APPENDIX 5.2: WHY ARE SO MANY DISTRIBUTIONS NORMAL?

This question may have occurred to you while you were reading Chapter 4. The usual explanation for the prevalence of normal distributions is that the central limit theorem captures something essential about nature. The reasoning goes as follows. We know that many statistics are normally distributed when the scores contributing to the statistic are drawn independently from the same distribution. Therefore, it may be that something similar is at work at the level of individual scores. It might be that there are many independent random influences or factors that each contribute something to a given score. That is, any given score combines many independent random factors, just as a sample mean combines many independent random scores. Therefore, a distribution of scores, like a distribution of means, will tend toward normality. In the concrete example that follows, we will think about the *sum* of a number of random factors, rather than the mean, because the central limit theorem applies equally to sums of scores and means; the mean is just a sum divided by *n*.

© iStock.com/megaflopp

Let's think about the weight of a Fender Stratocaster guitar to illustrate this point. The mean weight of a Stratocaster is about 7.87 pounds. There are many components in the guitar, including the body, neck, frets, strings, tuning pegs, pickups, pick guard, and so forth. Each component has a weight, so the weight of each component is drawn from a population of weights. To build a guitar, one selects a body from the population of bodies, a neck from a population of necks, a tuning peg from a population of tuning pegs, and so on. The selection of each component is *independent* of the selection of all other components; that is, the neck you choose does not depend on the pickup you choose. Therefore, just as the scores drawn from a population can be independently selected, the components of the guitar can also be independently selected.

Of course, there will be some variability within each of these populations of components. Therefore, when one assembles these independently selected components to make instruments, the instruments will vary in weight because the components vary in weight. Now, let's suspend reality for a moment and imagine that the distributions corresponding to the components of the guitar have the same mean and standard deviation. If this were the case, then the weight of each guitar could be seen as a sum of scores that are independently drawn from the same distribution. If each component's weight was drawn from the same distribution of weights, then the central limit theorem tells us that the distribution of guitar weights will be normal if there are enough components (i.e., the number of components, *n*, is large).

There is a certain face validity to this idea. It's quite easy to think that the variability in the components is independent. As mentioned, there's no reason to think that the weight of the body is affected by—or dependent on—the weight of the low E string. However, it is not reasonable to imagine that weights of the components are drawn from populations having the same mean and variance. (The mean and standard deviation of the distribution of body weights will be greater than the mean and standard deviation of the distribution of high E string weights.) Therefore, the form of the central limit theorem that we're considering cannot explain the normality of distributions of scores.

Fortunately, there are variants on the central limit theorem that make the general idea more plausible. For example, if the variance of each component distribution is much smaller than the variance of the fully assembled guitar distribution, then (given certain other assumptions that won't be mentioned) the distribution of guitar weights will be normal, even if the component distributions are not identical.

We can think of naturally occurring phenomena in the same way. For example, there are many determinants of height. Clearly, one's genes play a role in height. However, there is not a single gene that determines height but many. Each may make an independent contribution to height. Although height may be determined primarily by genes, factors such as nutrition and physical activity will also make a contribution. In addition, you can imagine that heights measured in the

same individual will vary from time to time depending on levels of fatigue and perhaps motivation to "stand up straight." We can think of all of these contributions to height being characterized as distributions. Therefore, any particular measured height reflects a combination of values selected from these many distributions. The sum of these many factors contributes to the measurement of the height of an individual. The general result of this kind of contribution may be a normal distribution of heights in a given human population.

One has to be a little careful here. This account of normally distributed scores is speculative, to some degree, and has been criticized (Lyon, 2014). We have to keep in mind that not all distributions having many independent contributions are normally distributed. Reaction times are a classic example of this because they tend to be skewed, rather than normal. All things considered, however, variants of the central limit theorem provide a reasonable account of why distributions might be normal.

References

Lyon, A. (2014). Why are normal distributions normal? *British Journal for the Philosophy of Science*, *65*(3), 621–649. doi:10.1093/bjps/axs046

APPENDIX 5.3: THE SAMPLING DISTRIBUTION DEMO

In Chapter 5, we asserted many properties of the distribution of means, variances, and proportions. There are some very useful online tools that help make these concepts concrete. Among the very best is one that can be found here: http://onlinestatbook.com/stat_sim/sampling_dist/index.html.

When you click on this link, you will be presented with an instructions page. After reading the instructions, press Begin and the dialog in Figure 5.A3.1 will appear.

Figure 5.A3.1a shows a distribution from which scores can be drawn. It defaults to a Normal distribution, as indicated in the drop-down list on the right. You can use this drop-down list to change the distribution to a Skewed or Uniform distribution. Or, if you choose Custom, you can define your own distribution by dragging the mouse over the distribution.

The parameters of the distribution of scores are shown to the left. The mean and median are 16 and the standard deviation is 5. Because the distribution shown is normal, the skew and excess kurtosis are both 0.

The purpose of this demo is to illustrate the sampling distribution for various statistics. In Figure 5.A3.1, I've selected two statistics to consider. The first is the biased variance (Variance), shown in Figure 5.A3.1c, which we referred to as s^2_{pop}. The second is the unbiased variance [Var (U)], shown in Figure 5.A3.1d, which we referred to as s^2. In both cases, sample size has been set to N = 5.

There are four ways to draw samples in this demonstration. Choosing Animated randomly selects the number of scores in your sample (in this case, N = 5) and shows each score as a black rectangle (see Figure 5.A3.1b). You can also choose to draw 5, 10,000, or 100,000 samples.

Figures 5.A3.1c and 5.A3.1d show the sampling distributions of the statistics you've chosen to examine. To the left of each of these, the parameters of the distribution (mean, median, standard deviation, skew, and kurtosis) as well as the number of samples (Reps) that contribute to the distribution are shown. You will see that the two sampling distributions are based on 1,000,002 samples. To do this, I pressed the 100,000 button 10 times to generate 1,000,000 samples. Then I pressed the Animated button twice to generate the last sample, shown in Figure 5.A3.1b.

