Chapter 2

Bernoulli Trials

2.1 The Binomial Distribution

In Chapter 1 we learned about i.i.d. trials. In this chapter, we study a very important special case of these, namely Bernoulli trials (BT). If each trial yields has exactly two possible outcomes, then we have BT. B/c this is so important, I will be a bit redundant and explicitly present the assumptions of BT.

The Assumptions of Bernoulli Trials. There are three:

1. Each trial results in one of two possible outcomes, denoted success (S) or failure (F).
2. The probability of S remains constant from trial-to-trial and is denoted by p. Write $q = 1-p$ for the constant probability of F.
3. The trials are independent.

When we are doing arithmetic, it will be convenient to represent S by the number 1 and F by the number 0.

One reason that BT are so important, is that if we have BT, we can calculate probabilities of a great many events.

Our first tool for calculation, of course, is the multiplication rule that we learned in Chapter 1. For example, suppose that we have $n = 5$ BT with $p = 0.70$. The probability that the BT yield four successes followed by a failure is:

$$P(SSSSF) = ppppq = (0.70)^4(0.30) = 0.0720.$$

Our next tool is extremely powerful and very useful in science. It is the **binomial probability distribution**.

Suppose that we plan to perform/observe n BT. Let X denote the total number of successes in the n trials. The probability distribution of X is given by the following equation.

$$P(X = x) = \frac{n!}{x!(n-x)!}p^x q^{n-x}, \text{ for } x = 0, 1, \ldots, n. \tag{2.1}$$

To use this formula, recall that $n!$ is read 'n-factorial' and is computed as follows.

$$1! = 1; 2! = 2(1) = 2; 3! = 3(2)(1) = 6, 4! = 4(3)(2)(1) = 24;$$

and so on. By special definition, $0! = 1$.

(Note to the extremely interested reader. If you want to see *why* Equation 2.1 is correct, go to the link to the *Revised Student Study Guide* on my webpage and read the *More Mathematics* sections of Chapters 2, 3 and 5.)

I will do an extended example to illustrate the use of Equation 2.1.

Suppose that $n = 5$ and $p = 0.60$. I will obtain the probability distribution for X.

$$P(X = 0) = \frac{5!}{0!5!}(0.60)^0(0.40)^5 = 0.0102.$$

$$P(X = 1) = \frac{5!}{1!4!}(0.60)^1(0.40)^4 = 0.0768.$$

$$P(X = 2) = \frac{5!}{2!3!}(0.60)^2(0.40)^3 = 0.2304.$$

$$P(X = 3) = \frac{5!}{3!2!}(0.60)^3(0.40)^2 = 0.3456.$$

$$P(X = 4) = \frac{5!}{4!1!}(0.60)^4(0.40)^1 = 0.2592.$$

$$P(X = 5) = \frac{5!}{5!0!}(0.60)^5(0.40)^0 = 0.0778.$$

You should check the above computations to make sure you are comfortable using Equation 2.1.

Here are some guidelines for this class. If $n \leq 8$, you should be able to evaluate Equation 2.1 'by hand' as I have done above for $n = 5$.

For $n \geq 9$, I recommend using a statistical software package on a computer or the website I describe later. For example, the probability distribution for X for $n = 25$ and $p = 0.50$ is presented in Table 2.1.

Equation 2.1 is called the **binomial probability distribution** with **parameters** n and p; it is denoted by Bin(n, p). With this notation, we see that my earlier 'by hand' effort was the Bin(5,0.60) and Table 2.1 is the Bin(25,0.50).

Sadly, life is a bit more complicated than the above. In particular, a statistical software package should not be considered a panacea for the binomial. For example, if I direct my computer to calculate the Bin(n,0.50) for any $n \geq 1023$ I get an error message; the computer program is smart enough to realize that it has messed up and its answer is wrong; hence, it does not give me an answer. Why does this happen?

Well, consider the computation of $P(X = 1000)$ for the Bin(2000,0.50). This involves a really huge number ($2000!$) divided by the square of a really huge number ($1000!$), and them multiplied by a really really small positive number ($(0.50)^{2000}$). Unless the computer programmer exhibits incredible care in writing the code, the result will be an overflow or an underflow or both.

