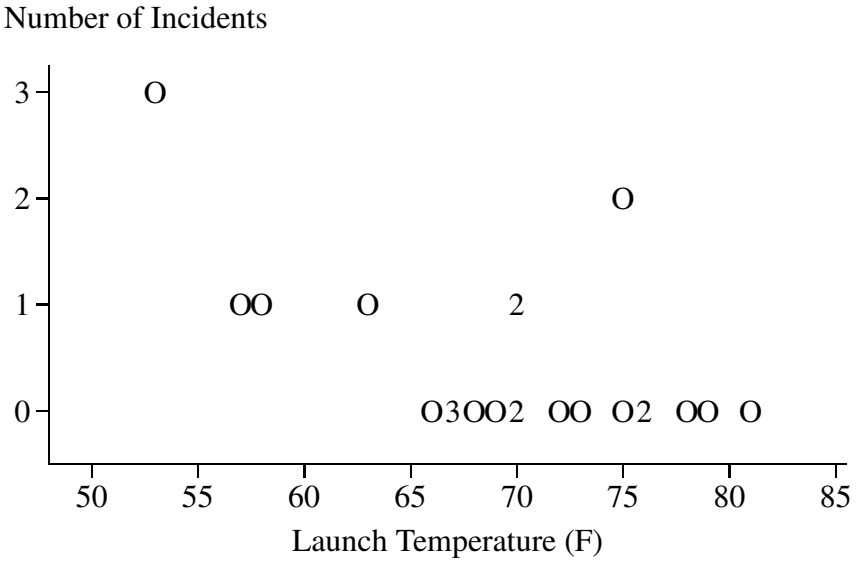


Figure 21.12: Scatterplot of the number of incidents of thermal distress to field joint O-rings versus launch temperature for 23 space shuttle flights before the Challenger accident.



21.7 Solutions to Practice Problems

1. (a) The slope and intercept of the regression line are:

$$b_1 = r(s_2/s_1) = 0.353(5.109/5.961) = 0.3025 \text{ and}$$

$$b_0 = \bar{y} - b_1\bar{x} = 92.35 - 0.3025(52.31) = 76.526.$$

Thus, the regression line is:

$$\hat{y} = 76.526 + 0.3025x.$$

(b) For $x = 46.0$, I get

$$\hat{y} = 76.526 + 0.3025(46.0) = 90.444 \text{ the old } \hat{y} \text{ is } 89.19.$$

For $x = 50.0$, I get

$$\hat{y} = 76.526 + 0.3025(50.0) = 91.654; \text{ the old } \hat{y} \text{ is } 91.00.$$

For $x = 55.0$, I get

$$\hat{y} = 76.526 + 0.3025(55.0) = 93.166; \text{ the old } \hat{y} \text{ is } 93.26.$$

For $x = 59.0$, I get

$$\hat{y} = 76.526 + 0.3025(59.0) = 94.376; \text{ the old } \hat{y} \text{ is } 95.06.$$

2. The relationship is clearly increasing and linear; thus, we may eliminate $r = -0.83$, $r = -0.36$ and $r = 0.00$. The relationship is weaker than our prototype for $r = 0.80$; thus, we may eliminate $r = 0.83$. By process of elimination, $r = 0.36$.

3. The relationship is clearly increasing and linear; thus, we may eliminate $r = 0.00$. The relationship is not perfect; thus, we may eliminate $r = 1.00$. The relationship is stronger than our prototype for $r = 0.40$; thus, we may eliminate $r = 0.052$ and $r = 0.302$. By process of elimination, $r = 0.724$.

4. (a) The relationship is definitely strong and decreasing. In my opinion, with a count response that takes on very few values, the idea of linear is somewhat meaningless. Excepting the three cases with $x > 65$ and $y \geq 1$, the pattern for the remaining 20 cases is **perfectly monotonic and deterministic**: for $x = 53$, $y = 3$; for $57 \leq x \leq 63$, $y = 1$; and for $x \geq 66$, $y = 0$. Looking at these 20 cases, it's hard to imagine that anyone would think that temperature doesn't matter!

   I conjecture that the case with $x = 75$ and $y = 2$ had a big role in the disaster. To me, this suggests that some O-rings had defects. It amazes me (not in a good way) that the observation that a warm temperature cannot *fix* a bad O-ring, could ever lead anyone to ignore the evidence a that cold temperature could damage a good O-ring!

   (b) As with my example of fish activity and water temperature, I don't see a data-based reason to worry about a launch at 31 degrees, simply because there are no data for any temperature remotely close to 31 degrees. I have tried throughout these notes to encourage you to use Statistics to *supplement* your scientific knowledge. As I understand it, simple physics suggests rather strongly that as temperature falls, the O-rings predictably will perform worse. (Sadly, my understanding of physics falls between none and simple.) Limitations of statistical methods should never be used as an excuse for ignoring scientific knowledge!

   (c) With only the seven cases with $y \geq 1$ and **no knowledge of physics**, I cannot argue with the decision. I am amazed (again not in a good way), however, that anyone would ever think that the cases with $y = 0$ were irrelevant.

5. (a) Every team played 82 games. Every game has a winner and a loser. Hence, the mean number of victories is $82/2 = 41$ each year.

   (b) First, the slope is

$$r(s_2/s_1) = 0.724(13.08/12.96) = 0.7307.$$

   Thus, the equation of the regression line is:

$$\hat{y} = 41 + 0.7307(x - 41).$$

   Here is my description. Every team plays 82 games. Thus, winning more than 41 games in 1990–91 is an above average performance. Thus, $(x - 41)$, if positive, measures how

many games above average a team was in 1990–91. The regression line predicts that only 73% (more precisely, 73.07%) of the *skill* exhibited in x is inherited by y. Thus, for example, a team that was 10 games better than average in 1990–91 is predicted to be only 7.3 games better than average in 1991–92.

Similarly, if $(x-41)$ is negative, we predict the team will win more games in 1991–92. Lest you *go all Galton* on me and predict that eventually every team will win one-half of its games, note that s_2 is slightly larger than s_1.

(c) First, Seattle is easy; because its x equals $\bar{x}$, its $\hat{y} = \bar{y} = 41$. The Law of the Preservation of Mediocrity at work!

$$\text{For Chicago: } \hat{y} = 41 + 0.7307(61 - 41) = 41 + 14.6 = 55.6.$$

$$\text{For Miami: } \hat{y} = 41 + 0.7307(24 - 41) = 41 - 12.4 = 28.6.$$

(d) The residuals are $e = y - \hat{y} = 47 - 41 = 6$ for Seattle; $e = 67 - 55.6 = 11.4$ for Chicago; and $e = 38 - 28.6 = 9.4$ for Miami.

21.8 Appendix: Optional Material

21.8.1 Properties 3 and 5 of the Correlation Coefficient

I will begin with property 3, which I will state again for convenience:

3. The value of the correlation coefficient is always between -1 and $+1$. It equals $+1$ if, and only if, all data points lie on a straight line with positive slope; it equals -1 if, and only if, all data points lie on a straight line with negative slope.

In particular, let's suppose that for all cases,

$$y_i = a + bx_i, \text{ with } b > 0.$$

In words, all of the data lie exactly on a straight line with positive slope. We need the following useful facts about means and standard deviations, which I could have proven in Chapter 1, but they are not much fun to prove; thus, it is optional for you to view it.

$$\bar{y} = a + b\bar{x} \text{ and } s_2 = bs_1.$$

The proof, described below, is not intellectually challenging, but is a bit messy algebraically.

First, I note that

$$\sum y_i = \sum(a + bx_i) = na + b\sum x_i.$$

Dividing both sides by n, we obtain

$$\bar{y} = a + b\bar{x},$$

the first of our results. Next, we look at the deviations for the y's:

$$y_i - \bar{y} = a + bx_i - a - b\bar{x} = b(x_i - \bar{x}).$$

Thus,

$$\sum (y_i - \bar{y})^2 = \sum [b(x_i - \bar{x})]^2 = b^2 \sum (x_i - \bar{x})^2.$$

Thus, dividing both sides by the degrees of freedom, $(n-1)$, we find

$$s_2^2 = b^2 s_1^2.$$

Taking square roots of both sides we get the desired result.

Now, we are ready to revisit the definition of the correlation coefficient for the current situation:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_1 s_2} = \frac{\sum (x_i - \bar{x})(a + bx_i - a - b\bar{x})}{(n-1)s_1 b s_1} = \frac{\sum b(x_i - \bar{x})(x_i - \bar{x})}{b(n-1)s_1 s_1} = \frac{s_1^2}{s_1 s_1} = 1.$$

If, instead, let's suppose that for all cases,

$$y_i = a + bx_i, \text{ with } b < 0.$$

The above proof works with the following change. The relationship between standard deviations becomes

$$s_2 = |b| s_1 = -b s_1 \text{ because } b < 0.$$

This changes the denominator of my earlier proof. The numerator is unaffected and eventually, after much rewriting and canceling, we have $r = b/(-b) = -1$.

Although I will not provide details, a slight modification of the algebra above will prove property 5 of the correlation coefficient in Result 21.1.

21.8.2 The Principle of Least Squares

My goal is to find the **best line** for describing my exam score data. Stating the obvious, there are two ways for us to go about this task:

- We can specify a way to measure how **good** each possible line is. Which ever line has the most goodness, is the best line.
- We can specify a way to measure how **bad** each possible line is. Which ever line has the least badness, is the best line.

It turns out that it is more fruitful to measure how bad a line is and then find the line with the least amount of badness. The particular method we use to do this is obtained by adopting the **Principle of Least Squares**. Take a moment and look at this name; in particular, note the word **principle**. This word reminds us that in the work below we are making a **value judgment**; with a different value judgment, an analyst would likely find a different best line.

Using the Principle of Least Squares to find the best line for a set of bivariate data is a big task. I will ease you into it in the following subsection.

21.8.3 The Principle of Least Squares for One Variable

Suppose that we have $n = 5$ numbers that, after sorting, are:

$$0, 1, 2, 7, 10.$$

In Chapter 1 you learned two ways to summarize these numbers: by their mean $\bar{x} = 4$ or by their median $\tilde{x} = 2$. We are now going to spend a few minutes looking at these five numbers and their two summaries in the context of this section.

Let me pose the following question:

> Which number c is best at describing these five numbers?

Of course, this question is meaningless until we decide how to measure badness. To be concrete, let me start by guessing $c = 4$. I want to know how badly $c = 4$ does, overall, at describing my five numbers. I decide that when I describe a number x by $c = 4$, I incur a **loss** of magnitude:

$$|x - c| = |x - 4|.$$

This loss (function) is called **absolute error loss**. It tells us, for example, that when I describe $x = 1$ by $c = 4$, I incur a loss of $|x - c| = |1 - 4| = 3$; when I describe $x = 10$ by $c = 4$, I incur a loss of $|x - c| = |10 - 4| = 6$; and so on. I measure the overall badness of $c = 4$ by summing the errors over all five data points:

$$|0 - 4| + |1 - 4| + |2 - 4| + |7 - 4| + |10 - 4| = 4 + 3 + 2 + 3 + 6 = 18.$$

For comparison, the overall badness of $c = 2$ is:

$$|0 - 2| + |1 - 2| + |2 - 2| + |7 - 2| + |10 - 2| = 2 + 1 + 0 + 5 + 8 = 16.$$

We see that for these data, the median, 2, is better than the mean, 4, at describing these data. Well, more precisely, the median is better than the mean if we use absolute error to measure badness.

It can be shown that for any set of data, when we use absolute error to measure badness, then the best descriptor of the numbers in the data set is the median. For an odd sample size, the median is the unique best descriptor; for an even sample size, there can be an interval of best descriptors. The interested reader may prove this fact.

I could argue that absolute error is the **natural** choice for measuring loss, but I don't want to be so restrictive; after all, many great discoveries in science were considered *unnatural*, at least at first.

There is another popular way to measure loss; it is called squared error. As the name suggests, when the data point x is described by the number c, then the loss incurred is

$$(x - c)^2.$$

For this choice of loss the overall badness in a set of data is

$$\sum (x_i - c)^2.$$

Our goal is to find the number c that minimizes this overall badness. I will show you two ways to obtain the value of c, one way uses algebra and the other calculus. Obviously, if you have never studied calculus, you are free to ignore my calculus argument and it won't affect your performance on the final.