The example here was chosen to illustrate the notion of bias. Remember from Chapter 5 that the biased variance is defined this way:

$$s^2_{pop} = \frac{\sum (y-m)^2}{n}.$$

The unbiased variance is defined this way:

$$s^2 = \frac{\sum (y-m)^2}{n-1}.$$

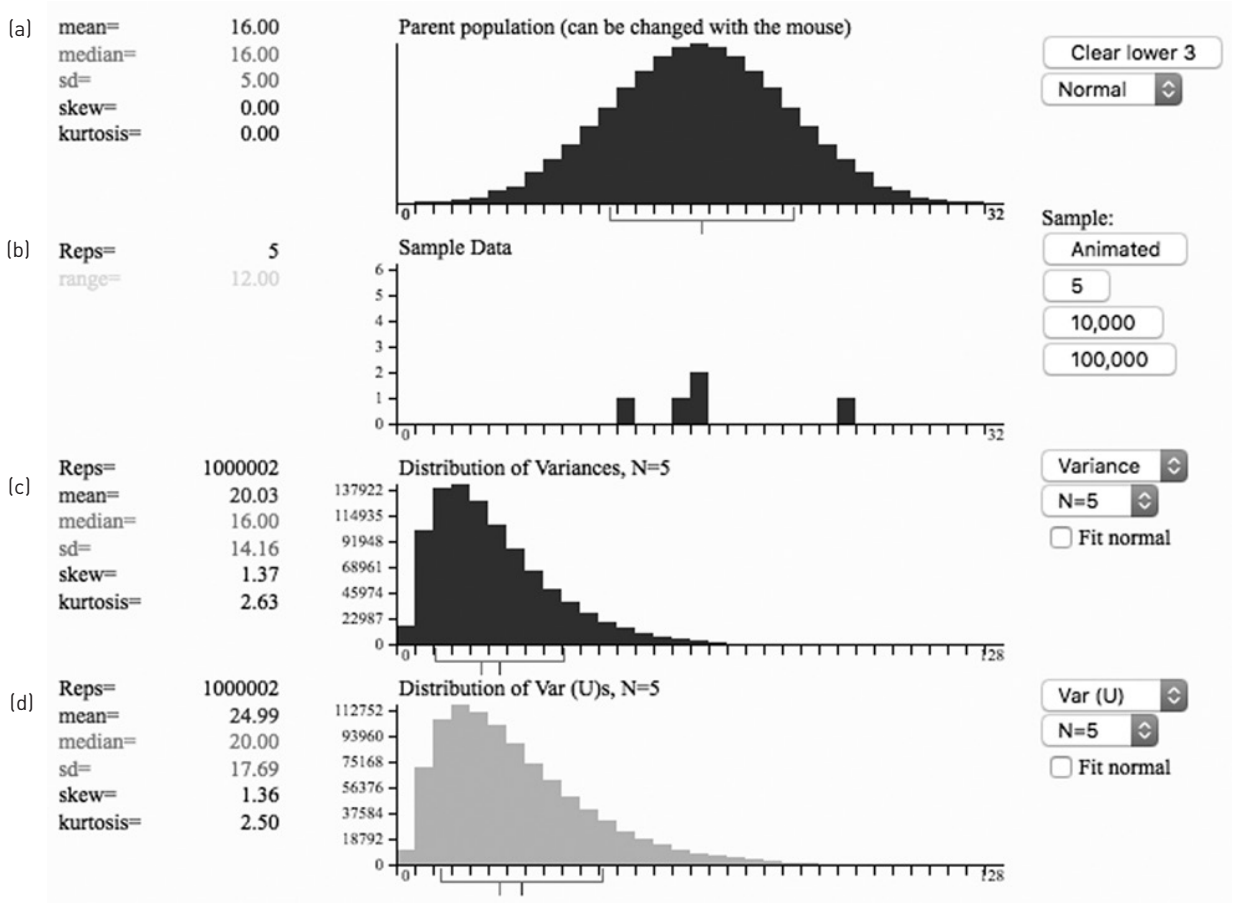
The distribution of s^2_{pop} is shown in 5.A3.1c, and the distribution of s^2 is shown in 5.A3.1d. How do these figures illustrate bias? First, note that the standard deviation of the population is 5, so the variance of the population is 25. We said that a statistic is unbiased if the mean of its sampling distribution (or its expected value) equals the parameter that it estimates. If we look to the left of Figures 5.A3.1c and 5.A3.1d, we see that the *mean* of s^2_{pop} is almost exactly 20, and the mean of s^2 is almost exactly 25. Therefore, s^2_{pop} is biased and s^2 is not.

Finally, in Appendix 5.4 we will see that when a distribution of scores is normal, the distribution of variances (s^2) has a variance of

$$\sigma^2_{s^2} = \frac{2(\sigma^2)^2}{n-1},$$

and a standard error of

FIGURE 5.A3.1 ■ The Sampling Distribution Demo



This extremely useful tool was written by David Lane at Rice University.

$$\sigma_{s^2} = \sqrt{\frac{2(\sigma^2)^2}{n-1}}.$$

This means that the standard error of the distribution of s^2 should be

$$\sigma_{s^2} = \sqrt{\frac{2(5^2)^2}{n-1}} = \sqrt{\frac{2*25^2}{4}} = 17.68.$$

On the left of Figure 5.A3.1d [the sampling distribution of s^2, Var (U)], we see the quantity sd = 17.69 (i.e., $\sigma_{s^2} = 17.69$), which is almost exactly what our equation says it should be. We will see that the most important point about σ_{s^2} is that it decreases as n increases, making estimates of σ^2 from large samples more precise than estimates from small samples.