Table 2.1: The Binomial Distribution for $n = 25$ and $p = 0.50$.

x	$P(X = x)$	$P(X \leq x)$	$P(X \geq x)$
0	0.0000	0.0000	1.0000
1	0.0000	0.0000	1.0000
2	0.0000	0.0000	1.0000
3	0.0001	0.0001	1.0000
4	0.0004	0.0005	0.9999
5	0.0016	0.0020	0.9995
6	0.0053	0.0073	0.9980
7	0.0143	0.0216	0.9927
8	0.0322	0.0539	0.9784
9	0.0609	0.1148	0.9461
10	0.0974	0.2122	0.8852
11	0.1328	0.3450	0.7878
12	0.1550	0.5000	0.6550
13	0.1550	0.6550	0.5000
14	0.1328	0.7878	0.3450
15	0.0974	0.8852	0.2122
16	0.0609	0.9461	0.1148
17	0.0322	0.9784	0.0539
18	0.0143	0.9927	0.0216
19	0.0053	0.9980	0.0073
20	0.0016	0.9995	0.0020
21	0.0004	0.9999	0.0005
22	0.0001	1.0000	0.0001
23	0.0000	1.0000	0.0000
24	0.0000	1.0000	0.0000
25	0.0000	1.0000	0.0000
Total	1.0000		

Figure 2.1: The Bin(100, 0.5) Distribution.

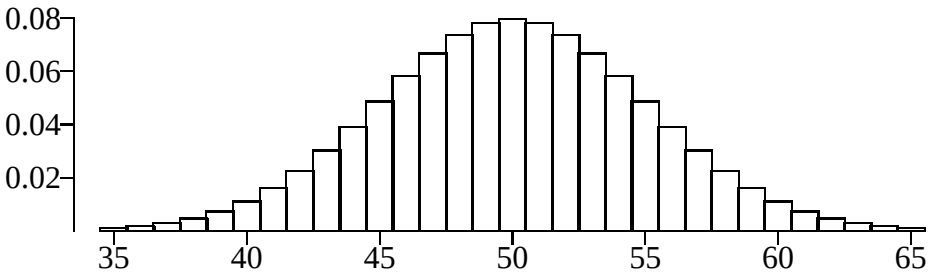


Figure 2.2: The Bin(100, 0.2) Distribution.

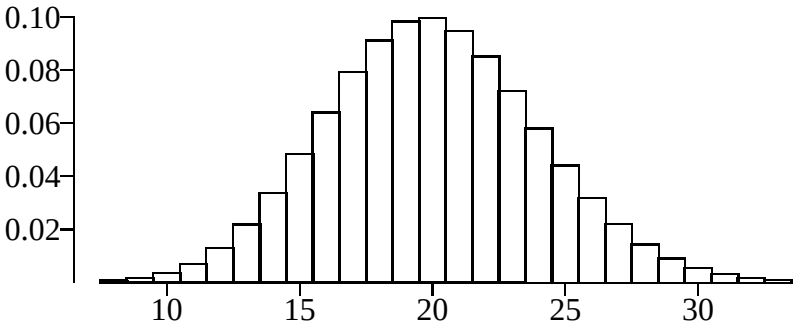


Figure 2.3: The Bin(25, 0.5) Distribution.

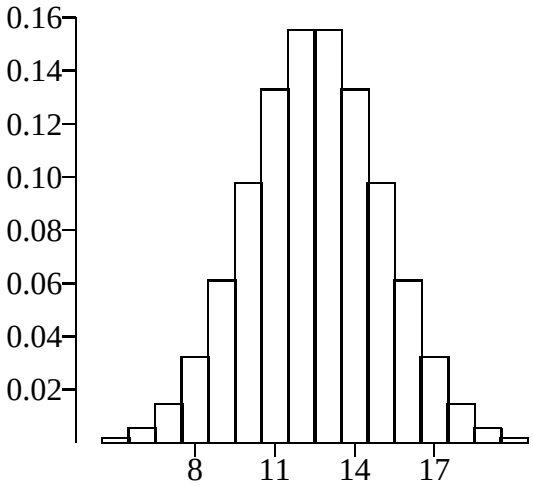
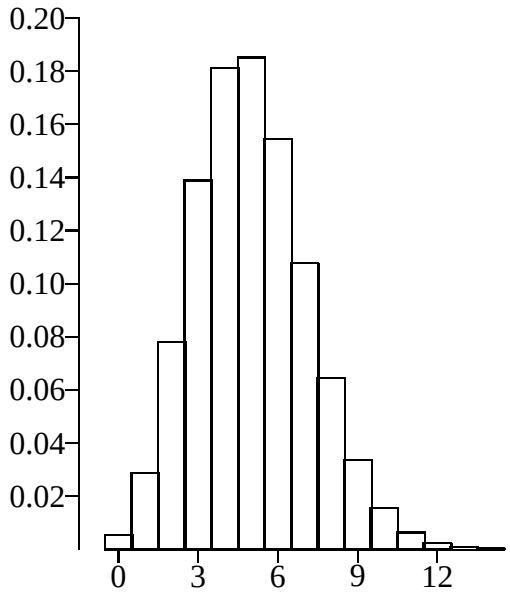