First, I will do algebra. We rewrite our overall badness as follows:

$$\sum(x_i - c)^2 = \sum(x_i - \bar{x} + \bar{x} - c)^2 = \sum[(x_i - \bar{x}) + (\bar{x} - c)]^2 =$$

$$\sum(x_i - \bar{x})^2 + 2\sum(x_i - \bar{x})(\bar{x} - c) + \sum(\bar{x} - c)^2 = d_1 + d_2 + d_3, \text{ respectively.}$$

Let's look at these three pieces separately. Remember: Our goal is to determine the value of c that minimizes $d_1 + d_2 + d_3$. First, we can ignore d_1 because it is unaffected by the value of c. Next,

$$d_2 = +2\sum(x_i - \bar{x})(\bar{x} - c) = 2(\bar{x} - c)\sum(x_i - \bar{x}) = 2(\bar{x} - c) \times 0 = 0;$$

first, because the term $(\bar{x} - c)$ can be factored outside the summation because its value does not depend on i and, second, because the sum of the deviations always equals zero. Thus, d_1 is unaffected by c and $d_2 = 0$ regardless of the value of c. Thus, we minimize the overall badness by minimizing d_3. Obviously, $d_3 \geq 0$ is minimized when by $c = \bar{x}$ which makes it equal to 0.

The calculus argument is much easier. Define the function

$$f(c) = \sum(x_i - c)^2.$$

Take the derivative of f with respect to c and set it equal to 0:

$$-2\sum(x_i - c) = 0.$$

Solving this equation for c gives $c = \bar{x}$. The second derivative of f is the constant 2, which means that $c = \bar{x}$ minimizes f. (We don't really need to find the second derivative; it is obvious that the function f has no maximum.)

To summarize, for a single set of data, the median is the best descriptor for absolute error and the mean is the best descriptor for squared error.

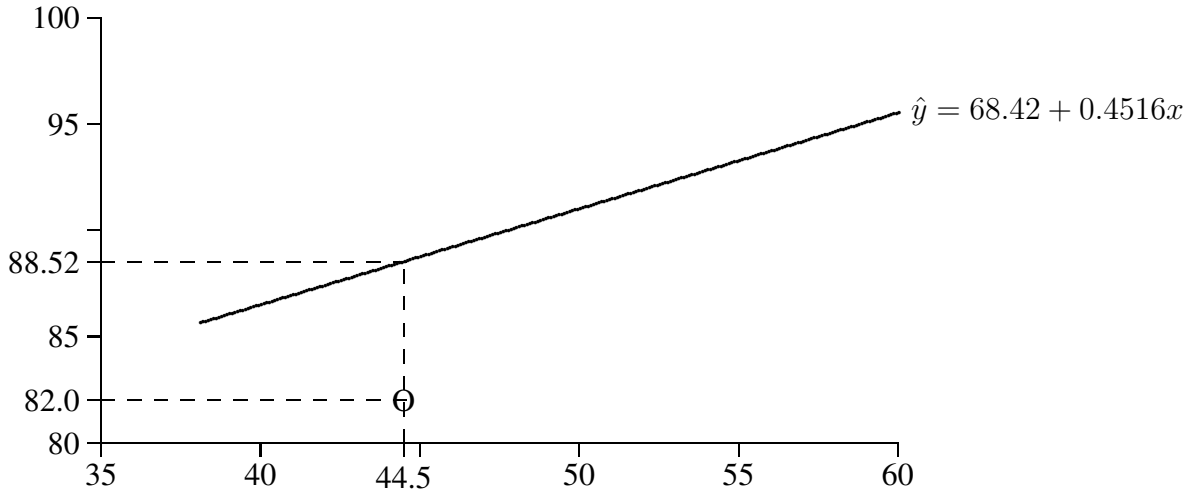
21.8.4 Back to Finding the Best Line

I want to generalize the above ideas relating the mean and median to two different loss functions to the problem of finding the best line for describing a scatterplot.

Our first principle is that if a particular case in a scatterplot lies exactly on a line, then the line's description of that point is perfect and no loss is incurred. (This is analogous to the idea that if a particular x_i is equal to the describer c, then with either absolute error or squared error loss, no loss is incurred.) The obvious question is: How do we measure the badness of the line for a point that does not lie on it?

To this end, please look at Figure 21.13. This figure shows just one case from our data set on exam scores, namely the student who scored $x = 44.5$ and $y = 82.0$. This figure also shows the regression line from Figure 21.6 and various dashed lines.

Figure 21.13: The regression line for the final exam score versus midterm exam score for $n = 35$ students, with the case $x = 44.5$ and $y = 82.0$.



Look at the graph of the regression line above the value $x = 44.5$. The height of the line above $x = 44.5$ is the value of

$$\hat{y} = 68.42 + 0.4516(44.5) = 88.52,$$

because this is what the graph presents: all pairs $(x, \hat{y})$ that satisfy the equation of the regression line. By contrast, the height of the circle above $x = 44.5$ is that student's actual final exam score, $y = 82.0$. In words, the actual y does not agree very well with the predicted y. We measure this lack of agreement by calculating the residual, $e = (y - \hat{y})$ which for this case is $(82.0 - 88.52) = -6.52$. The Principle of Least Squares tells us to measure the badness in this value by squaring it:

$$(-6.52)^2 = 42.5104.$$

Thus, the squared residual, usually called squared error, *suffered* (statisticians like to be a bit dramatic, at times) by using the regression line to predict the final for the student we have been considering is equal to 42.5104.

It would be incredibly tedious, but we could repeat the above argument to the other 34 cases in the data set. This would give us 35 squared errors. The Principle of Least Squares says to sum these 35 squared errors. We call this sum of squared errors, rather noncreatively, **the sum of squared errors** for the regression line. With the help of Minitab, I find that for the regression line, these 35 squared errors sum to 708.9464.

Why am I so enamored of our regression line? Answer: For the following mathematical fact that I will prove shortly:

Suppose that I use any line other than

$$\hat{y} = 68.42 + 0.4516x.$$

For this *new line*, for every case I calculate the predicted value of y, call it $\dot{y}$. Then I calculate the squared error of the difference between the actual and predicted response: $(y - \dot{y})^2$. I sum all of these squared errors. I **will obtain** a total of squared errors that is **larger than 708.9464**. In words, the regression line is the best line because it minimizes the sum of the squared errors; i.e., using the Principle of Least Squares, it is the winner!

Using calculus, the proof of this fact is quite simple. I will briefly give the details below.

Define the function f with two arguments by:

$$f(a_0, a_1) = \sum (y_i - a_0 - a_1 x_i)^2.$$

Take two *partial derivatives* of f, one with respect to a_0 and one with respect to a_1. Solve for the values b_0 and b_1 of a_0 and a_1, respectively, that make both equations equal 0. The resultant equations are:

$$\sum (y_i - b_0 - b_1 x_i) = 0 \text{ and } \sum x_i (y_i - b_0 - b_1 x_i) = 0. \tag{21.8}$$

Look at the first of these equations. Rewriting it, we obtain

$$\sum y_i = n b_0 + b_1 \sum x_i.$$

Divide both sides by n to obtain

$$\bar{y} = b_0 + b_1 \bar{x} \text{ or } b_0 = \bar{y} - b_1 \bar{x};$$

our familiar result in Equation 21.4.

This turns our second equation into

$$\sum x_i (y_i - \bar{y} + b_1 \bar{x} - b_1 x_i) = 0 \text{ or}$$

$$\sum x_i (y_i - \bar{y}) = b_1 \sum x_i (x_i - \bar{x}). \tag{21.9}$$

Note that

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i (y_i - \bar{y}) - \bar{x} \sum (y_i - \bar{y}) = \sum x_i (y_i - \bar{y}),$$

because deviations sum to zero. Similarly,

$$\sum (x_i - \bar{x})(x_i - \bar{x}) = \sum x_i (x_i - \bar{x}) - \bar{x} \sum (x_i - \bar{x}) = \sum x_i (x_i - \bar{x}),$$

Thus, Equation 21.9 can be written as

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = b_1 (n-1) s_1^2.$$

Divide both sides by $(n-1)s_1 s_2$ and obtain:

$$r = b_1 (s_1 / s_2) \text{ or, our familiar } b_1 = r(s_2 / s_1).$$

Thus, the regression line given in these *Course Notes* is, indeed, the best line according to the Principle of Least Squares.

If you are still reading—no mean task given the messiness of the above algebra—you might wonder: What line do we get if we use the principle of absolute error? After all, for univariate data, the Principle of Least Squares gives us the mean as the best descriptor and the principle of minimizing total absolute error gave any median as the best descriptor.

The answer? Sadly, there is no closed-form representation of the best line unless we square errors. This is not interpreted, however, as a major disappointment because, as you will see in Chapter 22, the best line according to the Principle of Least Squares is very useful in science.

Bibliography

[1] Carrel, J.E. and Heathcoat, R. D., "Heart Rate in Spiders: Influence of Body Size and Foraging Energetics," *Science*, July 9, 1976, pp 148–150.

[2] Reichler, J.L., Ed., *The Baseball Encyclopedia*, MacMillan Publishing Company, New York, 1988.

[3] Galton, Francis, "Regression Towards Mediocrity in Hereditary Stature," *The Journal of the Anthropological Institute of Great Britain and Ireland*, Vol. 15(1886), pp. 246-263.

[4] Gould, S.J., *The Mismeasure of Man*, Norton, New York, 1981.

[5] Witmer, J. A., *Data Analysis An Introduction*, Prentice Hall, Englewood Cliffs, New Jersey, 1992, pages 46–48.

Chapter 22

Simple Linear Regression: Continued

Chapter 21 presented quite a few methods for describing bivariate numerical data. Section 1 below presents a few more descriptive methods. (This material should have been in Chapter 21, but I decided that chapter was already too long!) Beginning with Section 2, this chapter will present methods of inference.

22.1 Is the Regression Line Any Good?

George Edward Pelham Box (1919–2013) was a great statistician and founded the Department of Statistics at the University of Wisconsin–Madison in 1960. He was my friend and taught me a great deal about what's important in Statistics. Being both famous and brilliant, George was often asked for pithy statements; one of my favorites of his is:

> Just because something is optimal, it doesn't mean it's any good!

George would especially direct this remark at mathematical statisticians who were solving problems that were not useful to scientists.

In the present context, we have found that the regression line is the best line for describing our data; but is it any good? I will explore this question in detail.

As stated in the previous chapter, each case enters the data set with two numbers: x_i and y_i, its values of the predictor and the response. After the regression line is determined, the case has two additional numbers associated with it: its predicted response $\hat{y}_i$ and its residual $e_i = y_i - \hat{y}_i$.

I want to examine the residuals. Each case has a residual; thus, there are n residuals:

$$e_1, e_2, e_3, \ldots, e_n.$$

With this one set of numbers, we are back in the realm of Chapters 1 and 2. We could draw a picture of them: a dot plot, a histogram or, perhaps, a kernel density histogram. Also, we could calculate their mean and their standard deviation.

To make this easier to follow, I will focus on two data sets from Chapter 21:

- The data on midterm and final exam scores for $n = 35$ students; and

Figure 22.1: Frequency histogram of the $n = 35$ residuals for the regression of final exam score on midterm exam score.