This outstanding online application offers a tremendous number of ways to explore sampling distributions. I highly recommend that you use it to check many of the points made about sampling distributions in Chapter 5.

APPENDIX 5.4: THE DISTRIBUTION OF SAMPLE VARIANCES

The Variance and Standard Error of the Distribution of Variances

We saw in Chapter 5 that, like all statistics, s^2 has a sampling distribution. The mean of this sampling distribution is σ^2. However, this distribution also has a variance. That is, the last column of numbers in Table 5.3 has a variance, just like any other column of numbers. It just so happens that the numbers in this

column are sample variances. So, the variance in the numbers is just the variance of the sample variances.

This idea of the variance of variances might cause a headache and vertigo, so let's stop and think about it. If you were given the 16 numbers in the last column of Table 5.3 on a midterm and asked to compute the population variance, by now you should have no trouble doing so. The mean of the 16 numbers is 5. For each number in the column, you can compute a squared deviation, and then these can be summed. When you divide the sum of squared deviations by 16 (the number of squared deviations), you have computed the population variance for these numbers. So, these 16 numbers are just numbers and we can compute the population variance for any set of numbers. In fact, do this as an exercise right now.[1] (By the way, this is an example of why having Excel open on your computer can be a big help; use the **VAR.P** function shown in Figure 3.A1.1.)

This exercise shows that there is nothing conceptually difficult about computing the population variance for these numbers, but we might ask a few questions. For example, why are we calling the thing we computed a *population* variance? We are computing a population variance because these numbers represent all possible values of the statistic in question, because the statistic was computed for all possible samples of size n drawn from our population of N scores. So, what is the statistic of interest? The statistic of interest is simply the unbiased sample variance, s^2. This means we've computed a population variance (σ^2) for a sample statistic (s^2), so it seems that we should call the thing we've computed $\sigma^2_{s^2}$. This is the variance of the distribution of sample variances. And that's not as bad as it initially sounded.

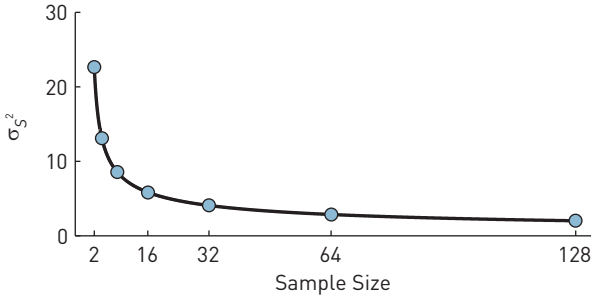
The main point to be made in this section is that $\sigma^2_{s^2}$ decreases as sample size increases, just as σ^2_m decreases as sample size increases. When a population of scores is normal, the variance of s^2 is given by the following equation:

$$\sigma^2_{s^2} = \frac{2(\sigma^2)^2}{n-1}, \qquad (5.A4.1)$$

as mentioned in Appendix 5.3. You absolutely do not have to memorize equation 5.A4.1 and will (probably) never be asked to reproduce it. This equation does not apply to our small population because it is not normal. However, the main point to note in equation 5.A4.1 is that the variance of the sample variance ($\sigma^2_{s^2}$) depends on both the population variance (σ^2) and sample size (n). As sample size increases, $\sigma^2_{s^2}$ decreases.

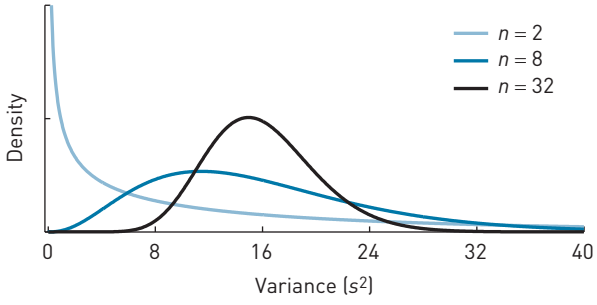
The standard error of s^2 is simply the square root of equation 5.A4.1 and is therefore denoted σ_{s^2}. Figure 5.A4.1 shows how the standard error of the sample variance (σ_{s^2}) changes with sample size. The solid black line plots σ_{s^2} for sample sizes ranging from 2 to 128. The population variance in this example is $\sigma^2 = 16$. The filled circles show a subset of sample sizes (i.e., 2, 4, 8, 16, 32, 64, and 128). As with the sample mean

FIGURE 5.A4.1 ■ Sample Size and σ_{s^2}



The distribution from which the samples were drawn had a mean of $\mu = 10$ and variance of $\sigma^2 = 16$. As sample size increases, the standard error of the sample variance (σ_{s^2}) decreases. This means that as sample size increases, s^2 becomes a more precise estimate of σ^2.

FIGURE 5.A4.2 ■ Sample Size and Shape



The distribution from which the samples were drawn had a mean of $\mu = 10$ and variance of $\sigma^2 = 16$. The light blue line plots the distribution of sample variance for $n = 2$. The dark blue line plots the distribution of sample variance (s^2) for $n = 8$. The black line plots the distribution of sample variance for $n = 32$. As sample size increases, the distribution of sample variances becomes increasingly normal, as we would expect from the central limit theorem. Because s^2 is an unbiased statistic, the mean of each distribution equals $\sigma^2 = 16$, the parameter that s^2 estimates.

1. The answer is 33.

(Figure 5.1), the sample variance s^2 becomes a more precise estimate of the parameter it estimates (σ^2) as sample size increases.

The Shape of the Distribution of Variances

From everything we've covered so far, you might expect the shape of the distribution s^2 to be normal. However, this is generally not the case. Because the sample variance can never be less than 0, the conditions are right for the distribution of s^2 to be positively skewed, which it is. Figure 5.A4.2 shows the sampling distribution of s^2 for sample sizes of 2, 8, and 32. In all cases, the scores were drawn from a normal distribution with mean $\mu = 10$ and variance $\sigma^2 = 16$. The distributions are clearly skewed to the right for sample sizes 2 and 8. However, as sample size increases, the distributions become increasingly normal. If you would like to further explore the distribution of sample variances, you can spend some time with the sampling distribution demo described in Appendix 5.3 (available at sagepub.com/gurnsey).