Figure 2.4: The Bin(50, 0.1) Distribution.



Before we condemn the programmer for carelessness or bemoan the limits of the human mind, note the following. We do not need to evaluate Equation 2.1 for large n's b/c there is a very easy way to obtain a good approximation to the exact answer.

Figures 2.1–2.4 **probability histograms** for several binomial probability distributions. Here is how they are drawn. The method I am going to give you works only if the possible values of the random variable are equally spaced on the number line. This definition can be modified for other situations, but we won't need to do this.

1. On a horizontal number line, mark all possible values of X. For the binomial, these are 0, 1, 2, $\ldots n$.

2. Determine the value of δ (Greek lower case delta) for the random variable of interest. The number δ is the distance between any two consecutive values of the random variable. For the binomial, $\delta = 1$, which makes life much simpler.

3. Above each x draw a rectangle, with its center at x, its base equal to δ and its height equal to $P(X = x)/\delta$. Of course, in the current case, $\delta = 1$, so the height of each rectangle equals the probability of its center value.

For a probability histogram (PH) the area of a rectangle equals the probability of its center value, b/c

$$\text{Area } = \text{ Base } \times \text{ Height } = \delta \times \frac{P(X = x)}{\delta} = P(X = x).$$

A PH allows us to 'see' a probability distribution. For example, for our pictures we can see that the binomial is symmetric for $p = 0.50$ and not symmetric for $p \neq 0.50$. This is not an accident of

the four pictures I chose to present; indeed, the binomial is symmetric if, and only if, $p = 0.50$. By the way, I am making the tacit assumption that $0 < p < 1$; that is, that p is neither 1 nor 0. Trials for which successes are certain or impossible are not of interest to us. (The idea is that if $p = 1$, then $P(X = n) = 1$ and its PH will consist of only one rectangle. Being just one rectangle, it is symmetric.)

Here are some other facts about the binomial illustrated by our pictures.

1. The PH for a binomial always has exactly one peak. The peak can be one or two rectangles wide, but never wider.

2. If np is an integer, then there is a one-rectangle wide peak located above np.

3. If np is not an integer, then the peak will occur either at the integer immediately below or above np; or, in some cases, at both of these integers.

4. If you move away from the peak in either direction, the heights of the rectangles become shorter. If the peak occurs at either 0 or n this fact is true in the one direction away from the peak.

It turns out to be very important to have ways to measure the center and spread of a PH. The center is measured by the center of gravity, defined as follows.

Look at a PH. Imagine that the rectangles are made of a material of uniform mass and that the number line has no mass. Next, imagine placing a fulcrum that supports the number line. Now look at the PH for the Bin(100,0.50). If the fulcrum is placed at 50, then the PH will balance. B/c of this, 50 is called the center of gravity of the Bin(100,0.50) PH. Similarly, for any binomial PH, there is a point along the number line that will make the PH balance; this point is the center of gravity.

The center of gravity is also called the **mean** of the PH. The mean is denoted by the Greek letter mu: μ. It is an algebraic fact that for the binomial, $\mu = np$.

It is difficult to explain our measure of spread, so I won't. There are two related measures of spread for a PH: the variance and the standard deviation. The variance is denoted by σ^2 and the standard deviation by σ. As you have probably noticed the variance is simply the square of the standard deviation. Thus, we don't really need both of these measures, yet in some settings it is more convenient to think of the variance and in others it is more convenient to focus on the standard deviation. In any event, for the binomial,

$$\sigma^2 = npq \text{ and } \sigma = \sqrt{npq}.$$

Note the following very important fact. Whereas the computation of exact binomial probabilities can be very difficult—especially for large values of n—the formulas for the mean and variance of the binomial are easy to evaluate.

