Chapter 10

Sampling Distributions and the Central Limit Theorem

In the previous chapter we explained the differences between sample, population and sampling distributions and we showed how a sampling distribution can be constructed by repeatedly taking random samples of a given size from a population.

A quick look back at Table 9-8 shows the outcome of the process. To summarize, when all possible samples of *N* = 2 are taken from a population with the values 5, 6, 7, 8 and 9, a total of 25 different samples with 9 different means can be observed. When the sample size is increased to 3, 125 different samples with 13 different means are produced. If we were to take samples of *N* = 4 from this same population of values, we would find that 625 distinct combinations of 4 numbers could be observed. It is quite obvious that constructing a sampling distribution by the method we used in the preceding chapter becomes extremely laborious as soon as sample size and population size are larger than 10. In fact, the number of different samples (S) of any given size (*N*) taken from a population of size (*P*) when we sample with replacement can be expressed by the following equation:

$$S = P^N$$

For populations of 5 units (*P* = 5) like our example, this yields the number of samples that were illustrated in Chapter 9:

For *N* = 2, S = 52 = 25

For N = 3, S = 53 = 125
For N = 4, S = 54 = 625

Imagine constructing a sampling distribution by drawing samples of N = 5 students out of a population of 50 students in an undergraduate communication research methods class. If we sample with replacement, the number of different samples is 50 raised to the power 5, or 50 x 50 x 50 x 50 x 50 or 312,500,000 different samples. Computing the mean, sampling variance and standard error of such a sampling distribution would be a monumental assignment. Furthermore, determining the probabilities associated with each of the various sample means would be an equally enormous task. But we need these probabilities to make a valid statistical generalization from a random sample. We obviously need to use another approach to obtaining these probabilities. Fortunately such an approach exists.

The Central Limit Theorem

The Central Limit Theorem provides us with a shortcut to the information required for constructing a sampling distribution. By applying the Theorem we can obtain the descriptive values for a sampling distribution (usually, the mean and the standard error, which is computed from the sampling variance) and we can also obtain probabilities associated with any of the sample means in the sampling distribution.

The mathematics which prove the Central Limit Theorem are beyond the scope of this book, so we will not discuss them here. Instead we will focus on its major principles. They are summarized below:

If we randomly select all samples of some size N out of a population with some mean M and some variance of ó^2, then

- The mean of the sample means ($\bar{\bar{X}}$) will equal M, the population mean;
- The sampling variance will be σ^2/N here (the population variance divided by N, the sample size). The standard error will equal the square root of the sampling variance;
- The sampling distribution of sample means will more closely approximate the Normal Distribution as N increases.

We'll discuss each of these points separately in the following sections of this chapter. But before we do, let us be sure to emphasize the three assumptions of the Central Limit Theorem: we know what size sample (N) we would like to draw from the population, and, more importantly, that we know the two population parameters M and ó^2.

The Mean of the Sampling Distribution of Means: Parameter Known

According to the Central Limit Theorem, the mean of the sampling distribution of means is equal to the population mean. We have already observed this in the examples given in the previous chapter. Our population, consisting of the values 5, 6, 7, 8 and 9, has a mean of 7. When we took all samples of N = 2 or N = 3 out of this population, the mean of all the resulting sample means ($\bar{\bar{X}}$) in the two sampling distributions were both equal to 7.

Therefore, if we know the parameter mean, we can set the mean of the sampling distribution equal to M. This allows us to avoid two massively difficult steps: (1) calculating sample means for all possible samples that can be drawn from the population and (2) calculating the sampling distribution mean from this mass of sample means.

The Variance of the Sampling Distribution of Means: Parameter Known

According to the Theorem, the variance of the sampling distribution of means equals the population variance divided by N, the sample size. The population variance (ó^2) and the size of the samples (N) drawn from that population have been identified in the preceding chapter as the two key factors which influence the variability of the sample means. As we saw in the examples in that chapter, the larger the variance of the values in the population, the greater the range of values that the sample means can take on. We also saw that the sample size was inversely related to the variability of sample means: the greater the sample size, the narrower the range of sample means. The effect of both factors is thus captured by computing the value of the sampling variance as ó^2/*N*. If we know the variance of the population as well as the sample size, we can determine the sampling variance and the standard error.

This aspect of the theorem can be illustrated by using our running example. As you can see in Table 10-1, the variance of the population equals 2.00.