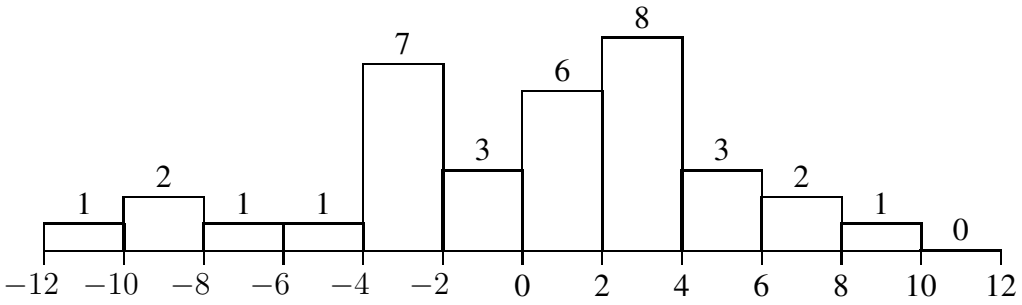
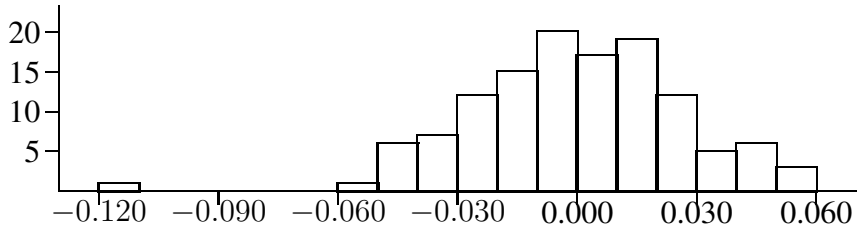


Figure 22.2: Frequency Histogram of the Residuals for the Batting Averages Data.



- The batting averages data for $n = 124$ American League baseball players.

Figure 22.1 presents a frequency histogram of the residuals for the exam scores and Figure 22.2 presents a frequency histogram of the residuals for the batting averages data. Let me make a few comments about these histograms.

1. For the exam scores' residuals, there are no outliers—the dot plot, not shown here, also revealed no outliers. Otherwise, the shape of the histogram is not recognizable to me. If I combine adjacent class intervals and have six class intervals instead of 12 (12 is a large number of intervals for $n = 35$) the histogram becomes a bit smoother.

2. For the batting averages data, I note one small outlier, a residual equal to approximately -0.120. Do you remember who this is?

   Answer: A negative residual means that the 'O' for the case is below the regression line. The fact that it's an outlier means that its distance below the regression line is much larger than any other case's distance below the regression line. It's Floyd Rayford!

   Excluding Mr. Rayford, the remainder of the histogram is approximately bell-shaped and symmetric.

3. Each histogram gives us a picture of the *sizes* of the residuals. For example:

(a) For the exam scores, 69% ($7 + 3 + 6 + 8 = 24$ of 35) of the residuals are between -4 and $+4$. Remembering what a residual measures, this means that by using the midterm score to predict the final score via the regression line, the predicted final is within four points of the actual final for 69% of the students. I view this as a *glass half full* statement. As a *glass half empty guy*, I prefer to say that for 31% of the students, the actual final deviates from the predicted final by more than four points.

(b) For the batting averages data, with the exception of Mr. Rayford, all predictions are within 60 points (remember, this is *baseball speak* for 0.060) of the actual value of the corresponding y. Although you cannot verify the following from Figure 22.2:

- 77% (95 of 124) residuals are between -30 and $+30$ points;
- 68% (84 of 124) residuals are between -25 and $+25$ points;
- 57% (71 of 124) residuals are between -20 and $+20$ points; and
- 30% (37 of 124) residuals are between -10 and $+10$ points;

Given that I **always** perform a regression analysis by using a computer software package, the counts illustrated in item 3 above are easy to obtain. They are very helpful for the researcher who is trying to decide whether the regression line is of value to the scientific problem being considered. Regarding the batting averages study, as a baseball fan I am disappointed with the residuals. It seems to me that predicting a batting average within 20 points is not particularly accurate and it is disappointing to learn that for 43% of the players the prediction fails to meet this modest goal. If you are a baseball fan, form your own opinion; you need not agree with me. Regarding the exams scores, I conjecture that as a student you have stronger feelings about exam scores than I do. Thus, I will leave it to you to decide whether predicting within four points for 69% of the students is a noteworthy achievement.

Let me make one more comment about the above; this is an issue that sometimes causes confusion. In Chapter 21, when I examined the scatterplot for the batting averages data, I labeled three players as being *isolated cases*. They are isolated in comparison to the other cases in the scatterplot. Now, we see that one case is an outlier. Thus, my first remark is to note that being isolated and being an outlier are two different notions. Thus, be careful about your use of these terms; sadly, many persons are careless and use them interchangeably, which leads to confusion. Below are some comments on how to keep these ideas separate.

1. We use the term *isolated* when considering two (or more) variables simultaneously; we use the term *outlier* for one variable. Note that we will **not** consider more than two variables simultaneously in these *Course Notes*.

2. A case is *isolated* if it is far away from other cases (with the possible exception that it could be close to another isolated case or cases). A case's residual is an *outlier* if it is unusually far away—in terms of vertical deviation—from the regression line. Often people are lazy and say *a case is an outlier* instead of the more accurate *a case's residual is an outlier*. Do not interpret *lazy* as a pejorative; if I am talking with a statistician—or if I simply forget—I will call a case an outlier.

3. If you think about the previous item, you will note that if a case's residual is an outlier, then it is an isolated case, but a case can be isolated without its residual being an outlier. For example, the residual for Wade Boggs [Don Mattingly] is 29 [52] points. They are far from other cases, but close to the line. Well, Boggs is reasonably close to the line; Mattingly had the second largest residual (third largest absolute residual) among the 124 players; thus, he is *almost* an outlier.

22.1.1 The Mean and Standard Deviation of the Residuals

It can be shown that for every set of data, there are two restrictions on the values of the residuals.

$$\sum e_i = 0 \text{ and } \sum x_i e_i = 0. \tag{22.1}$$

(If you read the optional Appendix in Chapter 21, the above is simply Equation 21.8.) From the first of these equations, we see that the mean of the residuals, $\bar{e}$, equals 0 for every set of data. In symbols,

$$\bar{e} = 0, \text{ for every set of data.}$$

Let's **not** rush on to the next topic; this is a very important equation. Similar to my feelings about the Law of the Preservation of Mediocrity in Chapter 21, I would not like the regression line very much if this equation were not true. Here is why.

The fact that $\sum e_i = 0$ means that the regression line passes through the *center* of the data in the sense that: some cases are above the line and some are below, but the sum of the distances above the line cancel exactly the sum of the distances below the line.

Now that we know the center—mean—of the distribution of the residuals, as in Chapter 1 we turn to the determination of the amount of spread. In Chapter 1, for each x_i in the data set, we compared it, via subtraction, to its mean in order to obtain its deviation: $(x_i - \bar{x})$. Residual e_i is its own deviation because $\bar{e} = 0$. Thus, the sum of squared deviations of the residuals is simply the sum of squared residuals, which we denote by SSE:

$$\text{SSE } = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - b_0 - b_1 x_i)^2. \tag{22.2}$$

Enrichment: Note that statisticians are a bit confused about their usage of the letters 'e' and 'r.' Why do I say this?

- We use r to denote the correlation coefficient and, hence, cannot use it for residuals. We use e for residuals, that, historically, were also called errors. Hence, the sum of the squared residuals is denoted SSE, where the E is for the word error. One reason statisticians replaced the name *error* with residual is we got tired of the following exchanges between a statistician (S) and a client (C):

  S: Let's look at a list of your errors.

  C: I didn't make any errors!

  or

  S: Here is a list of the errors from our regression analysis.

  C: Fix those errors, then get back to me!

Also, we use SSE for sum of squared residuals because SSR—see below—is reserved for the *sum of squares due to regression*.

Back in Chapter 1, I remarked that statisticians and mathematicians disagree on what to do with the sum of squared deviations: the mathematicians divide it by n and the statisticians divide it by the degrees of freedom, $(n-1)$. Recall, also, that there are $(n-1)$ degrees of freedom for the deviations because they are subject to one constraint: they must sum to zero. As stated above in Equation 22.1, the residuals have two constraints; hence, they have $(n-2)$ degrees of freedom. As a result, statisticians define the variance and standard deviation of the residuals as follows.

Definition 22.1 (The variance and standard deviation of the residuals.) *The variance of the residuals is*

$$s^2 = \frac{SSE}{n-2}, \tag{22.3}$$

where SSE is defined in Equation 22.2.

The standard deviation of the residuals is

$$s = \sqrt{s^2} = \sqrt{\frac{SSE}{n-2}}. \tag{22.4}$$

We now have three standard deviations associated with a regression analysis:

- The standard deviation of the values of x, denoted by s_1.
- The standard deviation of the values of y, denoted by s_2.
- The standard deviation of the residuals, denoted by s.

Note that it would *make sense* to denote the last of these by s_e with the subscript e to remind us that we are measuring the spread in the residuals. Statisticians don't do this, for at least two reasons:

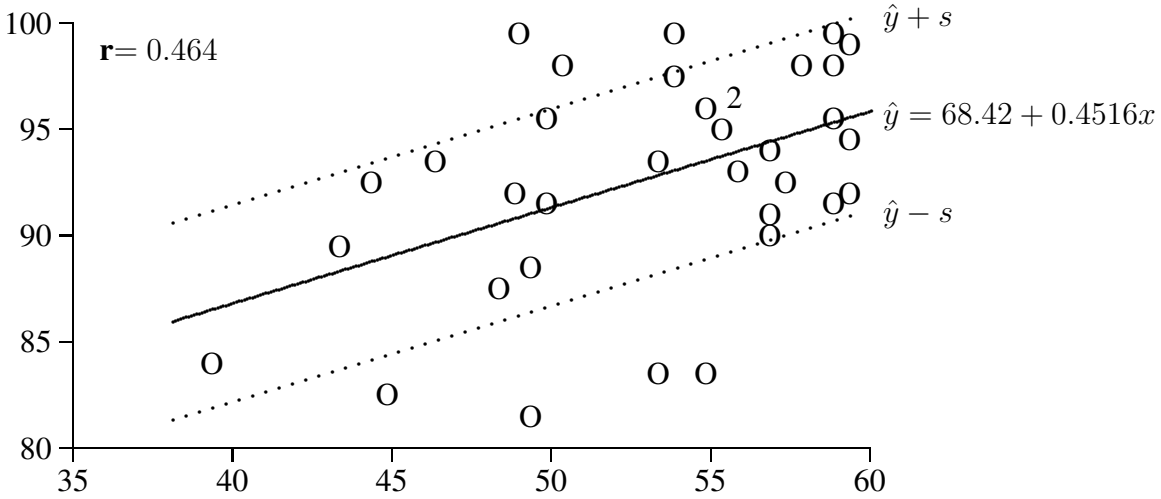
1. We are lazy and want to save the effort of typing a subscript whenever possible.
2. The lack of a subscript *honors* the standard deviation of the residuals as being the most important of the three standard deviations. This is analogous to how certain pop stars—and numerous Brazilian soccer players—have only one name—Cher, Prince, Pelé, etc. (I apologize for my pop culture references being so dated. Also, of course, Prince became the favorite of all math-ophiles when he replaced his name with a symbol!)

For the exam scores data, $s = 4.635$. Because $\bar{e} = 0$, the Empirical Rule from Chapter 2 (Result 2.2) tells us that approximately 68% of the residuals are between -4.635 and $+4.635$; i.e., approximately 68% of the predicted values are within 4.635 of the actual response.

The Empirical Rule is actually a pretty bad approximation for the exam scores data. As we saw above, 69% of the residuals are between -4 and $+4$. Here's another way to see that it is bad; by actual count—details not given— fully 80% (28 of 35) residuals are between -4.635 and $+4.635$. As noted, in Chapter 2, the Empirical Rule is unreliable for small values of n or for distributions that are not bell-shaped; and the exam scores' residuals suffer both of these maladies.

By the way, if you prefer pictures to counting, please refer to Figure 22.3. This figure presents the scatterplot of the 35 pairs of exams scores with three parallel lines superimposed:

Figure 22.3: Final Exam Score Versus Midterm Exam Score for 35 Students.



- The regression line $\hat{y} = 68.42 + 0.4516x$, drawn as a solid line.
- The line $\hat{y} + s$, drawn as a dotted line.
- The line $\hat{y} - s$, drawn as a dotted line.

A quick examination of this picture shows that four cases fall below the line $\hat{y} - s$ and three cases fall above the line $\hat{y} + s$; thus, the remaining 28 cases fall between the dotted lines. In words, 28 cases have a residual between the value $-s$ and $+s$, as I reported earlier.