The mean and standard deviation are keys to the approximate method that we will learn in the next section.

2.2 The Family of Normal Curves

Remember π, the famous number from math which is the area of a circle with radius equal to 1. Another famous number from math is e, which is the limit as n goes to infinity of $(1 + 1/n)^n$. As decimals, $\pi = 3.1416$ and $e = 2.7813$, both approximations. If you want to learn more about π or e, go to Wikipedia.

Let μ denote any real number, positive, zero or negative. Let σ denote any positive real number. In order to avoid really small type, when t is complicated we write $e^t = \exp(t)$. Consider the following function.

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp(\frac{-(x-\mu)^2}{2\sigma^2}), \text{ for all real numbers } x \tag{2.2}$$

The graph of the function f is called the **normal curve** with parameters μ and σ; it is pictured in Figure 2.5. By allowing μ and σ to vary, we generate the family of normal curves. We use the notation N(μ, σ) to designate the normal curve with parameters μ and σ. Here is a list of important properties of normal curves.

1. The total area under a normal curve is one.

2. A normal curve is symmetric about the number μ. Clearly, μ is the center of gravity of the curve, so we call it the mean of the normal curve.

3. It is possible to talk about the spread in a normal curve just as we talked about the spread in a PH for the binomial. In fact, one can define the standard deviation as a measure of spread for a curve and if one does, then the standard deviation for a normal curve equals its σ.

4. You can now see why we use the symbols μ and σ for the parameters of a normal curve.

5. A normal curve has points of inflection at $\mu + \sigma$ and $\mu - \sigma$. (If you don't know what a point of inflection is, here goes: it is a point where the curve changes from 'curving downward' to 'curving upward.' I only mention this b/c: If you see a picture of a normal curve you can immediately see μ. You can also see σ as the distance between μ and either point of inflection.)

Statisticians often want to calculate areas under a normal curve. (We will see one reason why in the next section.) Fortunately, there exists a website that will do this for us. It is:

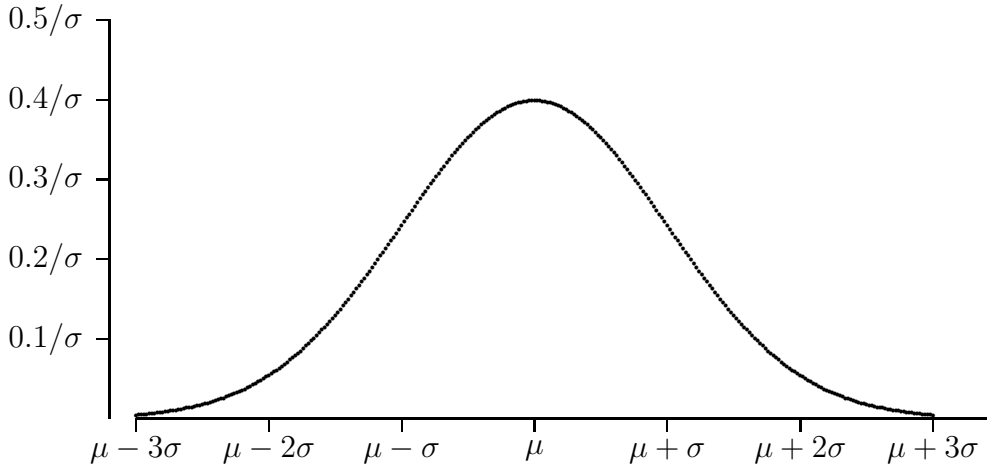
`http://davidmlane.com/hyperstat/z_table.html`

Our course webpage contains a link to this site:

- Click on 'Calculators for Various Statistical Problems.'
- Click on 'Normal Curve Area Calculator.'

Below are some examples of using this site. You should check to make sure you can do this and you will be asked to do similar things on homework.

Figure 2.5: The Normal Curve with Parameters μ and σ; i.e. the N(μ, σ) Curve.



1. **Problem:** I want to find the area under the N(100,15) curve between 95 and 118.