LEARNING CHECK 1

1. The variance of the distribution of variances increases as sample size increases. [True, False]
2. What is the expected value of the unbiased sample variance?
3. If a normal distribution has a mean of 10 and standard deviation of 2, what is the mean of the distribution of sample variances if sample size = 8?

Answers

1. False. The variance of the distribution of variances decreases as sample size increases.
2. σ^2.
3. Sample size is irrelevant. Because the sample variance is an unbiased estimator, the mean of its sampling distribution will always equal σ^2, which in this case is 4.

Excel Functions Related to the Variance

CHISQ.DIST

The distribution of sample variances is related to the χ^2 distribution. χ is the Greek letter *chi*, which is pronounced like the first syllable in *kay*ak. Therefore, χ^2 is pronounced chi-squared. χ^2 is a *statistic* defined as follows:

$$\chi^2 = \frac{s^2}{\sigma^2}(n-1). \qquad (5.A4.2)$$

The relationship between χ^2 and s^2 is like the relationship between z and m. The sampling distribution of s^2 can be transformed to the χ^2 distribution just as the sampling distribution of m can be transformed to the z-distribution. This means that the proportion of sample variances below any given value of s^2 is the same as the proportion of the χ^2 distribution below the corresponding value of χ^2.

The χ^2 distribution factors out the population variance (σ^2) just as the z-distribution factors out the population mean and variance. Therefore, χ^2 distributions depend only on n–1. When we transform s^2 to χ^2, we can use the **CHISQ.DIST** function to determine the proportion of the distribution of variances below s^2. We use **CHISQ.DIST** as follows:

CHISQ.DIST(χ^2, n–1, cumulative),

which is illustrated in Figure 5.A4.3.

CHISQ.INV

We noted above that the distribution of variances is related to the χ^2 distribution. χ^2 was defined above as

$$\chi^2 = \frac{s^2}{\sigma^2}(n-1).$$

With a little rearrangement, we find that

$$s^2 = \frac{\sigma^2\chi^2}{n-1}. \qquad (5.A4.3)$$

FIGURE 5.A4.3 ■ CHISQ.DIST

	A	B	C
1	Quantities	Values	Formulas
2	n	21	
3	n-1	20	=B2-1
4	σ^2	25	
5	s^2	30	
6	χ^2	24	=B5/B4*B3
7			
8	CHI.DIST (χ^2)	0.0437	=CHISQ.DIST(B6,B3,0)
9	CHI.DIST (χ^2)	0.7576	=CHISQ.DIST(B6,B3,1)

Illustration of the χ^2 distribution $\chi^2 = s^2/\sigma^2(n-1)$. In this illustration, $n = 21$, $\sigma^2 = 25$, and $s^2 = 30$. The proportion of the s^2 distribution below a given value of s^2 is the same as the corresponding proportion of the χ^2 distribution below the corresponding value of χ^2. Therefore, the proportion of the distribution of s^2 below $s^2 = 30$, when $\sigma^2 = 25$ and $n = 21$, is $P(\chi^2) = .7576$.

FIGURE 5.A4.4 ■ CHISQ.INV

	A	B	C
1	Quantities	Values	Formulas
2	n	21	
3	n-1	20	=B2-1
4	σ^2	25	
5	α	0.05	
6	α/2	0.025	=B5/2
7	1-α/2	0.975	=1-B6
8			
9	$\chi^2_{(\alpha/2)}$	9.5908	=CHISQ.INV(B6,B3)
10	$\chi^2_{(1-\alpha/2)}$	34.1696	=CHISQ.INV(B7,B3)
11			
12	$s^2_{(\alpha/2)}$	11.9885	=B4*B9/B3
13	$s^2_{(1-\alpha/2)}$	42.7120	=B4*B10/B3

Using **CHISQ.INV** to determine $\chi^2_{\alpha/2}$ and $\chi^2_{1-\alpha/2}$. These values are transformed to $s^2_{\alpha/2}$ and $s^2_{1-\alpha/2}$ in cells **B12** and **B13**.

With this simple relationship in mind, we can find the two values of s^2 that enclose the central (1−α)100% of the distribution of sample variances, which we'll call $s^2_{\alpha/2}$ and $s^2_{1-\alpha/2}$. First we find the values of χ^2 that enclose the central (1−α)100% of the distribution of χ^2 ($\chi^2_{\alpha/2}$ and $\chi^2_{1-\alpha/2}$) and then convert these values to $s^2_{\alpha/2}$ and $s^2_{1-\alpha/2}$.

In Figure 5.A4.4, we've entered values for n, σ^2, and α in cells **B2**, **B4**, and **B5**, respectively. From these values, we've computed n−1, α/2, and 1−α/2 in cells **B3**, **B6**, and **B7**, respectively. The function **CHISQ.INV** takes two arguments as shown:

CHISQ.INV(P(χ^2), n–1).

The first argument is a proportion, $P(\chi^2)$, which indicates the proportion of the χ^2 distribution below the χ^2 value of interest. The second argument is n−1. In cell **B9**, we've asked **CHISQ.INV** to return the value of χ^2 having (α/2)100% of the χ^2 distribution below it. We call this $\chi^2_{\alpha/2}$. In cell **B10**, we've asked **CHISQ.INV** to return the value of χ^2 having (α/2)100% of the χ^2 distribution above it. We call this $\chi^2_{1-\alpha/2}$. The values returned are $\chi^2_{\alpha/2} = 9.5908$ and $\chi^2_{1-\alpha/2} = 34.1696$. These values are transformed to values of $s^2_{\alpha/2}$ and $s^2_{1-\alpha/2}$ in cells **B12** and **B13** using the formula in equation 5.A4.3. For this example, we see that 95% of the distribution of sample variances, for $\sigma^2 = 20$ and $n = 21$, lies in the interval [11.9885, 42.7120].