For the batting averages data, $s = 0.0268$, just under 27 points. By actual count—details not given—slightly more than 72% (87 of 124) of the residuals are between -0.0268 and $+0.0268$. Here, n is pretty large and, excepting the outlier, the residuals have a bell-shaped distribution. The 72% is larger than the Empirical Rule's 68% because Mr. Rayford's residual inflates the value of s. In particular, if one deletes Mr. Rayford from the data set, and runs the regression program again, s becomes 0.0248, a reduction of 7.5% from the earlier $s = 0.0268$. Sadly, however, for the new data set with $n = 123$, only 65% (80 of 123) of the residuals are between -0.0248 and $+0.0248$; approximations can be so annoying! To further confuse matters, if we round s to 25 points, then 67% (82 of 123) of the residuals fall between $-s$ and $+s$. The moral: The Empirical Rule is a useful guide, but it's not exact; if you want exact, look at the list of residuals and count!

To summarize the above, for the exam scores and batting average data sets, we look at the value of s and, using the Empirical Rule, can make a subjective assessment as to whether the predictions from the regression line are scientifically useful.

Another popular approach is to measure how well the regression line predictions—which are obtained from the best line for using X to predict Y—compare to predictions that ignore X.

Well, if we ignore X, then our data set becomes a collection of n values of Y. Based on the Principle of Least Squares the best predictor of each y_i using only the values of

$$y_1, y_2, y_3, \ldots, y_n,$$

is $\bar{y}$. (Allow me to casually call $\bar{y}$ the best predictor. As I show in the optional Appendix to Chapter 21, for data on one variable, according to the Principle of Least Squares the mean is the best predictor/describer of the data. If you did not read this material, that is fine, but you will need to take my use of $\bar{y}$ on faith.)

If I predict y_i by $\bar{y}$, the difference is $(y_i - \bar{y})$. Squaring these differences and summing them, we get

$$\text{SSTO} = \sum (y_i - \bar{y})^2, \tag{22.5}$$

where the symbol SSTO is called the *total sum of squares*. (In very old textbooks, our SSTO is called the *adjusted* total sum of squares.) From Chapter 2 we know that

$$s_2^2 = \text{SSTO}/(n-1) \text{ and } s_2 = \sqrt{\text{SSTO}/(n-1)}.$$

Recall that for the exam scores data, $s_2 = 5.154$ and for the batting averages data, $s_2 = 0.0320$. Also, as discussed above, for the exam scores data, $s = 4.635$ and for the batting averages data, $s = 0.0268$. For the exam scores data:

$$s/s_2 = 4.635/5.154 = 0.899;$$

thus, using the midterm to predict the final reduces the standard deviations of the errors in the predictions by 10.1%. For the batting averages data:

$$s/s_2 = 0.0268/0.0320 = 0.838;$$

thus, using the 1985 batting average to predict the 1986 batting average reduces the standard deviations of the errors in the predictions by 16.2%.

In my opinion, comparing s to s_2, as I have done above, is a good way to measure the relative usefulness of using X versus not using X. I need to mention, however, that this is not the only comparison that scientists make. Indeed, I must admit that my personal evidence is overwhelming that the comparison below is more popular than my favored comparison of s versus s_2. First, however, we must make a side trip into the next subsection.

22.1.2 The Analysis of Variance Table

In Table 21.3 in Chapter 21 I presented *edited* Minitab output for the regression analysis of the exam scores data for $n = 35$ students. This table, with an additional column added—SE(Fit)—is reproduced in Table 22.1 because we will need it often later in this chapter. In both of these tables I *deleted* the **Analysis of Variance Table** given in the Minitab output. By the way, we abbreviate the *Analysis of Variance Table* as the ANOVA table. We could get through this chapter without

Table 22.1: Edited Minitab output for the regression of final exam score on midterm exam score for 35 students.

The regression equation is: Final = 68.4 + 0.452 Midterm

Predictor	Coef	SE(Coef)	T	P
Constant	68.420	7.963	8.59	0.000
Midterm	0.4516	0.1501	3.01	0.005

S = 4.635 R-Sq = 21.5%

Obs	Midterm	Final	Fit	SE(Fit)	Residual
1	39.0	83.5	86.032	2.213	-2.532
2	43.0	89.0	87.838	1.665	1.162
3	44.0	92.0	88.290	1.534	3.710
4	44.5	82.0	88.515	1.470	-6.515
5	46.0	93.0	89.193	1.285	3.807
6	48.0	87.0	90.096	1.063	-3.096
7	48.5	91.5	90.322	1.014	1.178
8	49.0	99.5	90.548	0.968	8.952
9	49.0	88.0	90.548	0.968	-2.548
10	49.0	81.0	90.548	0.968	-9.548
11	49.5	91.0	90.773	0.926	0.227
12	49.5	95.0	90.773	0.926	4.227
13	50.0	97.5	90.999	0.888	6.501
14	53.0	83.0	92.354	0.784	-9.354
15	53.0	93.0	92.354	0.784	0.646
16	53.5	97.0	92.580	0.791	4.420
17	53.5	99.0	92.580	0.791	6.420
18	54.5	95.5	93.031	0.825	2.469
19	54.5	83.0	93.031	0.825	-10.031
20	55.0	94.5	93.257	0.851	1.243
21,22	55.5	96.0	93.483	0.883	2.517
23	55.5	92.5	93.483	0.883	-0.983
24	56.5	93.5	93.934	0.962	-0.434
25	56.5	89.5	93.934	0.962	-4.434
26	56.5	90.5	93.934	0.962	-3.434
27	57.0	92.0	94.160	1.007	-2.160
28	57.5	97.5	94.386	1.056	3.114
29	58.5	99.0	94.838	1.162	4.162
30	58.5	95.0	94.838	1.162	0.162
31	58.5	91.0	94.838	1.162	-3.838
32	58.5	97.5	94.838	1.162	2.662
33	59.0	91.5	95.063	1.218	-3.563
34	59.0	98.5	95.063	1.218	3.437
35	59.0	94.0	95.063	1.218	-1.063

Table 22.2: Analysis of Variance table from the Minitab analysis of the regression of final exam score on midterm exam score for 35 students.

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	194.38	194.38	9.05	0.005
Residual Error	33	708.81	21.48		
Total	34	903.19			

mentioning ANOVA tables, but if you do any statistical analyses beyond this chapter, you may well run into them.

Table 22.2 is the ANOVA table from Minitab's regression analysis of the exam scores data. All statistical software packages I have seen give a similar ANOVA table for regression. Let me take a few minutes to explain the connection between this table and our current work.

First, ignore the last three columns—those headed *MS*, *F* and *P*. The *DF* column presents degrees of freedom and the *SS* column presents various sum of squares, both identified with the *feature* in the *Source* column. If you recall that $n = 35$, you see that the degrees of freedom for the total sum of squares (SSTO) is indeed $(n - 1) = (35 - 1) = 34$, as Minitab states. Also, the degrees of freedom for the error sum of squares, SSE, is $(n - 2) = (35 - 2) = 33$, as Minitab states. This table leads to some obvious questions:

1. What is the *regression sum of squares*? Why does it have one degree of freedom?

2. Why do the first two sums of squares sum to the third? ($194.38 + 708.81 = 903.19$.)

To answer these questions, I must go back to the beginning. We start with an arbitrary case 'i' in our data set and look at the deviation of its response y_i from the mean response, $\bar{y}$:

$$(y_i - \bar{y}).$$

We rewrite this as

$$(y_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}). \tag{22.6}$$

In words, this equation states:

> The deviation of a response from its mean is the sum of two terms: The deviation of the response from its predicted value **and** the predicted value minus the (overall) mean response.

This equation is so important that I want to illustrate it with two cases from our exam scores data:

- Consider the case with $x = 57.5$ and $y = 97.5$. You may verify that for this case, $\hat{y} = 94.386$. (You may verify this by plugging $x = 57.5$ into the equation of the regression line, or, more

easily, locate this case as observation 28 in Table 22.1.) Also, recall that $\bar{y} = 92.257$. For this case, Equation 22.6 becomes:

$$(97.5 - 92.257) = (97.5 - 94.386) + (94.386 - 92.257) \text{ or}$$

$$5.243 = 3.114 + 2.129 = 5.243.$$

- Consider the case with $x = 56.5$ and $y = 89.5$. You may verify that for this case, $\hat{y} = 93.934$. (This case is observation 25 in Table 22.1.) For this case, Equation 22.6 becomes:

$$(89.5 - 92.257) = (89.5 - 93.934) + (93.934 - 92.257) \text{ or}$$

$$-2.757 = -4.434 + 1.677 = -2.757.$$

If you take Equation 22.6 and sum both sides over all values of i, then obviously the equality is preserved:

$$\sum(y_i - \bar{y}) = \sum(y_i - \hat{y}_i) + \sum(\hat{y}_i - \bar{y}).$$

If we square the three terms in Equation 22.6 we on longer get equality; in my cases above:

$$(5.243)^2 = 27.489 \text{ does not equal } (3.114)^2 + (2.129)^2 = 14.230 \text{ and}$$

$$(-2.757)^2 = 7.601 \text{ does not equal } (-4.434)^2 + (1.677)^2 = 22.473.$$

If, however, we square the three terms and then sum over all cases, the equality is preserved:

$$\sum(y_i - \bar{y})^2 = \sum(y_i - \hat{y}_i)^2 + \sum(\hat{y}_i - \bar{y})^2. \tag{22.7}$$

This seems to be a magical result, but it is simply the n dimensional version of the Pythagorean Theorem. You should recognize two of the terms in Equation 22.7; it becomes:

$$\text{SSTO} = \text{SSE} + \sum(\hat{y}_i - \bar{y})^2.$$

I name this last term the sum of squares due to regression and write it as SSR. Thus, we see that the identity

$$\text{SSTO} = \text{SSE} + \text{SSR}$$

in Table 22.2 is not an accident; this equation will be true for every regression analysis.

By the way, I will ask you to take it on faith that the degrees of freedom for SSR is 1. It's easy to remember: both sum of squares and degrees of freedom sum in an ANOVA table.

For any regression analysis, all three of our sums of squares must be nonnegative numbers. In addition:

- SSTO is positive, because $s_2 > 0$.
- SSE is zero if, and only if, all points lie on a straight line which happens if, and only if $r = \pm 1$.

- SSR is zero if, and only if, the regression line has slope equal to zero which happens if, and only if, $r = 0$.

Thus, for example,

$$\text{SSE} \leq \text{SSTO}\,.$$

We can also see this result by applying *logic*, as follows. SSE minimizes the sum of squared errors around all lines; thus, it cannot exceed SSTO, which is the sum of squared errors around a *particular line*, namely, the horizontal line with intercept equal to $\bar{y}$.

The above considerations lead to the following definition.

Definition 22.2 (The coefficient of determination, R^2.) *The coefficient of determination is denoted by R^2 and is given by:*

$$R^2 = (\frac{\textit{SSTO} - \textit{SSE}}{\textit{SSTO}}) \textit{ which also equals } (\frac{\textit{SSR}}{\textit{SSTO}}) \textit{ or } (1 - \frac{\textit{SSE}}{\textit{SSTO}}). \tag{22.8}$$

Let's look at the first expression for R^2. The numerator is the total squared error when ignoring X (SSTO) minus the the total squared error when using X via the best possible line (SSE). Various picturesque ways to describe this difference include:

- SSTO minus SSE measures the amount of squared error in the y's that can be _____

  (Choose one): explained, removed, accounted for, explained, ...

  by a linear relationship with the x's.

Thus, for example if you have regression data and obtain SSTO $= 100$ and SSE $= 20$, then 80 of the 100 total squared error is explained (my choice of verb) by a linear relationship between y and x. But 80 what? How do we interpret this number? **Answer:** Look at the denominator of R^2. We compare, by dividing, the *squared error removed* with the original amount of squared error. Thus, for my current fictional numerical example,

$$R^2 = \frac{\text{SSTO} - \text{SSE}}{\text{SSTO}} = \frac{100 - 20}{100} = \frac{80}{100} = 0.80.$$

Usually, R^2 is reported as a percentage, for a reason you will see in a moment. Thus, instead of saying $R^2 = 0.80$ it is standard to say $R^2 = 80\%$. Returning to my picturesque statement for this value, we get:

> Eighty percent of the squared error in the y's can be explained by x.