   Solution: Go to the website and enter 100 for 'Mean;' enter 15 for 'Sd;' choose 'Between;' enter 95 in the left box; and enter 118 in the right box. The answer appears below: 0.5155.

2. **Problem:** I want to find the area under the N(100,15) curve to the right of 95.

   Solution: Go to the website and enter 100 for 'Mean;' enter 15 for 'Sd;' choose 'Above;' and enter 95 in the box. The answer appears below: 0.6306.

3. **Problem:** I want to find the area under the N(100,15) curve to the left of 92.

   Solution: Go to the website and enter 100 for 'Mean;' enter 15 for 'Sd;' choose 'Below;' and enter 92 in the box. The answer appears below: 0.2969.

2.2.1 Using a Normal Curve to Obtain Approximate Binomial Probabilities.

Suppose that $X \sim$ Bin(100,0.50). I want to calculate $P(X \geq 55)$. With the help of my computer, I know that this probability equals 0.1841. We will now approximate this probability using a normal curve. First, note that for this binomial $\mu = np = 100(0.50) = 50$ and $\sigma = \sqrt{npq} = \sqrt{100(0.50)(0.50)} = 5$. We use N(50,5) to approximate the binomial; i.e. we pair the binomial with the normal curve with the same mean and standard deviation.

Look at Figure 2.1, the PH for the Bin(100,0.50). The probability that we want is the area of the rectangle centered at 55 plus the area of all rectangles to the right of it. The rectangle centered at 55 actually begins at 54.5; thus, we want the sum of all the areas beginning at 54.5 and going to the right. This picture-based-conversion of 55 to 54.5 is called a **continuity correction** and it greatly improves the accuracy of the approximation. We proceed as follows:

$$P(X \geq 55) = P(X \geq 54.5); \text{ this is the continuity correction.}$$

We now approximate this probability by computing the area under the N(50,5) to the right of 54.5. With the help of the website we obtain 0.1841 (verify this). To four digits of precision, our approximation is perfect!

To summarize, in order to calculate probabilities for $X \sim$ Bin(n, p) we use the continuity correction and the normal curve with $\mu = np$ and $\sigma = \sqrt{npq}$.

The tricky part is the continuity correction. It is easy if you just visualize the rectangles of interest. Below are some examples. In these examples, I don't tell you n and p. They would be needed, of course, to obtain answers, but they are not needed for the continuity correction.

1. Suppose we want $P(X = 43)$. This is one rectangle, namely the one centered at 43. The boundaries of this rectangle are 42.5 and 43.5; thus, $P(X = 43) = P(42.5 \leq X \leq 43.5)$. As a result, our approximation is the area under the appropriate normal curve between 42.5 and 43.5.

2. Suppose we want $P(37 \leq X \leq 63)$. This is many rectangles, starting at 36.5 and extending up to 63.5. As a result, our approximation is the area under the appropriate normal curve between 36.5 and 63.5.

3. Suppose we want $P(37 < X < 63)$. This is many rectangles, starting at 37.5 and extending up to 62.5. As a result, our approximation is the area under the appropriate normal curve between 37.5 and 62.5.

Note that the above three examples do **not** provide an exhaustive list of possible questions that could arise. It is better to *understand* how to do these than to memorize how to do them.

2.2.2 Is the Approximation Any Good?

Approximations are tricky. They don't claim to be exact, so saying that an approximation is not exact—i.e. is *wrong*—is not the point. The idea is that if the approximate answer is *close* to the exact answer, then it is a good approximation and if it is not close to the exact answer, then it is a bad approximation. The difficulty, of course, is deciding what it means to be close. Unfortunately, when we use approximations based on the normal curve, there is no 'pretty certain interval' like we had for computer simulation in Chapter 1.

The current situation is analogous to the following. Your Mom calls (emails?) you and says, "Tell me about your Stats teacher." You answer, "Well he looks approximately like Brad Pitt." Is this a good approximation? Well, before you shout "No!" consider the following idea. Suppose that Mr. Pitt and I are standing next to each other and you are standing, say, 1000 yards from us. At that distance it will be difficult to tell us apart, so, in that sense, the approximation is good. Granted, at 10 yards the approximation is absolutely horrible. Here is the point. Distance is an important factor in determining the quality of the approximation. I maintain: the greater the distance the better the approximation. Many textbooks advocate what I call the *magic threshold approach*. To this way of thinking, there is a magic number, call it T yards. At a distance of less than T yards I am a bad approximation to Mr. Pitt and at a distance of more than T yards I am a

good approximation to him. Books like to claim there are magic thresholds, but that is simply not true.