Applying the Central Limit Theorem to sample sizes of $N = 2$ and $N = 3$ yields the sampling variances and standard errors shown in Table 10-1. For $N = 2$ and $N = 3$, these are exactly the same values for the sampling variance and standard error as were computed from the full set of sample means shown in Table 9-5. Table 10-1 also shows the sampling variance and standard error for a sampling distribution based on a sample size of $N = 4$ drawn from the same population. Had we calculated these values from the set of 625 sample means, we would have obtained exactly the same results for the variance and standard error.

But what do we do when the population parameters are unknown? For example, assume that we are interested in studying the population of newly married couples. Specifically, we are interested in the amount of time they spend talking to each other each week about their relationship. It is highly unlikely that any parameters for this population would be available. As we have already mentioned several times, the absence of known parameters is very common in communication research. How are we to proceed under these conditions? In the absence of known parameters we will have to make do with reliable estimates of these parameters. Such reliable estimates can be obtained when we take random samples of sufficient size from the population.

Suppose that we draw a random sample of $N = 400$ from this population. After measuring the amount of time these newlyweds spend talking to one another about their relationship we observe the mean to be 2 hours per week and the sample standard deviation is 1 hour per week. We will use this information to estimate the mean, the variance and the standard error or the sampling distribution.

The Mean of the Sampling Distribution of Means: Parameter Unknown

Since we have only a single sample mean, we can't compute the mean of the means. But we can make a simple assumption, based on probability, that will allow us to work from the results of this single sample.

We know that the most probable mean found in the sampling distribution is the true population mean (you can see this in Table 9-5 in the previous chapter), and that this mean is at the center

Table 10-1 Sampling Distribution Variances Computed from Population Variance

X_i	$(X_i - M)^2$
5	4
6	1
7	0
8	1
9	4

$$10 = \sum (X_i - M)^2$$

$$\sigma^2 = \frac{10}{5} = 2.00\text{, population variance}$$

	N=2	*N*=3	*N*=4
Sampling Variance =	$\frac{\sigma^2}{N} = \frac{2}{2} = 1.00$	$\frac{2}{3} = .672$	$\frac{2}{4} = .500$
Standard Error =	$\sqrt{1.00} = 1.00$	$\sqrt{.672} = .819$	$\sqrt{.500} = .707$

of the sampling distribution. So if we have only one sample from a population, the assumption that the value of the sample mean is the same as the value of the population mean is more likely to be correct than any other assumption we could make. When we do this, we place the center of the sampling distribution right at the sample mean. That is, we arrange our sampling distribution around the computed value of our sample mean (see Figure 10-1).

It is important to note that the sample mean of 2.0 is the best estimate of the unknown population (or true) mean. But we have to realize that there is also the possibility that the true population mean is somewhat higher or lower than that figure. We can use the sampling distribution to describe how probable it is that the real population mean falls somewhere other than the computed sample mean. We will return to this point once we are able to fully describe the sampling distribution.