Let's return to real data. Using the sums of squares in Table 22.2 we obtain

$$R^2 = \text{SSR}/\text{SSTO} = 194.38/903.19 = 0.215.$$

Recall that for the exam scores data, the correlation coefficient equals 0.464. Thus, $r^2 = (0.464)^2 = 0.215$, rounded to three digits. Is this agreement between r^2 and R^2 an accident? No.

Result 22.1 *For every regression analysis,*

$$r^2 = R^2. \tag{22.9}$$

Be careful with this result. Its main benefit is that it gives us another interpretation of the correlation coefficient r; namely, the square of correlation coefficient has the same interpretation as R^2 in terms of *explaining* squared errors. This helps us see the validity of property four of the correlation coefficient given in Result 21.1. Among other things, this property said that $r = +0.60$ reflects the same strength of a linear relationship as $r = -0.60$; we can see that this is true in the sense they both give the same value, 36%, of R^2.

There is a bizarre misinterpretation of Result 22.1 and I must comment on it. Let me illustrate this with a fictional conversation between Researchers A and B.

> A: I published my regression analysis; now all the world knows that my r is 60%.
>
> B: You are a bad person!
>
> A: Huh?
>
> B: You deliberately deceive people. It would be more honest to report that R^2 equals 36%. You are trying to trick people into believing you have somehow accounted for 60% when, in fact, you have accounted for only 36%. Shame, shame on you!

If you haven't guessed my position, Researcher B is misguided. I disagree with Researcher B for two reasons:

1. There is nothing *natural* about squaring errors. Indeed, when we squared deviations in Chapters 1 and 2 to obtain the variance, we quickly found that to get a summary that *has meaning* we need to take the square root to obtain the standard deviation. Thus, arguably, R^2 is not a natural measure.

2. When we discussed the regression effect in Chapter 21, we found that Equation 21.6 justifies referring to $r = 0.60$ as 60%; because 60% of the advantage in x is inherited by $\hat{y}$.

Don't get me wrong; I think that R^2 is an interesting summary of a regression, but it is not the whole story. If you decide that you prefer R^2 to r, that is fine; just don't use a questionable argument to bully people!

One final comment on this section. I have shown you two ways to decide whether using X via the regression line is better than not using it. The first was to compare s to s_2; and the second was to compute R^2. You may have noticed that these ways are mathematically equivalent. I will spare you the algebra, but note that

$$(s/s_2)^2 = [(n-1)/(n-2)](1 - R^2).$$

For the exam scores data,

$$(s/s_2)^2 = (4.635/5.154)^2 = 0.815057 \text{ and}$$

$$[(n-1)/(n-2)](1 - R^2) = (34/33)(1 - (194.38/903.19)^2 = 0.808566,$$

which are the same, except for my round-off error. (Sorry.)

22.2 The Simple Linear Regression Model

We now turn to inference for simple linear regression. As you might imagine, inference will require us to make one of the following three assumptions:

- We have a smart random sample from a finite population;
- We have a dumb random sample from a finite population; or
- We assume that we have i.i.d. trials.

Indeed, our approach in Part II of these notes always has been to make one of these assumptions in order to perform inference. Recall also that:

- Having a dumb random sample yields i.i.d. trials;
- Provided the sample size, n, is 5% or fewer of the population size, N, probabilities for dumb sampling are a good approximation to probabilities for smart sampling; and
- Literal random samples are very rare in science. Usually, the best we can do is feel comfortable in making the WTP (Willing to Pretend) assumption of Chapter 10 (Definition 10.3) on page 240.

Before I talk about random sample assumptions, however, there is another feature of regression that I need to introduce. This feature is similar to the observational study versus experimental study dichotomy of Chapter 15.

In the three main studies of Chapter 21—spiders, baseball players and Statistics 371 students—each case entered the study with two numerical features that the researcher measured/determined. In the vernacular, both X and Y are random. While Y is always random in regression analysis, in many studies the n values of X are selected by the researcher and assigned by randomization to the n cases available for study. Let me give you an example:

Example 22.1 (A hypothetical study on crop yield.) *A researcher has n one-acre* plots *of land available for study. The goal is to investigate the effect of a particular fertilizer on the yield of a particular crop. Before conducting the study, the researcher selects n values of the* concentration of the fertilizer *for study. The concentration values need not be distinct. The n concentration values are assigned to plots by randomization. Let x_i denote the concentration assigned to plot number i and let y_i be the yield—say, in bushels or pounds—of the crop from the plot.*

To summarize, there are n cases (one-acre plots) and each case yields two numbers: the concentration of the fertilizer, x, and the crop yield, y. In other words, we have regression data *with Y random and X not random.*

I am now ready to show you the **simple linear regression model**. We assume the following relationship between X and Y:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \text{ for } i = 1, 2, 3, \ldots n. \tag{22.10}$$

Note the following features of this equation.

1. I use upper case letters—i.e., Y_i and X_i—to emphasize that we write down this model **before** we collect any data. This means, in particular, that we will consider **probabilities** for the values taken on by
$$Y_1, Y_2, Y_3, \ldots, Y_n.$$

2. Note that even though we use upper case letters for the values of the predictor, we **do not** consider **probabilities** for the values taken on by
$$X_1, X_2, X_3, \ldots, X_n,$$
for one of the following two reasons:

   (a) In the non-random X case introduced above, it makes no sense to calculate probabilities for the values of the predictor; they are deliberately (i.e., non-randomly) selected by the researcher.

   (b) In the random X case, our inference **conditions** on the n values of X in the data set.

   A thorough exploration of why statisticians and scientists condition on the values of X in the *random X situation* is beyond the scope of these notes. I will remark, however, that while the reasons include mathematical necessity—some of the formulas later in this chapter become invalid for random X without conditioning—they also include scientific usefulness. For example, in most scientific applications the main interest is on the behavior of the response, conditional on the value of the predictor. Suppose that you want to use height (X) to predict weight (Y) for a *particular randomly selected man of interest* then you want to condition on (i.e., use) *his height* to make your prediction; i.e., you want to condition on his being 76 inches tall or 66 inches tall rather than have an *unconditional prediction* over all possible heights!

   Let me note that the idea of conditioning is familiar to the reader of these notes. The Skeptic's Argument in Part I has the effect of conditioning on the response values actually obtained. Even statisticians who don't like randomization-based inference condition on the marginal totals in the 2×2 contingency table in order to perform Fisher's Test. I could list more examples of conditioning, but I prefer to *stick to our task*.

3. After collecting data, the researcher will have the n observed pairs of values of the predictor and response.

4. The values β_0 and β_1 are parameters of the model; this means, of course, that they are numbers and by changing either or both of their values we get a different model. The actual numerical values of β_0 and β_1 are known by Nature and unknown to the researcher; thus, the researcher will want to estimate both of these parameters and, perhaps, test hypotheses about them.

5. The ϵ_i's are random variables with the following properties. They are i.i.d. with mean 0 and variance σ^2. Thus, σ^2 is the third parameter of the model. Again, its value is known

by Nature but unknown to the researcher. It is very important to note that we assume these ϵ_i's, which are called **errors**, are statistically independent. In addition, we assume that every case's error has the same variance.

Oh, and by the way, not only is σ^2 unknown to the researcher, the researcher does not get to observe the ϵ_i's. Remember this: all that the researcher observes are the n pairs of values of (X, Y).

6. Our inference procedures below are based on the assumption that the simple linear regression model is **true** or **correct**. A scientist won't know whether the model is correct. Indeed, another of George Box's pithy statements is:

   All models are wrong, but some are useful.

   This reflects the belief that we should be careful about using the simple linear regression model, as well as more complicated models. If you study regression beyond this chapter, I hope that a good amount of time is spent on how to check whether the assumptions of the model are close enough to being correct for the model to be useful.

Now, we look at some consequences of our model. The results below follow quite easily from the rules of means and variances familiar to the undergraduate Statistics major. In these notes, I have made only vague references to these rules; thus, don't worry if the algebra below is confusing.

Remember, the Y_i's are random variables; the X_i's are viewed as constants. The mean of Y_i given X_i is denoted by $\mu_{Y_i|X_i}$. First, we note that

$$\mu_{Y_i|X_i} = \beta_0 + \beta_1 X_i + \mu_{\epsilon_i},$$

because the mean of a constant ($\beta_0 + \beta_1 X_i$) is the constant. Finally, remembering the mean of the error term, ϵ_i equals 0, we get:

$$\mu_{Y_i|X_i} = \beta_0 + \beta_1 X_i. \tag{22.11}$$

The variance of Y_i is

$$\sigma^2_{Y_i} = \sigma^2_{\epsilon_i} = \sigma^2, \tag{22.12}$$

because the variance of a constant ($\beta_0 + \beta_1 X_i$) is 0.

The important facts for you to know are:

- The relationship between X and Y is such that the mean of Y given the value of X is a linear function of X with y-intercept given by β_0 and slope given by β_1.

- The variance of the Y's around their means (remember the mean depends on X) is σ^2, for every case regardless of its value of X.

22.2.1 Point Estimates of the Slope, Intercept and Variance

The first issue we turn to is: How do we use data to estimate the values of β_1, β_0 and σ^2? This turns out to be quite easy.

First, I will refer to Equation 22.11 as **the equation of the population regression line**. We estimate the slope, β_1, and intercept, β_0, by using the Principle of Least Squares, as we did in Chapter 21. In particular, the point estimate of β_1 is

$$b_1 = r(s_2/s_1), \tag{22.13}$$

and the point estimate of β_0 is

$$b_0 = \bar{y} - b_1\bar{x}. \tag{22.14}$$

The point estimate of σ^2 is a bit trickier. First, by rewriting Equation 22.10 we obtain:

$$\epsilon_i = Y_i - \beta_0 - \beta_1 X_i. \tag{22.15}$$

If we were Nature, then we could calculate the n observed values of ϵ and calculate their variance. Sadly, we are not Nature and must proceed as follows. Replace Y_i and X_i by their observed values and replace the unknown β_0 and β_1 by their point estimates. After these four replacements, the right side of Equation 22.15 becomes:

$$y_i - b_0 - b_1 x_i = y_i - \hat{y}_i = e_i.$$

Thus, in a sense, the residual, e_i, is the *sample version* of the error, ϵ_i. Thus, it makes sense to use the variance of the residuals to estimate the variance of ϵ_i. We do this and our result is that the point estimate of σ^2 is

$$s^2 = \frac{\text{SSE}}{n-2}.$$

22.3 Three Confidence Intervals, a Prediction Interval and a Test

It is possible to estimate β_1, β_0 and σ^2 with confidence. In this section I will address these problems along with two closely related problems. Finally, I will present a test of hypotheses for the slope of the population regression line.

First, let's deal with σ^2. There is a confidence interval formula for σ^2, but I will not present it, for a variety of reasons that (sadly?) I have no time to discuss.

There exist algebraic formulas for confidence intervals for both β_1 and β_0, but, frankly, they are no fun and nobody uses them by hand. Instead, I will show you how to obtain these confidence intervals from Minitab output.

Table 22.1 on page 592 presents our familiar Minitab output for the regression of final exam score on midterm exam score for $n = 35$ students. I will use this output to illustrate the following result.

Result 22.2 (Confidence interval estimates of the slope and intercept.) *The confidence interval estimates of the slope β_1, and intercept, β_0, of the population regression line are*

$$b_1 \pm t^*(SE(b_1)) \text{ and} \tag{22.16}$$

$$b_0 \pm t^*(SE(b_0)) \tag{22.17}$$

respectively.

In these formulas, t^* is obtained from the t-curve with $df = (n-2)$, with our usual method, first introduced for the t-curve in Chapter 17. The expression "SE(b_1)" ["SE(b_0)"] is the *estimated standard error of* b_1 [b_0] and its value must be obtained from computer output.

Based on our earlier work in these notes, it is reasonable to wonder whether the confidence level in the above result is exact or an approximation. **For all of the inference procedures in this chapter, including the above result, the confidence levels, prediction probabilities and P-values are exact with the additional assumption that the error terms, ϵ_i, have a Normal curve for their pdf.** Without this assumption, the results are approximations. Sadly, we do not have time to explore the quality of these approximations.