I will try to discuss this issue in an intellectually honest manner. Look again at Figure 2.4, the PH for the Bin(50,0.1) distribution. This picture is strikingly asymetrical, so one suspects that the normal curve will not provide a good approximation. This observation directs us towards the situation in which the normal curve might provide a bad approximation, namely when p is close to 0 (or, by symmetry, close to 1). But for a fixed value of p, even one close to 0 or 1, the normal curve approximation improves for n large. A common magic threshold for this problem is the following:

If both $np \geq 15$ and $nq \geq 15$ then the normal approximation to the binomial is good.

In the next section we will see that there is a website that calculates exact binomial probabilities, so the above magic threshold does not cause any practical problems.

2.3 A Website and Standardizing

There is website that calculates binomial probabilities. Its address is

`http://stattrek.com/Tables/Binomial.aspx#binomialnormal`

It is linked to our webpage, reached by first clicking on 'Calculators for Various Statistical Problems' and then 'Binomial Probability Calculator.' Below I illustrate how to use it. You should verify these results.

To use the site, you must enter: p in the box 'Probability of success on a single trial;' n in the box 'Number of trials;' and your value of interest for the number of successes (more on this below) in the box 'Number of successes (x).' As output, the site gives you the values of:

- $P(X = x), P(X < x), P(X \leq x), P(X > x)$ and $P(X \geq x)$.

Here is an example. If I enter 0.50, 100 and 55, I get:

- $P(X = 55) = 0.0485, P(X < 55) = 0.8159, P(X \leq 55) = 0.8644, P(X > 55) = 0.1356$ and our previously found $P(X \geq 55) = 0.1841$.

Depending on our purpose, we might need to be algebraically clever to obtain our answer. For example, suppose that for $X \sim$ Bin(100,0.50) we want to determine $P(43 \leq X \leq 55)$. Our website does not give us these 'between probabilities,' so we need to be clever. Write

$$P(43 \leq X \leq 55) = P(X \leq 55) - P(X \leq 42) = 0.8644 - P(X \leq 42),$$

from our earlier output for $x = 55$. To finish our quest, we need to enter the website with 0.5, 100 and 42. The result is $P(X \leq 42) = 0.0666$. Thus, our final answer is

$$P(43 \leq X \leq 55) = 0.8644 - 0.0666 = 0.7978.$$

2.3.1 Can We Trust This Website?

Buried in the description of the binomial calculator website is the following passage.

> Note: When the number of trials is greater than 20,000, the Binomial Calculator uses a normal distribution to estimate the cumulative binomial probability. In most cases, this yields very good results.

I am skeptical about this. As stated earlier, the computer software package that I use—which has been on the market for over 35 years—does not work for $n \geq 1023$ if $p = 0.5$. Thus, I find it hard to believe that the website works at n =20,000. Let's investigate this.

Let's take $n = 16000$ and $p = 0.5$. This gives $\mu = np = 8000$ and $\sigma = \sqrt{npq} = 63.246$. Suppose I want to find $P(X \leq 8050)$. According to the website, this probability is 0.7877 and the normal curve approximation (details not shown) is also 0.7877. I am impressed. For theoretical math reasons, I trust the normal curve approximation and am quite impressed that the exact answer appears to be, well, correct. Thus, it appears that whoever programmed the website was careful about it.

To summarize, it seems to me that you can trust this website provided n is 20,000 or smaller. As I will show you below, **do not use it for larger values of** n.

As mentioned earlier, the normal curve approximation should not be trusted if $np < 15$. Let's look at some examples.

1. Let n =20,000 and $p = 0.00005$ (one in 20,000). Then $\mu = 1$ and $\sigma = 1.0000$. I want to know $P(X = 0)$. The exact answer is $q^{20000} = 0.3679$. The website gives the exact answer 0.3679, which is correct. The normal curve approximation—which should not be used b/c $np = 1 < 15$— gives 0.3085, which, in my opinion, is a bad approximation.