APPENDIX 5.5: THE DISTRIBUTION OF SAMPLE PROPORTIONS

Many important questions in psychology rely on proportions. For example, what proportion of children are autistic? What proportion of individuals living in a house with a gun die of a gunshot wound? What proportion of adults have a phobia? In this section, we will see that the distribution of proportions is really a special case of the distribution of means.

In the examples given above, we divided our populations into two groups: those who possess some property ("is autistic," "died of a gunshot wound," "has a phobia") and those who do not ("is not autistic," "did not die of a gunshot wound," "does not have a phobia"). Such variables are called *dichotomous variables*. (Dichotomous means "divided into two parts.") When we select an individual from a population and find that the individual possesses the property of interest, we refer to that outcome as a *success*. If the individual does not possess the property of interest, we refer to that outcome as a *failure*.

The number of individuals in a population that possess the property of interest can be denoted by N_{success}. The number of individuals in a population that do not possess the property of interest can be denoted by N_{failure}. Because one either has the property or not, $N_{\text{success}} + N_{\text{failure}} = N$, which is the number of individuals (scores) in the population. The *proportion* of individuals in a population that possess the property of interest can be calculated by dividing N_{success} by N. We will use the symbol π (the Greek letter *p*) to indicate the proportion of individuals in a population possessing the property of interest:

$$\pi = \frac{N_{\text{success}}}{N}. \qquad (5.A5.1)$$

Proportions as Means

Dichotomous variables can be coded with ones (successes) and zeros (failures). Therefore, each of the N individuals in a population can be assigned a 1 or 0 depending on whether they possess the property of interest. If the variable y is dichotomous, then we can use the following equation to define the proportion of individuals in the population having the property of interest as a mean:

$$\pi = \frac{\sum y}{N}, \tag{5.A5.2}$$

where $\sum y = N_{success}$. If the scores in y represent a sample of n scores drawn from a population, then we can define the sample mean in the same way:

$$p = \frac{\sum y}{n}. \tag{5.A5.3}$$

Therefore, if $y = \{1, 0, 1, 0, 1, 1, 0, 1, 1, 1\}$, then $\sum y = 7$ and $\pi = p = .7$.

The Variance in Populations and Samples of Dichotomous Variables

The variance in a dichotomous population can be defined exactly as any other variance:

$$\sigma^2 = \frac{\sum (y - \pi)^2}{N}. \tag{5.A5.4}$$

The sample variance can be defined as before:

$$s^2 = \frac{\sum (y - p)^2}{n - 1}. \tag{5.A5.5}$$

Let's continue with a small population of 10 scores, $y = \{1, 0, 1, 0, 1, 1, 0, 1, 1, 1\}$, with $\pi = .7$. Now, when we compute the population variance in the usual way, we discover something interesting and useful:

$$\begin{aligned}
\sigma^2 &= \frac{\sum (y - \pi)^2}{N} = \frac{\sum \left(\{1,\ 0,\ 1,\ 0,\ 1,\ 1,\ 0,\ 1,\ 1,\ 1\} - .7\right)^2}{10} \\
&= \frac{\sum \{.3,\ -.7,\ .3,\ -.7,\ .3,\ .3,\ -.7,\ .3,\ .3,\ .3\}^2}{10} \\
&= \frac{\sum \{.09,\ .49,\ .09,\ .49,\ .09,\ .09,\ .49,\ .09,\ .09,\ .09\}}{10} \\
&= \frac{7(.09) + 3(.49)}{10} = \frac{2.1}{10} = .21.
\end{aligned}$$

So, what's so interesting and useful here? Well, the answer is that

$$\sigma^2 = \pi(1 - \pi). \tag{5.A5.6}$$

In this case, $\sigma^2 = \pi(1 - \pi) = (.7)(.3) = .21$. Therefore, the variance of a dichotomous population of scores coded as 1s and 0s is simply the product of the proportion of 1s and the proportion of 0s.

Computing the variance of the sample is just a little more complicated, because it requires a correction factor:

$$s^2 = p(1 - p)\frac{n}{n - 1}. \tag{5.A5.7}$$

This is the same correction factor used in Chapter 3 (equation 3.7c) for the sample variance.

The Sampling Distribution of *p*

At this point, we can put together a number of points that were made in Chapter 5 to understand the sampling distribution of p. First, because there are just two values of the variable, the distribution of scores is very nonnormal. Second, according the central limit theorem, the sampling distribution of the mean (i.e., the sampling distribution of p) will be a normal distribution if sample size is big enough. That is, the distribution of all possible values of the statistic (p), computed from all possible samples of size n, will be a normal distribution if n is big enough. Furthermore, we know that the mean of the sampling distribution of p will be

$$\mu_p = \pi, \tag{5.A5.8}$$

and the variance of the distribution will be

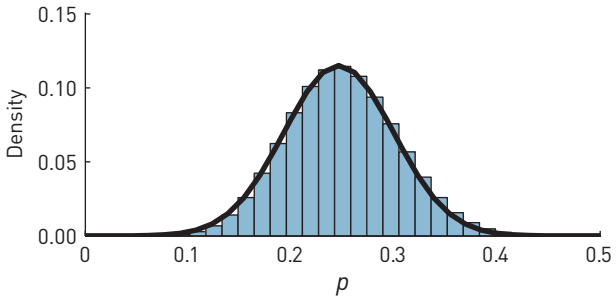
$$\sigma_p^2 = \frac{\sigma^2}{n} = \frac{\pi(1 - \pi)}{n}. \tag{5.A5.9}$$

Therefore, the standard error of the sampling distribution of p is

$$\sigma_p = \sqrt{\sigma_p^2} = \sqrt{\frac{\sigma^2}{n}} = \sqrt{\frac{\pi(1 - \pi)}{n}}. \tag{5.A5.10}$$