Suppose that I choose 95% for my confidence level; because $df = n - 2 = 35 - 2 = 33$, you may verify by using our t-curve website that $t^* = 2.035$. For the slope, using Table 22.1, we find that the estimated standard error of the estimated slope is 0.1501; it's in the *SE(Coef)* column in the *Midterm* row. Thus, the 95% confidence interval estimate of β_1, the population slope, is:

$$0.4516 \pm 2.035(0.1501) = 0.4516 \pm 0.3055 = [0.1461, 0.7571].$$

First, I conclude, qualitatively, that the population slope is a positive number. Recalling that the slope measures the change in the mean of Y for a unit change in X, this qualitative result is unsurprising; it states that a higher midterm score yields a higher mean score on the final. How much higher? This is where the endpoints of the confidence interval come into play: if the score on the midterm increases by one point, then the mean score on the final increases by at least 0.1461 and at most 0.7571 points. Subjectively, I consider this interval to be very wide; perhaps not very useful.

Next, let's consider the intercept, β_0. Minitab presents the relevant output in the *Constant* row, with *Constant* falling under the heading *Predictor*; admittedly, the terminology is a bit confusing. Anyways, for expediency, I will stick to 95% for my confidence level; thus, t^* remains equal to 2.035. Therefore, the 95% confidence interval estimate of β_0, the y-intercept of the population regression line, is:

$$68.42 \pm 2.035(7.963) = 68.42 \pm 16.20 = [52.22, 84.62].$$

I conjecture that I know what you are thinking: Bob is going to interpret this interval. Well, I am not. Here is why.

Literally, the population intercept is the mean value of Y given $X = 0$. The smallest midterm score in the data set is $X = 39.0$; thus, we have no idea whether the linear relationship in the data set extends down to $X = 0$. (Sound familiar? We talked about this idea in the Fish Activity study

in Chapter 21.) Also, as a teacher, I am not particularly interested in predicting a final exam score for a student who scores 0 on my midterm! (Nobody has ever scored 0 on any of my tests!) In particular, if a student said,

> Yeah, I scored 0 on the midterm. The final is inconvenient for me, so why not give me my predicted score, $b_0 = 68.42$?

Even though the gift of 68.42 would not save the student from an F in the course, as a matter of principle I would never seriously consider such a request! (And I have no evidence any student would make this request; my students are serious about their educations.)

To summarize the above, if $X = 0$ is outside the range of the data and/or not of scientific interest, then, even though it is possible to calculate a confidence interval estimate of β_0, I don't do it.

Be careful in your reading of the previous paragraph. I am **not saying** that estimating β_0 is unimportant. It is **important** because it is part of the population regression line. What I am saying is that unless $X = 0$ is both in the range of the data **and** of scientific interest to the researcher, then β_0, **by itself**, it is not important. If this distinction is confusing, I hope that the next confidence interval formula will help.

22.3.1 Confidence Interval for the Mean Response for a Given Value of the Predictor

Let us consider a specific possible value of X, call it x_0. Now given that $X = x_0$, the mean of Y is $\beta_0 + \beta_1 x_0$; call this μ_0. We can use the computer output to obtain a point estimate and confidence interval estimate of μ_0.

For example, suppose we select $x_0 = 49.0$. Then, the point estimate is

$$b_0 + b_1(49.0) = 68.42 + 0.4516(49.0) = 90.548$$

If, however, you look at the output in Table 22.1, you will find this value, 90.548, in the column *Fit* and the row *Midterm* $= 49.0$. (This is observation 8, 9 or 10.) Of course, this is not a huge aid because, frankly, the above computation of 90.548 was pretty easy. But it's the next entry in the output that is useful. Just to the right of *Fit* is *SE(Fit),* where, as before, SE is the abbreviation for estimated standard error. Thus, we are able to calculate a confidence interval estimate of μ_0:

$$\text{Fit } \pm t^*(\text{SE(Fit)}). \tag{22.18}$$

For the current example, the 95% confidence interval estimate of the mean of Y given $X = 49.0$ is

$$90.548 \pm 2.035(0.968) = 90.548 \pm 1.970 = [88.578, 92.518].$$

The obvious question is: We were pretty lucky that $X = x_0 = 49.0$ was in our computer output; what do we do if our x_0 isn't there? It turns out that the answer is easy: trick the computer. Here is how.

Suppose we want to estimate the mean of Y given $X = 51.0$. An inspection of the *Midterm* column in the computer output in Table 22.1 reveals that there is no row for $X = 51.0$. Go back to the data set and add a 36th *student*. For this student, enter 51.0 for Midterm and a **missing value** for Final. (For Minitab, my software package, this means you enter a *.) Then rerun the regression analysis. In all of the computations the computer will ignore the 36th student because it has no value for Y and therefore cannot be used in all of the needed computations. Thus, it is not used in any computations. As a result, the computer output is unchanged by the addition of this extra student. But, and this is the key point, the computer output includes observation 36 in the last section of the output, creating the row:

```
Obs  Midterm  Final     Fit SE(Fit)  Residual
 36     51.0      *  91.451   0.828         *
```

From this we see that the point estimate of the mean of Y given $X = 51.0$ is 91.451. (This, of course, is easy to verify.) But now we also have the *SE(Fit)*, so we can obtain the 95% confidence interval estimate of the mean of Y given $X = 51.0$:

$$91.451 \pm 2.035(0.828) = 91.451 \pm 1.685 = [89.766, 93.136].$$

22.3.2 Prediction of the Response for a Given Value of the Predictor.

Suppose that beyond our n cases for which we have data, we have an additional case. For this new case, we know that $X = x_{n+1}$, for a known number x_{n+1}, and we want to predict the value of Y_{n+1}. Now, of course,

$$Y_{n+1} = \beta_0 + \beta_1 x_{n+1} + \epsilon_{n+1}.$$

We assume that ϵ_{n+1} is independent of all previous ϵ's, and, like the previous errors, has mean 0 and variance σ^2.

The natural prediction of Y_{n+1} is obtained by replacing the β's by their estimates and ϵ_{n+1} by its mean, 0. The result is

$$\hat{y}_{n+1} = b_0 + b_1 x_{n+1}.$$

We recognize this as the Fit for $X = x_{n+1}$; as such, its value and its SE are both presented in (or can be made to be presented in) our computer output.

Following our approach to prediction in Chapter 14, we compare the actual response to the predicted response via subtraction, giving us:

$$W = Y_{n+1} - \hat{y}_{n+1}.$$

Our Result 14.1 tells us that the estimated variance of W is

$$s^2 + [\text{SE(Fit)}]^2.$$

(It is ok if you don't bother with verifying that this result applies here; it's late in the semester! This variance, however, has a nice interpretation. The estimated variance of the prediction is the sum of two terms: The estimated variance of an observation around the population regression line:

s^2; and the estimated variance due to the population regression line being estimated at the point x_{n+1}: [SE(Fit)]^2.)

The prediction interval for Y_{n+1} is:

$$\text{Fit } \pm t^* \sqrt{s^2 + [\text{SE(Fit)}]^2}. \tag{22.19}$$

For example, suppose that $x_{n+1} = 49.0$. From our computer output, and our earlier work, the point prediction of y_{n+1} is 'Fit,' which is 90.548. The estimated variance of W is

$$(4.635)^2 + (0.968)^2 = 22.4202;$$

thus, the estimated standard error is $\sqrt{22.4202} = 4.735$. Thus, the 95% prediction interval for Y_{n+1} is

$$90.548 \pm 2.035(4.735) = 90.548 \pm 9.636 = [80.912, 100.184].$$

In words, for an additional student who scores 49.0 on the midterm, at the 95% probability level, we predict that this student's final exam score will be between, roughly, 81.0 and 100 points, inclusive. This is not a particularly useful prediction interval.

22.3.3 A Test of Hypotheses

The results of this short subsection are similar to the results we had in Section 18.1.

In many regression analyses, a key question is whether a regression is needed. In particular, researchers often want to test the null hypothesis that the slope of the population regression line, β_1, equals zero:

$$H_0 : \beta_1 = 0.$$

There are, as usual, three possible alternatives:

$$H_1 : \beta_1 > 0; H_1 : \beta_1 < 0; \text{ and } H_1 : \beta_1 \neq 0.$$

I recommend using the *Inconceivable Paradigm* to select the appropriate alternative.

The obvious starting point for the test statistic is the point estimator of β_1, denoted, when viewed as a random variable, by B_1, which has observed value b_1. As in our earlier work, the standardized version of B_1 is:

$$\frac{B_1 - \beta_1}{\sqrt{\text{Variance}(B_1)}}.$$

To obtain our test statistic we replace β_1 by its hypothesized value, 0, and estimate the variance in the denominator. The result is our test statistic, denoted by T with observed value t:

$$T = \frac{B_1}{\text{SE}(B_1)} \text{ and}$$

$$t = \frac{b_1}{\text{SE}(b_1)}. \tag{22.20}$$

In the formulas below, t is given above (Equation 22.20) and areas are computed under the t-curve with $df = (n - 2)$.

1. For the alternative $>$, the P-value equals the area to the right of t.

2. For the alternative $<$, the approximate P-value equals the area to the left of t.

3. For the alternative $\neq$, the approximate P-value equals twice the area to the right of $|t|$.

The Minitab output *anticipates* that you will want to do this test. Let's look at part of the output again for the exam scores' data:

Predictor	Coef	SE(Coef)	T	P
Constant	68.420	7.963	8.59	0.000
Midterm	0.4516	0.1501	3.01	0.005

The observed value of the test statistic

$$t = 0.4516/0.1501 = 3.01,$$

is given in the fourth column (headed 'T'). The P-value for the alternative $\neq$ is given in the fifth column (headed 'P') and is equal to 0.005. Thus, for the alternative $>$, the P-value equals $0.005/2 = 0.0025$. More precisely, using our t-curve website, the area under the t-curve with $df = 33$ to the right of 3.01 is 0.00249.

Often in regression, the researcher has a special possible value of interest for the population slope, denoted by β_{10} and read as *beta-one-zero* or *beta-one-naught*, but never as *beta-ten*. In this situation the null hypothesis becomes:

$$H_0 : \beta_1 = \beta_{10}.$$

The three possible alternatives are:

$$H_1 : \beta_1 > \beta_{10}; H_1 : \beta_1 < \beta_{10}; \text{ and } \beta_1 \neq \beta_{10}.$$

(Obviously, if $\beta_{10} = 0$ this new problem reduces to the problem we solved above.)

For this more general situation, the test statistic is

$$T = \frac{B_1 - \beta_{10}}{\text{SE}(B_1)}$$

with observed value

$$t = \frac{b_1 - \beta_{10}}{\text{SE}(b_1)}. \tag{22.21}$$

The earlier rules for the P-value apply to this new situation.

For this course—and our final exam—the above is the only test you are required to learn. Let me mention in passing that a researcher could adapt the above method to test a null hypothesis on the value of the intercept, β_0, or on the mean value of the response for a given value of the predictor. Indeed, the Minitab output gives the observed value of the test statistic and the two-sided P-value for the test of the null hypothesis that the intercept equals 0.

22.4 Extensions

The simple linear regression model begins with Equation 22.10, reproduced below:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

for $i = 1, 2, 3, \ldots n$. The model includes the assumptions about the sequence of ϵ's given earlier; namely, they are independent, mean equal to zero and variance equal to the unknown σ^2.

The word *simple* signifies that there is only one predictor. Also—although this fact often causes confusion—*linear* refers to the model being linear in the parameters β_0 and β_1 and **not** to the fact that the mean of Y is a linear function of X. This leads to the first generalization of our model.

Suppose that for some reason, you believe that the relationship between a nonnegative response Y and a nonnegative predictor X is

$$Y_i = \beta_0 + \beta_1 X_i^2 + \epsilon_i. \tag{22.22}$$

Literally, this is not the same as the simple linear regression model, but it can be **transformed** into the simple linear regression model quite easily. All we do is define a new predictor X_i^* to equal X_i^2. With this substitution, Equation 22.22 becomes

$$Y_i = \beta_0 + \beta_1 X_i^* + \epsilon_i,$$

which **is** the simple linear regression model with predictor denoted by X_i^*.

In Chapter 13 you learned about the family of Poisson distributions. Recall that for a Poisson distribution the mean equals the variance. In many applications to science the variance of the response will increase with the mean, although they won't necessarily be equal. In particular, consider the following modification of the simple linear regression model for a predictor that must be positive.