2. Let n =25,000 and $p = 0.00004$ (one in 25,000). Then $\mu = 1$ and $\sigma = 1.0000$. I want to know $P(X = 0)$. The exact answer is $q^{25000} = 0.3679$. The website gives 0.2417 and the normal curve again gives 0.3085. Thus, we can see that the website's writer is less than honest. This is not a good approximation. The website should tell the user to beware for np or nq small. Frankly, I don't know how the website obtained 0.2417! It seems to be using some bizarre continuity correction.

In Chapter 4 we will learn a way to obtain good approximations for the binomial when n is large and either np or nq is small.

For our work in Chapter 3 and later in these notes, I need to tell you about standardizing a random variable. Let X be any random variable with mean μ and standard deviation σ. Then the **standardized version** of X is denoted by Z and is given by the equation:

$$Z = \frac{X - \mu}{\sigma}.$$

Before I further discuss standardizing, let's do a simple example.

Suppose that $X \sim$ Bin(3,0.25). You can verify the following facts about X.

1. Its mean is $\mu = np = 3(0.25) = 0.75$ and its standard deviation is $\sigma = \sqrt{3(0.25)(0.75)} = 0.75$.

2. Its probability distribution is:

   $$P(X = 0) = 0.4219; P(X = 1) = 0.4219; P(X = 2) = 0.1406; \text{ and } P(X = 3) = 0.0156.$$

3. The formula for Z is
   $$Z = \frac{X - 0.75}{0.75}.$$
   Thus, the possible values of Z are: -1, $1/3$, $5/3$ and 3. Thus, the probability distribution for Z is:

   $$P(Z = -1.00) = 0.4219; P(Z = 1/3) = 0.4219; P(Z = 5/3) = 0.1406;$$

   $$\text{and } P(Z = 3) = 0.0156.$$

Earlier in this chapter I argued that if $X \sim$ Bin(n, p) then probabilities for X can be approximated by using a normal curve with $\mu = np$ and $\sigma = \sqrt{npq}$. (I also discussed situations in which this approximation is bad.) This result can be stated in terms of the standardized version of X too. Namely, if $X \sim$ Bin(n, p), then probabilities for

$$Z = \frac{X - \mu}{\sigma} = \frac{X - np}{\sqrt{npq}}$$

can be approximated by using the N(0,1) curve, called the **standard normal curve**. The situations when this latter approximation is good or bad coincide exactly with the conditions for the approximation without standardizing.

2.4 Finite Populations

A finite population is a well-defined collection of individuals. Here are some examples:

- All students registered in this course.
- All persons who are currently registered to vote in Wisconsin.
- All persons who voted in Wisconsin in the 2008 presidential election.

Associated with each individual is a **response**. In this section we restrict attention to responses that have two possible values; called **dichotomous responses**. As earlier, one of the values is called a success (S or 1) and the other a failure (F or 0).

It is convenient to visualize a finite population as a box of cards. Each member of the population has a card in the box, called the **population box**, and on the member's card is its value of the response, 1 or 0. The total number of cards in the box marked '1' is denoted by s (for success) and

the total number of cards marked '0' is denoted by f (for failure). The total number of cards in the box is denoted by $N = s + f$. Also, let $p = s/N$ denote the proportion of the cards in the box marked '1' and $q = f/N$ denote the proportion of the cards in the box marked '0.'

For example, one could have: $s = 60$ and $f = 40$, giving $N = 100$, $p = 0.60$ and $q = 0.40$. Clearly, there is a great deal of redundancy in these five numbers; statisticians prefer to focus on N and p. Knowledge of this pair allows one to determine the other three numbers. We refer to a population box as **Box(N,p)** to denote a box with N cards, of which $N \times p$ cards are marked '1.'

Consider the CM: Select one card at random from Box(N, p). After operating this CM, place the selected card back into the population box. Repeat this process n times. This operation is referred to as **selecting n cards at random with replacement**. Viewing each selection as a trial, we can see that we have BT:

1. Each trial results in one of two possible outcomes, denoted success (S) or failure (F).
2. The probability of S remains constant from trial-to-trial.
3. The trials are independent.

Thus, everything we have learned (the binomial sampling distribution) or will learn about BT is also true when one selects n cards at random with replacement from Box(N, p). Below is a two-part example to solidify these ideas.

