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.

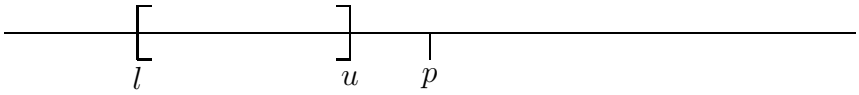
Before I get to my first example, it is convenient to have a not-so-brief digression. I want to introduce you to what I call the **Goldilocks metaphor**, a device that repeatedly will prove useful in these notes.

According to Wikipedia,

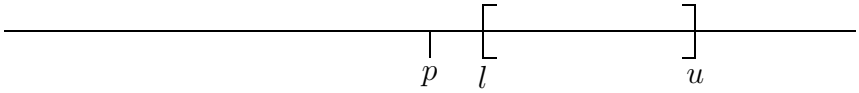
> *The Story of the Three Bears* (sometimes known as *The Three Bears*, *Goldilocks and the Three Bears* or, simply, *Goldilocks*) is a fairy tale first recorded in narrative form by British author and poet Robert Southey, and first published anonymously in a volume of his writings in 1837. The same 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.

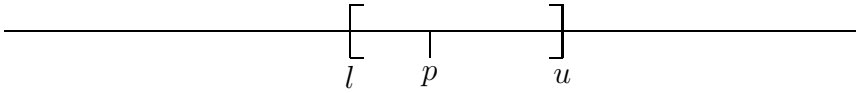
The CI is too small, $u < p$:



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 comeupance 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 in the precomputer 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,

$$h = z^* \sqrt{\frac{\hat{p}\hat{q}}{n}}. \tag{12.6}$$

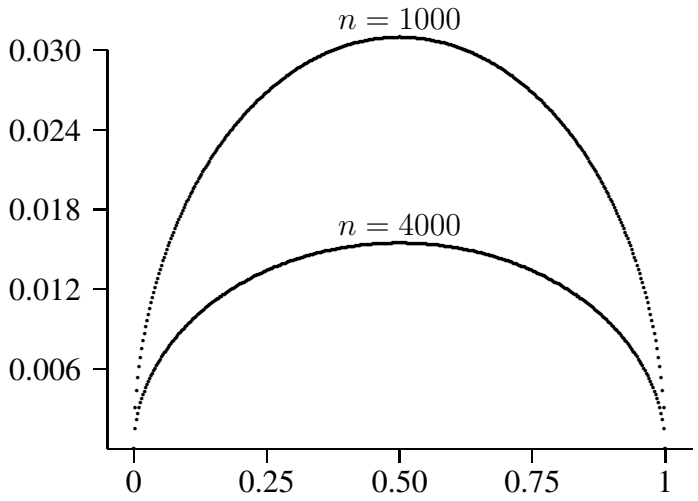
There are three numbers in this formula for h that we can vary: z^*, which is determined by the choice of confidence level; n; and $\hat{p}$, remembering that $\hat{q} = 1 - \hat{p}$. My first effort is to fix the confidence level at 95%—i.e. $z^* = 1.96$—and see how h varies as a function of n and $\hat{p}$. 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 run 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 run 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 run 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 run 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.