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i. \tag{22.23}$$

In this equation, conditional on the value of X_i, the error term, ϵ_i, has mean 0 and variance $\sigma^2 X_i^2$. Thus, this is **not** the simple linear regression model because the error terms do **not** have constant variance. How can we fix this? The answer is quite simple. Divide both sides of Equation 22.23 by X_i (remember, it must be positive), to get:

$$Y_i/X_i = \beta_0/X_i + \beta_1 + \epsilon_i/X_i.$$

Next, make the following definitions:

$$Y_i^* = Y_i/X_i; X_i^* = 1/X_i; \beta_0^* = \beta_1; \beta_1^* = \beta_0; \text{ and } \epsilon_i^* = \epsilon_i/X_i.$$

Thus, Equation 22.23 becomes

$$Y_i^* = \beta_0^* + \beta_1^* X_i^* + \epsilon_i^*,$$

the simple linear regression model for response Y_i^* and predictor X_i^* because the errors ϵ_i^* are independent, mean 0 with constant variance σ^2.

Sometimes, researchers assume a multiplicative error rather than the additive error in the simple linear regression model. In particular, define $\epsilon_i^* = \exp(\epsilon_i)$, where the ϵ_i's satisfy the assumptions of the simple linear regression model. Consider the new model:

$$Y_i = \exp(\beta_0 + \beta_1 X_i) \times \epsilon_i^*. \tag{22.24}$$

If we take the natural logarithm ($\ln$) of both sides of this equation, we get:

$$\ln(Y_i) = \beta_0 + \beta_1 X_i + \epsilon_i,$$

which is the simple linear regression model for response $Y_i^* = \ln(Y_i)$.

Finally, with the definition of ϵ_i^* in the previous paragraph, consider the model with $\beta_0 > 0$ and the predictor and response both constrained to be positive:

$$Y_i = \beta_0 X_i^{\beta_1} \times \epsilon_i^*. \tag{22.25}$$

If we take the natural logarithm ($\ln$) of both sides of this equation, we get:

$$\ln(Y_i) = \ln(\beta_0) + \beta_1 \ln(X_i) + \epsilon_i,$$

which is the simple linear regression model for:

$$Y_i^* = \ln(Y_i); X_i^* = \ln(X_i); \beta_0^* = \ln(\beta_0); \text{ and } \beta_1^* = \beta_1.$$

Let me end with two comments about the above list of examples.

1. Rather obviously, the list above is not an exhaustive list of models that can be easily transformed to the simple linear regression model.

2. Each example began with an equation relating the original response to the original predictor. A scientist typically *obtains* such an equation in one of two ways:

   - **Empirically:** By looking at a scatterplot of the data.
   - **Theoretically:** Some scientific theory leads to the belief that the relationship between Y and X should have the form given in the equation.

   Note that for either of these methods, not only should you focus on how Y varies with X, but on how the error terms enter the relationship: additive or multiplicative; constant or nonconstant variance.

22.5 Summary

In a regression analysis, each case has four numbers associated with it:

$$x, y, \hat{y} = b_0 + b_1 x \text{ and } e = y - \hat{y}.$$

In words, its predictor, response, predicted response and residual, respectively.

The researcher should draw a picture—dot plot or histogram—of the residuals. Note that a case's residual being an outlier is a different notion than a case being isolated, although the former implies the latter.

For any regression analysis, the mean of the residuals equals zero: $\bar{e} = 0$. The sum of squared residuals is denoted by SSE:

$$\text{SSE } = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2.$$

The residuals have $(n - 2)$ degrees of freedom and the variance of the residuals is

$$s^2 = \frac{\text{SSE}}{n - 2}.$$

The standard deviation of the residuals is, of course, $s = \sqrt{s^2}$.

For each individual case, the residual tells us *how well* the regression line performed, in the following sense.

- If the residual equals 0, then $y = \hat{y}$ and the regression line's prediction is perfect.
- If the residual is greater than 0, then $y > \hat{y}$ and the regression line's prediction is too small.
- If the residual is less than 0, then $y < \hat{y}$ and the regression line's prediction is too large.
- The farther the value of the residual is from 0, in either direction (positive or negative), the worse the regression line predicts the actual response.
- In short, a residual that is *close to zero*, positive or negative, indicates that the regression line did a *good job* predicting the response. A residual that is *far from zero*, positive or negative, indicates that the regression line did a *bad job* predicting the response. The distinction between being *close to zero* or *far from zero* should be based on the scientific goals of the study.

The bulleted items above are concerned with evaluating the regression line for individual cases. We also want an overall evaluation of the quality of the regression line.

If a scientist has a specific number that distinguishes between *close to zero* and *far from zero*, then the regression line can be evaluated with simple counting: count the number of cases for which the prediction is good and compare it to the number of cases for which the prediction is bad. I did this earlier in the chapter where I (subjectively) decided that the boundary between a good and bad prediction of a batting average was 20 points (0.020). I found that with this boundary, (only) 57% of the cases were predicted well.

Table 22.3: The ANOVA table for simple linear regression.

Source	DF	SS
Regression	1	SSR
Residual Error	$n-2$	SSE
Total	$n-1$	SSTO

However useful it is for a scientist in a particular study to specify the boundary between *good* and *bad* predictions, this activity does **not** lend itself to mathematical analysis of a general nature. Instead, we focus on the value of s, the standard deviation of the residuals, and use the Empirical Rule for interpreting s to measure the effectiveness of the regression line. For example, according to the Empirical Rule approximately 68% of the residuals will be between $-s$ and s; in other words, for approximately 68% of the cases, the actual response will be within s of the predicted response.

I recommend comparing the standard deviation of the residuals, s, to the standard deviation of the responses, s_2. This amounts to comparing the best predictions using X with the best predictions that ignore X.

Mostly in work beyond this chapter, the Analysis of Variance Table is useful. It is presented in Table 22.3. The various sum of squares are defined by:

$$\text{SSR } = \sum(\hat{y}_i - \bar{y})^2; \text{ SSE } = \sum(y_i - \hat{y}_i)^2; \text{ and SSTO } = \sum(y_i - \bar{y})^2.$$

Note that in this table, both the degrees of freedom and sums of squares *sum*, in the sense that:

$$1 + (n-2) = (n-1); \text{ and SSR } + \text{ SSE } = \text{ SSTO}.$$

The coefficient of determination, R^2 is a popular summary statistic for a regression analysis:

$$R^2 = \frac{\text{SSTO } - \text{ SSE}}{\text{SSTO}}.$$

The denominator of R^2 equals the total squared error in the responses. The numerator of R^2 equals the amount of squared error that is removed by using X to predict Y via the regression line. Thus, the ratio of R^2 equals the proportion (usually reported as a percentage) of the total squared error in the response that can be explained by a linear relationship with the predictor.

For every regression analysis, R^2 equals the square of the correlation coefficient:

$$R^2 = r^2.$$

This identity gives us another interpretation of the correlation coefficient, r.

For inference, we assume that the simple linear regression model is true:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \text{ for } i = 1, 2, 3, \ldots n.$$

In this equation, β_0 and β_1 are parameters whose values are known to Nature, but unknown to the researcher. The error terms,

$$\epsilon_1, \epsilon_2, \epsilon_3, \ldots, \epsilon_n,$$

are assumed to be independent random variables, each with mean equal to 0 and variance equal to σ^2, the third unknown parameter of the model. Statistical inference for this model conditions on the values of the predictor.

A consequence of this model is that the mean value of Y for a given value of $X = x$ is equal to:

$$\beta_0 + \beta_1 x.$$

The point estimates of β_1 and β_0 are obtained by applying the Principle of Least Squares, as we did in Chapter 21, yielding

$$b_1 = r(s_2/s_1); \text{ and } b_0 = \bar{y} - b_1\bar{x}, \text{ respectively.}$$

The point estimate of σ^2 is the variance of the residuals, $s^2 = \text{SSE}/(n-2)$.

I do not give you algebraic formulas for confidence intervals, prediction intervals and testing. Instead, I show you how to obtain answers—intervals and P-values—from Minitab computer output. In particular, confidence intervals for β_1 and β_0 are given in Result 22.2 on page 601.

For a specified value of the predictor, call it x_0, the mean value of the response is

$$\mu_0 = \beta_0 + \beta_1 x_0.$$

The point estimate of μ_0 is

$$b_0 + b_1 x_0,$$

which Minitab denotes as *Fit*. The estimated standard error of the *Fit* is denoted by SE(Fit). The confidence interval for μ_0 is

$$\text{Fit } \pm t^*(\text{SE(Fit)}).$$

Suppose that beyond our n cases for which we have data, we have an additional case. For this new case, we know that $X = x_{n+1}$, for a known number x_{n+1}, and we want to predict the value of Y_{n+1}. The point prediction is:

$$\hat{y}_{n+1} = b_0 + b_1 x_{n+1}.$$

We recognize this as the Fit for $X = x_{n+1}$; as such, its value and SE(Fit) are both presented in (or can be made to be presented in) our computer output. The prediction interval for Y_{n+1} is:

$$\text{Fit } \pm t^* \sqrt{s^2 + [\text{SE(Fit)}]^2}.$$

Tests of hypotheses also are possible for the simple linear regression model. Sometimes the researcher has a special possible value of interest for the population slope, denoted by β_{10}. The null hypothesis is:

$$H_0 : \beta_1 = \beta_{10}.$$

The three possible alternatives are:

$$H_1 : \beta_1 > \beta_{10}; H_1 : \beta_1 < \beta_{10}; \text{ and } \beta_1 \neq \beta_{10}.$$

The test statistic is

$$T = \frac{B_1 - \beta_{10}}{\text{SE}(B_1)}$$

with observed value

$$t = \frac{b_1 - \beta_{10}}{\text{SE}(b_1)}.$$

The rules for obtaining the P-value are given on page 605.

Finally, Section 22.4 presents several models that can be transformed easily into the simple linear regression model.

22.6 Practice Problems

1. A simple linear regression analysis with $n = 5$ yields the numbers in the following table.

x:	-2	-1	0	1	2
e:	1	-2	$+2$	b	c

   Determine the values of b and c. (Hint: Use Equation 22.1.)

2. A simple linear regression analysis with $n = 10$ yields the following (partial) ANOVA table.

Source	DF	SS
Regression	a	496
Residual Error	b	800
Total	c	d

   (a) Determine the values of a–d in the ANOVA table.

   (b) Calculate the values of s and s_2. Explain what you have found.

   (c) Calculate the value of R^2; interpret the number you obtain.

   (d) What can you say about the value of the correlation coefficient, r?

3. A simple linear regression analysis yields the following (partial) ANOVA table.

Source	DF	SS
Regression	a	b
Residual Error	20	c
Total	d	2000

   In addition, $R^2 = 0.800$.

   (a) Determine the values of a–d in the ANOVA table.

   (b) Calculate the values of s and s_2. Explain what you have found.

4. Table 22.4 presents edited Minitab regression output for the exam scores data for all $n = 36$ students; i.e., it includes the isolated case $(35.5, 95.5)$.

   (a) Calculate the 95% confidence interval estimate of the slope of the regression line. Compare your answer to the answer earlier in this chapter for $n = 35$ students and comment.

   (b) Calculate the P-value for the alternative $\beta_1 > 0$. Compare your answer to the answer earlier in this chapter for $n = 35$ students and comment.

   (c) Calculate the P-value for the alternative $\beta_1 < 0.75$.

(d) Calculate the 95% confidence interval estimate of the mean response for $X = 49.0$. Compare your answer to the answer earlier in this chapter for $n = 35$ students and comment.

(e) Calculate the 95% prediction interval for a future response for $X = 49.0$. Compare your answer to the answer earlier in this chapter for $n = 35$ students and comment.

(f) Determine the ANOVA table for this analysis.

5. Table 22.5 presents edited Minitab regression output for the batting averages data for $n = 123$ players, after Floyd Rayford has been deleted from the data set. Note also that for arithmetic convenience (for me!) I have multiplied all batting averages by 1000 that, for example, converts $x = 0.269$ to $x = 269$.

(a) Calculate the 95% confidence interval estimate of the slope of the regression line.

(b) Calculate the P-value for the alternative $\beta_1 > 0$.