1. **Problem:** In the 2008 presidential election in Wisconsin, Barack Obama received 1,677,211 votes and John McCain received 1,262,393 votes. In this example, I will ignore votes cast for any other candidates. (Eat your heart out, Ralph Nader.) The finite population size is N = 1,677,211 + 1,262,393 = 2,939,604. I will designate a vote for Obama as a success, giving $p = 0.571$ and $q = 0.429$.

   Imagine a lazy pollster named Larry. Larry plans to select $n = 5$ persons at random with replacement from the population. He counts the number of successes in his sample and calls it X. He decides that if $X \geq 3$, then he will declare Obama to be the winner. If $X \leq 2$, then he will declare McCain the winner.

   What is the probability that Larry will correctly predict the winner?

   Solution: We could use the website, but I will take this opportunity to give you some practice calculating by hand.

$$P(X \geq 3) = P(X = 3) + P(X = 4) + P(X = 5) =$$

$$\frac{5!}{3!2!}(0.571)^3(0.429)^2 + \frac{5!}{4!1!}(0.571)^4(0.429) + \frac{5!}{5!0!}(0.571)^5 =$$

$$0.3426 + 0.2280 + 0.0607 = 0.6313.$$

2. **Problem:** Refer to the previous problem. Larry decides that the answer we obtained, 0.6313, is too small. So he repeats the above with n =601 instead of $n = 5$. He will declare Obama the winner if $X \geq 301$.

What is the probability that Larry will correctly predict the winner?

Solution: Using the website, the answer is 0.99977. For practice, I will obtain the answer with the normal curve approximation. First, $\mu = 601(0.571) = 343.17$ and $\sigma = \sqrt{343.17(0.429)} = 12.13$. Using the website, the normal curve approximation is 0.99975

We see above that if we sample at random with replacement from a finite population then we get BTs. But suppose that we sample at random without replacement, which, of course, seems more sensible. In this new case, we say that we have a **(simple) random sample** from the finite population. Another way to say this is the following. A sample of size n from a finite population of size N is called a (simple) random sample if, and only if, every subset of size n is equally likely to be selected.

Here is a common error. Some people believe that you have a (simple) random sample if every member of the population has the same probability of being in the sample. This is wrong. Here is a simple example why. Suppose that a population consists of $N = 10$ members and we want a sample of size $n = 2$. For convenience, label the members $a_1, a_2, \ldots, a_{10}$. A **systematic random sample** is obtained as follows. Select one of the members $a_1, a_2, \ldots, a_5$ at random. Denoted the selected member by a_s. Then let the sample be a_s and a_{s+5}. Each member of the population has a 20% chance of being in the sample, but most possible subsets (40 out of 45) are impossible; hence, not a (simple) random sample.

Now there are many situations in practice in which one might prefer a systematic random sample to a (simple) random sample (typically for reasons of ease in sampling). My point is *not* that one is better than the other, simply that they are different. Another popular way of sampling is the **stratified random sample** in which the researcher divides the population into two or more strata, say males and females, and then selects a (simple) random sample from each strata.

The common feature of (simple) random samples, systematic random samples and stratified random samples is that they are **probability samples**. As such, they are particularly popular with scientists, statisticians and probabilists b/c they allow one to compute, in advance, probabilities of what will happen when the sample is selected. There are a number of important ways to sample that are not probability samples; the most important of these are: judgment sampling, convenience sampling and volunteer sampling. There are many examples of the huge biases that can occur with convenience or volunteer sampling, but judgment sampling, provided one has good judgment, can be quite useful.

Sadly, the above topics are beyond the scope of these notes, primarily b/c they are mainly of interest in social science applications. In this course we will focus on (simple) random sampling. (Well, with one exception, clearly stated, much later, of stratified sampling.) As a result I will drop the adjective simple and refer to it as random sampling.

Much of this chapter has been devoted to showing you how to compute probabilities when we have BTs. If, instead, we have a random sample, the formulas for computing probabilities are much messier and, in fact, cannot be used unless we know N exactly; and often researchers don't know N. Here is an incredibly useful fact; it will be illustrated with an example in the lecture notes. Provided $n \leq 0.05N$, the probability distribution of X, the total number of successes in a sample of size n for a random sample, can be well-approximated by the Bin(n, p). In words, sample either way, but when you calculate probabilities for X you may use the binomial.