Figure 5.A5.1 illustrates the sampling distribution of p for $\pi = .25$ and sample size $n = 64$. Because sample size is 64, there are only 65 possible events (values of p) because $n_{success}$ can only take on integer values from 0 to 64. This means that there can only be 65 events defined as $p = n_{success}/n$. Therefore, the sampling distribution

FIGURE 5.A5.1 ■ The Sampling Distribution of *p*



The bars show the sampling distribution of p for $\pi = .25$ and sample size $n = 64$. The distribution has a mean of $\pi = .25$ and a standard error of $\sigma_p = \sqrt{\pi(1-\pi)/n} = .1919$. The continuous black line shows the best-fitting normal distribution, which also has a mean of $\pi = .25$ and a standard error of $\sigma_p = .1919$.

must be shown as a histogram (see the bars in Figure 5.A5.1).

The solid black line in Figure 5.A5.1 shows the probability density function for a normal distribution having a mean of $\pi = .25$ and standard error of

$$\sigma_p = \sqrt{\pi(1-\pi)/n} = \sqrt{.25(1-.25)/64} = .1919.$$

This normal distribution provides a reasonably good fit to the sampling distribution of p. Therefore, we can use this normal approximation to answer area-under-the-curve questions just as we were able to do with sample means because, after all, *p is* a sample mean.

We need to address one point before we look at an example problem. The sample size required to make sure that the distribution of p is normal depends on π. The closer π is to 0 or 1, the larger the sample needs to be. Therefore, sample size (n) must be large enough so that either $n\pi > 15$, if $\pi < .5$, or $n(1-\pi) > 15$, if $\pi > .5$.

To test whether the normality assumption is met, we take the following steps:

Step 1. Choose π or $1 - \pi$, whichever is smaller. This means we'll choose π if it's less than .5, and we'll choose $1 - \pi$ if π is greater than .5.

Step 2. Divide 15 by the number you've chosen; i.e., $15/\pi$ or $15/(1 - \pi)$.

Step 3. If this quotient is smaller than your sample size, then you can assume that the distribution of p is normal.

For the case in which $\pi = .25$, we would need at least $n = 15/\pi = 15/0.25 = 60$ individuals in the sample to be safe in assuming normality. If π is greater than .5, we would need sample size such that $n > 15/(1 - \pi)$.

Critical Values of $z_{\alpha/2}$

Let's think about a question that is similar to the hypothetical jar of jelly beans that we first met in the early pages of Chapter 1. In that case, the dichotomous variable had values "red" and "not red," and we wondered about the probability that the proportion of red jelly beans in a handful would be in a particular interval. We will ask a related question about psychology students who are male.

> Let's say that 25% of psychology students are male. What two sample proportions enclose the central $(1-\alpha)100\%$ of the distribution of proportions when sample size is $n = 1024$ and $\alpha = .05$?

Step 1. Calculate $z_{\alpha/2}$. $\alpha/2 = .05/2 = .025$; therefore, $z_{\alpha/2} = 1.96$.

Step 2. Compute the standard error of the proportion. We were given that $\pi = .25$ and $n = 1024$, so we know that

$$\sigma_p = \sqrt{\pi(1-\pi)/n} = \sqrt{.25(.75)/1024} = 0.01353.$$

Step 3. Compute the two values of p. We use $\pi \pm z_{\alpha/2}(\sigma_p)$ to determine our two values of p as follows:

$$\pi \pm z_{\alpha/2}(\sigma_p) = .25 \pm 1.96(.01353) = .25 \pm .0265 = [.2235, .2765].$$

Step 4. State the answer. The interval [.2235, .2765] encloses the central 95% of a distribution of proportions when $\pi = .25$ and $n = 1024$.

What does this conclusion mean in practical terms? It means that if we consider all possible random samples of 64 psychology students, then the proportion of male students will be between .2235 and .2765 in 95% of these samples.

A Caveat

You may have noticed that it is impossible to have .2235 ∗1024 = 228.864 people because there are no fractional people. This imprecision means that our example is an

approximation. We will not work through a more precise solution at this point. We can note, however, that the precision of our solutions will improve as sample size increases, and questions involving proportions (e.g., proportion of a sample of voters preferring a particular party) typically involve large samples.

LEARNING CHECK 1

1. State whether the following statements are true or false.
   (a) Blue eyes is a dichotomous variable.
   (b) The sampling distribution of p is always normal.
2. If $\pi = .35$, and $n = 100$, calculate σ_p.
3. A recent report showed that the prevalence of the human papillomavirus (HPV) in North America is 11.3%. What is the probability that, in a random sample of 250 North Americans, 10% or more would have HPV?
4. If the prevalence of autism spectrum disorder (ASD) is 1 in 68 children in the United States, what is the probability that, in a random sample of 144 American children, 1% or more would have ASD?

Answers

1. (a) True. (b) False. It approaches the normal distribution as sample size increases.
2. $\sigma_p = \sqrt{.35(1-.35)/100} = .0477$.
3. $\sigma_p = .02$, $z = (.1 - .113)/.02 = -0.65$, $1 - P(-0.65) = .7419$.
*4. $\sigma_p = .01$, $z = (.01 - .0147)/.01 = -0.47$, $1 - P(-0.47) = .6808$.

NORM.DIST, BINOM.DIST, and HYPGEOM.DIST and the Distribution of Proportions in Excel

Figure 5.A5.2 is set up very much like Figure 5.A1.1, except that we are now considering a distribution of proportions. Figure 5.A5.2 shows how to compute the proportion of a distribution of proportions

- below p_1;
- below p_2;
- between p_1 and p_2; i.e., $P(p_1) - P(p_2)$, assuming $p_2 > p_1$; and
- outside the interval p_1 to p_2; i.e., $1 - [P(p_1) - P(p_2)]$.