(c) Calculate the P-value for the alternative $\beta_1 < 1$.

(d) Calculate the 95% confidence interval estimate of the mean response for $X = 309$.

(e) Calculate the 95% prediction interval for a future response for $X = 309$.

Table 22.4: Edited Minitab output for the regression of final exam score on midterm exam score for 36 students.

```
The regression equation is: Final = 76.5 + 0.302 Midterm

Predictor        Coef    SE(Coef)          T        P
Constant       76.536       7.239      10.57    0.000
Midterm        0.3023      0.1375       2.20    0.035
S = 4.850       R-Sq = 12.4%

Obs   Midterm   Final       Fit  SE(Fit)  Residual
  1      39.0    83.5    88.325    2.000    -4.825
  2      43.0    89.0    89.534    1.514    -0.534
  3      44.0    92.0    89.837    1.399     2.163
  4      44.5    82.0    89.988    1.344    -7.988
  5      46.0    93.0    90.441    1.186     2.559
  6      48.0    87.0    91.046    1.002    -4.046
  7      48.5    91.5    91.197    0.963     0.303
 10      49.0    81.0    91.348    0.927   -10.348
 12      49.5    95.0    91.499    0.896     3.501
 13      50.0    97.5    91.650    0.868     5.850
 15      53.0    93.0    92.557    0.814     0.443
 17      53.5    99.0    92.708    0.825     6.292
 19      54.5    83.0    93.011    0.863   -10.011
 20      55.0    94.5    93.162    0.889     1.338
 23      55.5    92.5    93.313    0.920    -0.813
 26      56.5    90.5    93.615    0.993    -3.115
 27      57.0    92.0    93.766    1.035    -1.766
 28      57.5    97.5    93.917    1.079     3.583
 32      58.5    97.5    94.220    1.174     3.280
 35      59.0    94.0    94.371    1.225    -0.371
 36      35.5    95.5    87.267    2.449     8.233
```

Table 22.5: Edited Minitab output for the regression of 1986 batting average multiplied by 1000 on 1985 batting average multiplied by 1000 for 123 baseball players. Note the Floyd Rayford has been deleted from the data set.

The regression equation is: 1986BA = 83.0 + 0.681 1985BA

Predictor	Coef	SE(Coef)	T	P
Constant	83.03	21.55	3.85	0.000
1985BA	0.68114	0.08055	8.46	0.000

S = 24.82 R-Sq = 37.1%

Analysis of Variance

Source	DF	SS
Regression	1	44060
Residual Error	121	74562
Total	122	118621

Obs	1985BA	1986BA	Fit	SE(Fit)	Residual
1	265	264	263.53	2.24	0.47
2	309	296	293.50	4.12	2.50
3	268	240	265.57	2.24	-25.57
4	243	229	248.54	2.91	-19.54
5	289	289	279.88	2.90	9.12
6	266	286	264.21	2.24	21.79
7	231	238	240.37	3.61	-2.37
8	275	309	270.34	2.35	38.66
9	304	300	290.09	3.78	9.91

22.7 Solutions to Practice Problems

1. Let's expand the given table:

x:	-2	-1	0	1	2	Total
e:	1	-2	2	b	c	0
xe:	-2	2	0	b	$2c$	0

   Thus,

$$1 + b + c = 0 \text{ and } b + 2c = 0.$$

   Rewrite the first of these equations as $b = -c - 1$ and substitute it into the second equation:

$$-c - 1 + 2c = 0 \text{ or } c = 1; \text{ which yields } 1 + b + 1 = 0 \text{ or } b = -2.$$

2. (a) We know that a always equals 1; $b = (n-2) = (10-2) = 8$; $c = a + b = 1 + 8 = 9$ or $c = (n-1) = (10-1) = 9$; and $d = 496 + 800 = 1296$.

   (b) First, $s^2 = 800/b = 800/8 = 100$; thus, $s = 10$. Next, $s_2^2 = 1296/9 = 144$; thus, $s_2 = 12$. Finally, $s/s_2 = 10/12 = 0.833$; thus, the standard deviation of prediction errors using the regression line is 16.7% smaller than the standard deviation of prediction errors not using the regression line.

   (c) $R^2 = 496/1296 = 0.383$. In words, 38.3% of the squared error in the response is explained by a linear relationship with the predictor.

   (d) The correlation coefficient $r = \pm\sqrt{0.383} = \pm 0.619$.

3. (a) We know that $a = 1$, thus $d = 1 + 20 = 21$.
   We know that $R^2 = 0.800$; thus:

$$0.800 = \text{SSR}/\text{SSTO} = \text{SSR}/2000; \text{ which gives SSR} = b = 1600.$$

   By subtraction, $c = 2000 - 1600 = 400$.

   (b) We know that

$$s^2 = 400/20 = 20; \text{ and } s_2^2 = 2000/21 = 95.24.$$

   Thus, $s = 4.472$, $s_2 = 9.759$ and $s/s_2 = 4.472/9.759 = 0.458$. Thus, the standard deviation of predictions using X is 54.2% smaller than the standard deviation of predictions without using X.

4. First, note that in order to obtain 95% confidence (for the estimates) or probability (for prediction), we have $t^* = 2.033$ for $df = (n-2) = (36-2) = 34$.

   (a) Using the computer output, the 95% confidence interval estimate of β_1 is:

$$0.3023 \pm 2.033(0.1375) = 0.3023 \pm 0.2795 = [0.0228, 0.5818].$$

   This interval is narrower than the earlier interval (half-width is 0.2795 versus 0.3055), but is much closer to zero (center is 0.3023 versus 0.4516).

(b) The easiest way to obtain the P-value is to realize that it equals one-half of the two-sided P-value given by Minitab: $0.035/2 = 0.018$.

Alternatively, the observed value of the test statistic is

$$t = 0.3023/0.1375 = 2.1985.$$

With the help of the website the area under the t-curve with $df = 34$ to the right of 2.1985 is 0.0174. (Paraphrasing Shakespeare, "Much ado about not much.")

For the earlier analysis with $n = 35$, the P-value is much smaller, 0.0025.

(c) The observed value of the test statistic is

$$t = \frac{0.3023 - 0.75}{0.1375} = -0.4477/0.1375 = -3.256.$$

With the help of the website the area under the t-curve with $df = 34$ to the left of -3.256 is 0.0013.

(d) Using the computer output for observation 10, the 95% confidence interval estimate the mean of Y given $X = 49$ is

$$91.348 \pm 2.033(0.927) = 91.348 \pm 1.885 = [89.463, 93.233].$$

This point prediction is almost one point larger ($91.348 - 90.548 = 0.800$) than the earlier point prediction and this interval is narrower than the earlier interval (half-width is 1.885 versus 1.970).

(e) Again using the computer output for observation 10, we find that the estimated variance of the predicted value is:

$$(4.850)^2 + (0.927)^2 = 24.3818.$$

Thus, the estimated standard error of the predicted value is:

$$\sqrt{24.3818} = 4.938.$$

Thus, the 95% prediction interval is:

$$91.348 \pm 2.033(4.938) = 91.348 \pm 10.039 = [81.309, 101.387].$$

This interval is slightly narrower than the earlier one, but both intervals have little practical value because they are so wide. A score of 81.5 on the final is very different than a score of 100.

(f) First, the easy part of the ANOVA table:

Source	DF	SS
Regression	1	
Residual Error	34	
Total	35	

Next,

$$(4.85)^2 = 23.5225 = s^2 = \text{ SSE}/34; \text{ or SSE } = 34(23.5225) = 799.765.$$

Next,

$$0.124 = R^2 = 1 - (799.765/\text{SSTO}) \text{ or SSTO } = 799.765/0.876 = 912.974.$$

Thus, the ANOVA table is:

Source	DF	SS
Regression	1	113.209
Residual Error	34	799.765
Total	35	912.974

5. First note that for $df = 123 - 2 = 121$, for 95% confidence or probability, $t^* = 1.980$.

   (a) The 95% confidence interval estimate of the slope is:

   $$0.6811 \pm 1.980(0.08055) = 0.6811 \pm 0.1595 = [0.5216, 0.8406].$$

   (b) The P-value equals one-half of the value in the table, which is one-half of 0.000, or 0.000. More precisely, using Minitab for $t = 8.46$ I obtain 3.62×10^{-14}. This is a really small P-value!

   (c) The observed value of the test statistic is

   $$t = \frac{0.6811 - 1}{0.08055} = -3.959.$$

   The area under the t-curve with $df = 121$ to the left of -3.959 is equal to—with the help of Minitab—0.0000368; or approximately 37 in one million. This is a very small P-value.

   (d) The 95% confidence interval estimate of the mean response given $X = 309$ is

   $$293.50 \pm 1.980(4.12) = 293.50 \pm 8.16 = [285.34, 301.66].$$

   (e) Again using the computer output for observation 2, we find that the estimated variance of the predicted value is:

   $$(24.82)^2 + (4.12)^2 = 633.0068.$$

   Thus, the estimated standard error of the predicted value is:

   $$\sqrt{633.0068} = 25.16.$$

   Thus, the 95% prediction interval is:

   $$293.50 \pm 1.980(25.16) = 293.50 \pm 49.82 = [243.68, 343.32].$$

   This interval is very wide. As a baseball fan, I consider it to be almost totally worthless; 343 is a great batting average and 243 is—while not horrible—pretty poor.

22.8 Homework Problems for Chapter 21

1. (Hypothetical data.) Fifty students in a Statistics class take midterm and final exams. Below are selected summary statistics for these data.

Exam	Mean	Stand. Dev.
Midterm	50.00	10.00
Final	70.00	15.00

   Also, the correlation coefficient of the the two exam scores is $r = 0.48$.

   (a) Determine the equation of the regression line for using the score on the midterm exam to predict the score on the final exam.

   (b) Determine the equation of the regression line for using the score on the final exam to predict the score on the midterm exam.

   (c) Sally scores 60 on the midterm exam. Use your equation from (a) to obtain her predicted score on the final exam.

   (d) Tom scores 80 on the final exam. Use your equation from (b) to obtain his predicted score on the midterm exam.

   (e) Refer to (c). Given that Sally actually scored 82 on the final exam, calculate her residual.

   (f) Refer to (d). Given that Tom actually scored 47 on the midterm exam, calculate his residual.

2. (Hypothetical data.) We have two measurements, x and y, on each of 500 children. We use these data to obtain the regression line for using x to predict y.

   The means for the 500 children are: 110 for x and 190 for y.

   Ron's value for x is 10 less than the mean of the x values. In addition, Ron's y is 20 less than predicted from his x.

   Given that Ron's $y = 140$, calculate the regression line.

22.9 Homework Problems for Chapter 22

Below is edited output from a study on the heights and weights of 10 members of the WNBA Chicago Sky team. For the purpose of the following questions, we will view these 10 women as a random sample from the population of all female professional basketball players.

```
The regression equation is
Weight = - 304 + 6.57 Height

Predictor      Coef   SE Coef
Height         6.57    0.9256

S = 12.36       R-Sq = 86.3%

  Ht       Wt      Fit  SE Fit
75.0   184.00   188.47    4.59
69.0   162.00   149.08    5.02
74.0   162.00   181.91    4.18
78.0   200.00   208.17    6.49
80.0   240.00   221.30    8.05
```

1. Briefly explain (this means in words) the meaning of the regression equation. Make sure you interpret the number 6.57.
2. Bert looks at the regression equation and states, "This is ridiculous! A woman cannot have a negative weight!"
   (a) Can you guess why Bert made this statement? If yes, explain.
   (b) Do you agree or disagree with Bert? Explain your answer.
3. Calculate the 95% confidence interval for the slope of the simple linear regression model.
4. Calculate the 95% confidence interval for the mean weight of women in the population who are 69 inches tall.
5. Calculate the 95% confidence interval for the mean weight of women in the population who are 78 inches tall.
6. I select a woman at random from the population and note that she is 74 inches tall. Given this information, obtain the 95% prediction interval for her weight.
7. I select a woman at random from the population and note that she is 80 inches tall. Given this information, obtain the 95% prediction interval for her weight.
8. Calculate the value of the correlation coefficient r.
9. (Tricky.) Calculate the slope of the regression line for using weight to predict height.