The mean of this distribution is $\pi = .25$. That is, we can think of a large population of 1s and 0s, with 25% 1s and 75% 0s. The variance of this distribution is $\sigma^2 = .25*(1 - .25) = .25*.75 = .1875$. This calculation is shown in cell **B3**. To do area-under-the-curve questions, we have to compute the standard error of a proportion (p), which requires knowing the sample size. For this example, the sample size is 64, given in cell **B4**. The standard error of p is $\sqrt{\sigma^2/n} = \sqrt{.1875/64} = .05413$. This calculation is shown in cell **B5**.

FIGURE 5.A5.2 ■ NORM.DIST

	A	B	C
1	**Quantities**	**Values**	**Formulas**
2	π	0.25	
3	σ^2	0.1875	=B2*(1-B2)
4	n	64	
5	σ_p	0.05413	=SQRT(B3/B4)
6			
7	p_1	0.2	
8	p_2	0.3	
9			
10	$P(p_1)$	0.1778	=NORM.DIST(B7,B2,B5,1)
11	$P(p_2)$	0.8222	=NORM.DIST(B8,B2,B5,1)
12	$P(p_2)$-$P(p_1)$	0.6444	=B11-B10
13	1-[$P(p_2)$-$P(p_1)$]	0.3556	=1-B12

A portion of an Excel spreadsheet showing the calculation of approximate values of $P(p_1)$, $P(p_2)$, $P(p_1) - P(p_2)$, and $1 - [P(p_1) - P(p_2)]$ when π and n are known.

All remaining calculations are the same as given for two sample means. We use **NORM.DIST** as follows:

NORM.DIST(p, π, σ_p, cumulative).

Cells **B7** and **B8** show two proportions, p_1 and p_2. Cell **B10** shows the approximate proportion below p_1 = .2, cell **B11** shows the approximate proportion below p_2 = .3, cell **B12** shows the approximate proportion between p_1 and p_2, and cell **B13** shows the approximate proportion outside the interval p_1 to p_2. Note that the calculations in cells **B10** to **B13** are exactly the same as the calculations in these cells when we were considering means m_1 and m_2. The only change is that μ and $\sigma_m = \sigma/\sqrt{n}$ from Figure 5.A1.1 have been replaced with π and $\sigma_p = \sqrt{\pi(1-\pi)/n}$ in Figure 5.A5.2.

It is important to keep in mind that we are using normal distribution approximations to compute probabilities associated with proportions. These approximations are imprecise for two reasons. First, distributions of proportions are not exactly normal, although they do approach normality as sample size increases. The second approximation, evident in Figure 5.A5.2, is that we are dealing with arbitrary proportions (e.g., .2 and .3) even though such proportions may not arise for the sample size (n) that we're working with. In the case of $n = 64$, no number divided by 64 is exactly .2 or exactly .3. Despite the inexactness of the normal approximation that we've reviewed, it is widely used because the approximation improves with sample size, and the imprecision is negligible for very large sample sizes.

There are two exact distributions related to proportions. The *binomial distribution* provides an exact distribution of successes when sampling from a dichotomous population *with replacement*. We use it in Excel as follows:

BINOM.DIST($n_{success}$, n, π, cumulative).

n is the number of scores (0s and 1s) in a sample, $\mathbf{n}_{success}$ is the number of successes (1s) in a sample, and **π** is the proportion of successes (1s) in the population. Figure 5.A5.3 shows how to use the **BINOM.DIST** function. n, $n_{success}$, and π are given in cells **B2**, **B3**, and **B4**, respectively; for completeness, we've computed the proportion of successes in our sample as $p = n_{success}/n$. Note, however, that p is not something we provide to **BINOM.DIST**. When *cumulative* is set to 0 (cell **B7**), **BINOM.DIST** returns the exact probability of $n_{success}$ in a sample of n scores drawn with replacement from a dichotomous population with a specified π. When *cumulative* is set to 1 (cell B8), **BINOM.DIST** returns the probability of $n_{success}$ or fewer in a sample of n scores drawn with replacement from a dichotomous population with a specified π.

FIGURE 5.A5.3 ■ BINOM.DIST

	A	B	C
1	**Quantities**	**Values**	**Formulas**
2	n	64	
3	$n_{success}$	18	
4	π	0.25	
5	p	0.28125	=B3/B2
6			
7	BINOM.DIST (p)	0.0938	=binom.dist(B3,B2,B4,0)
8	BINOM.DIST (P)	0.7682	=binom.dist(B3,B2,B4,1)

We use **BINOM.DIST** to determine the exact probability of $n_{success}$ in a sample of n scores drawn with replacement from a dichotomous population with a specified π. **BINOM.DIST** (**p**) refers to this probability. To compute the proportion (P) of the distribution at or below $n_{success}$, we set *cumulative* to 1. **BINOM.DIST** (**P**) refers to this probability.

The *hypergeometric distribution* provides an exact distribution of successes when sampling from a dichotomous population *without replacement*. We use it in Excel as follows:

HYPGEOM.DIST($n_{success}$, n, πN, N, cumulative).

n is the number of scores (0s and 1s) in a sample and $n_{success}$ is the number of successes (1s) in a sample. N is the number of scores in the population and πN is the number of successes (1s) in the population. π must correspond to some k/N where k is an integer between 0 and N.

Figure 5.A5.4 shows how to use the **HYPGEOM.DIST** function. n and $n_{success}$ are given in cells **B2** and **B3**, respectively; for completeness, we've computed the proportion of successes in our sample as $p = n_{success}/n$ in cell **B4**. π is the proportion of successes in the population. In this example, we set N = 1,000,000 and π = .25. Therefore, the number of successes in the population is πN = 250,000. When *cumulative* is set to 0 (cell **B7**), **HYPGEOM.DIST** returns the exact probability of $n_{success}$ in a sample of n scores drawn without replacement from a dichotomous population of size N with πN successes. When *cumulative* is set to 1 (cell **B8**), **HYPGEOM.DIST** returns the exact probability of $n_{success}$ or fewer in a sample of n scores drawn without replacement from a dichotomous population of size N with πN successes.

As population size increases, the hypergeometric distribution converges on the binomial distribution. That is, they become the same thing. This can be seen when we compare cells **B7** and **B8** in Figure 5.A5.3 with cells **B9** and **B10** in Figure 5.A5.4. Remember, in this case, the

FIGURE 5.A5.4 ■ HYPGEOM.DIST

	A	B	C
1	**Quantities**	**Values**	**Formulas**
2	n	64	
3	$n_{success}$	18	
4	p	0.28125	=B3/B2
5			
6	π	0.25	
7	N	1000000	
8	$N_{sucess} = \pi N$	250000	=B6*B7
9	HYPGEOM.DIST (p)	0.0938	=hypgeom.dist(B3,B2,B8,B7,0)
10	HYPGEOM.DIST (P)	0.7682	=hypgeom.dist(B3,B2,B7,B7,1)

We use **HYPGEOM.DIST** to determine the exact probability of $n_{success}$ in a sample of n scores drawn without replacement from a dichotomous population of size N having πN successes. **HYPGEOM.DIST** (**p**) refers to this probability. To compute the proportion (P) of the distribution at or below $n_{success}$, we set *cumulative* to 1. **HYPGEOM.DIST** (**P**) refers to this probability.

population contains 1,000,000 scores. When the population contains a small number of scores, the binomial and hypergeometric distributions diverge. This is shown in Figure 5.A5.5. In this case, the population contains only $N = 100$ scores, while $\pi = .25$. We can now see this divergence when we compare cells **B7** and **B8** in Figure 5.A5.3 with cells **B9** and **B10** in Figure 5.A5.5.

BINOM.INV in Excel

The **BINOM.INV** function in Excel takes the following arguments:

BINOM.INV(n, π, P($n_{success}$)).

n is the number of scores in the sample, π is the proportion of successes in the population, and $P(n_{success})$ is the proportion of the binomial distribution at or below

FIGURE 5.A5.5 ■ HYPGEOM.DIST

	A	B	C
1	**Quantities**	**Values**	**Formulas**
2	n	64	
3	$n_{success}$	18	
4	p	0.28125	=B3/B2
5			
6	π	0.25	
7	N	100	
8	$N_{sucess} = \pi N$	25	=B6*B7
9	HYPGEOM.DIST (p)	0.1240	=hypgeom.dist(B3,B2,B8,B7,0)
10	HYPGEOM.DIST (P)	0.8867	=hypgeom.dist(B3,B2,B7,B7,1)

Illustration of the hypergeometric distribution when the population size is only $N = 100$.

the value of $n_{success}$ that we'd like to obtain. **BINOM.INV** returns the smallest value of $n_{success}$ for which the *cdf* of the binomial distribution is greater than or equal to $P(n_{success})$. We can use **BINOM.INV** to determine the two values of $n_{success}$ such that approximately (α/2)100% of the binomial distribution lies at or below one value and approximately (α/2)100% of the distribution lies at or above the other. This means that (1–α)100% of the distribution lies between the two numbers. We can refer to these two values of $n_{success}$ as $n_{success(\alpha/2)}$ and $n_{success(1-\alpha/2)}$. We can then convert them to proportions by dividing them by n, the number of scores in the sample.

In Figure 5.A5.6, n, π, and α have been entered in cells **B2** to **B4**. Cell **B5** computes the proportion of the binomial distribution within the interval of interest. We've computed α/2 in cell **B6**. We want to find the value of $n_{success(\alpha/2)}$ such that $P(n_{success}) = \alpha/2$. This is accomplished in cell **B9** using **BINOM.INV**. The exact proportion of this binomial distribution ($n = 64$, $\pi = .25$) below 9 is .02524. Therefore, this is what we mean when we say **BINOM.INV** returns the smallest value of $n_{success}$ for which the *cdf* of the binomial distribution is greater than or equal to $P(n_{success})$. [The exact proportion of this binomial distribution ($n = 64$, $\pi = .25$) below 8 is .01113, which is not greater than $\alpha/2 = .025$.]

In cell **B7**, we've computed 1–α/2. We now want to find the value of $n_{success(1-\alpha/2)}$ such that $P(n_{success}) = 1-\alpha/2$. This value of $n_{success}$, computed in cell **B10**, has approximately (1–α/2)100% of the binomial distribution below it and (α/2)100% of the binomial distribution above it. These two values of $n_{success}$ are divided by sample size (n) in cells **B12** and **B13** to produce $p_{\alpha/2}$ and $p_{1-\alpha/2}$. For this example, approximately 95% of the distribution of proportions lies between .1406 and .3594.

FIGURE 5.A5.6 ■ BINOM.INV

	A	B	C
1	**Quantities**	**Values**	**Formulas**
2	n	64	
3	π	0.25	
4	α	0.05	
5	(1-α)*100%	95	=(1-B4)*100
6	α/2	0.025	=B4/2
7	1-α/2	0.975	=1-B6
8			
9	$n_{success(\alpha/2)}$	9	=BINOM.INV(B2,B3,B6)
10	$n_{success(1-\alpha/2)}$	23	=BINOM.INV(B2,B3,B7)
11			
12	$p_{(\alpha/2)}$	0.1406	=B9/B2
13	$p_{(1-\alpha/2)}$	0.3594	=B10/B2

Using **BINOM.INV** to determine the sample score [$n_{success(\alpha/2)}$] having (α/2)100% below it and the sample score [$n_{success(1-\alpha/2)}$] having (α/2)100% above it. $n_{success(\alpha/2)}$ and $n_{success(1-\alpha/2)}$ are converted to proportions by dividing them by n (sample size).