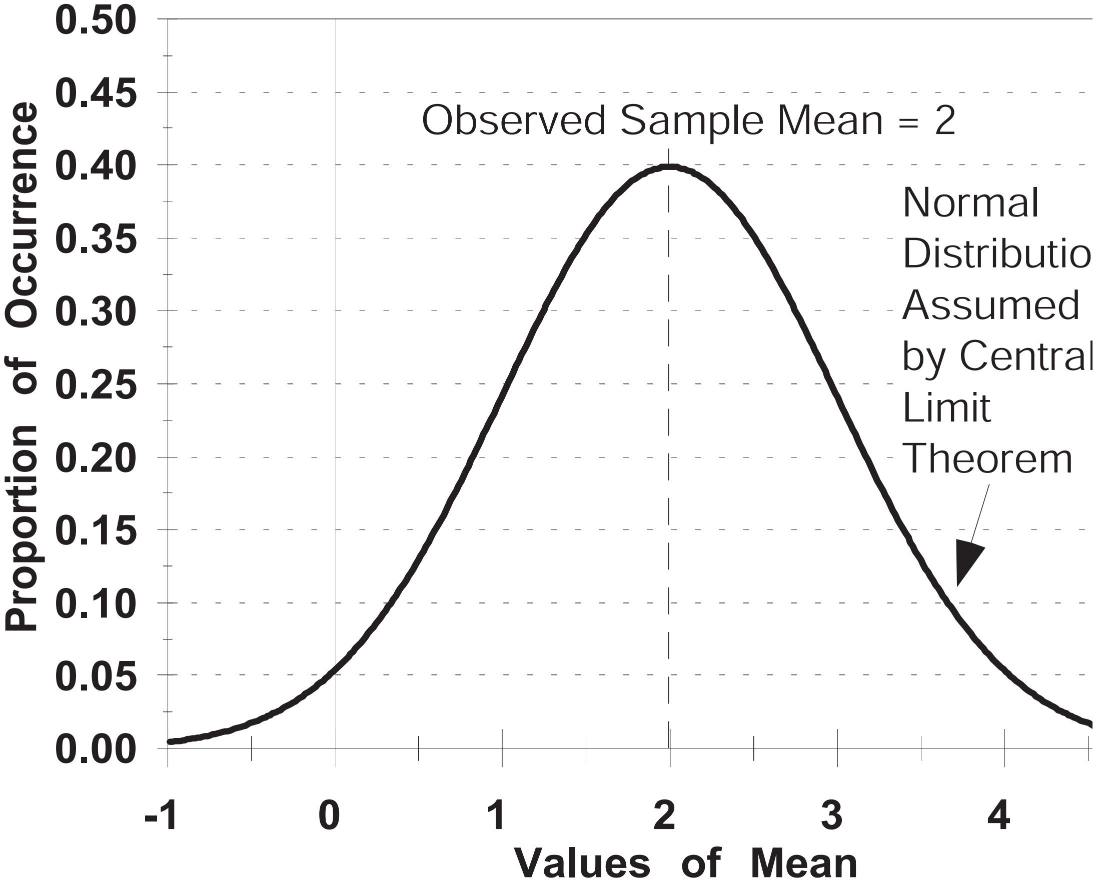
The Variance of the Sampling Distribution of Means: Parameter Unknown

The sampling variance (and hence the standard error) can be estimated from the sample variance if we are willing to make the following assumption. If we are willing to assume about the population variance what we assumed about the population mean, namely, that the most probable value for this unknown parameters is the one which we have computed from our sample, we can again work with the results of our single sample. After we make this assumption, we can estimate the measures of dispersion by the same method as we did in Table 10-1.

Sampling Variance= var / N = 1 / 400 = .0025

Std Error = $\sqrt{\text{var}/N} = \sqrt{1/400} = 0.05hrs$

Std Error = $sd\Big/\sqrt{N} = 1\Big/\sqrt{400} = 1/20 = 0.05hrs$



We now have a complete description of the sampling distribution, constructed from the information provided by a single random sample.

Once the sampling distribution has been identified, either by using known parameters, or by using estimates of these parameters obtained from samples, we can now use this distribution to carry out the next important step: computing the probabilities of the means in the sampling distribution. We need these probabilities to be able to make statements about the likelihood of the truth or falsity of our hypotheses, as we've already mentioned in the previous chapter.

Whether the sampling distribution was derived from known parameters or from estimates matters little in terms of the remainder of the discussion in this chapter, as long as one condition is met: if estimates are used, the sample from which the estimates are derived should be sufficiently large. A violation of this condition does not alter the logic, but requires some computational adjustments. Any statistics text will show you how to carry out these adjustments whenever small samples are encountered.

The Sampling Distribution of Means and Probability

The most important reason why we need to construct sampling distributions is obtaining the probabilities associated with varying amounts of sampling error. Once we have this information, we can use it for statistical generalization and for testing our hypotheses.

Recall that the sampling distribution of means, constructed as we did in the previous chapter, yields a number of means, along with the frequency with which these means were observed. From such a table of means we can determine, for any given amount of sampling error, the likelihood of occurrence of such a sampling error. This likelihood is determined by dividing the frequency with which a given sampling error occurred by the total number of samples.

The Central Limit Theorem allows us to directly compute the mean, variance, and standard error of the sampling distribution without going to the trouble of drawing all possible samples from the population. But we still need a way to describe the shape of the distribution of sample means. Fortunately, we can do this by using the third principle of the Central Limit Theorem. This principle states that the sampling distribution of means will tend to look like a normal distribution, when we select samples which contain relatively large numbers of observations.

What does this mean? Just this: if we were able to compute the thousands or millions of means from every different sample in a population, then collect these means into a classified frequency distribution, we would see that the distribution has the characteristics of the standard normal distribution. But we don't have to do this. We can simply assume a normal distribution (a "bell-shaped curve") will result, as the Central Limit Theorem tells us that we're going to find this distribution if our samples are large enough.

A standard normal distribution like our sampling distribution can be described by the following equation:

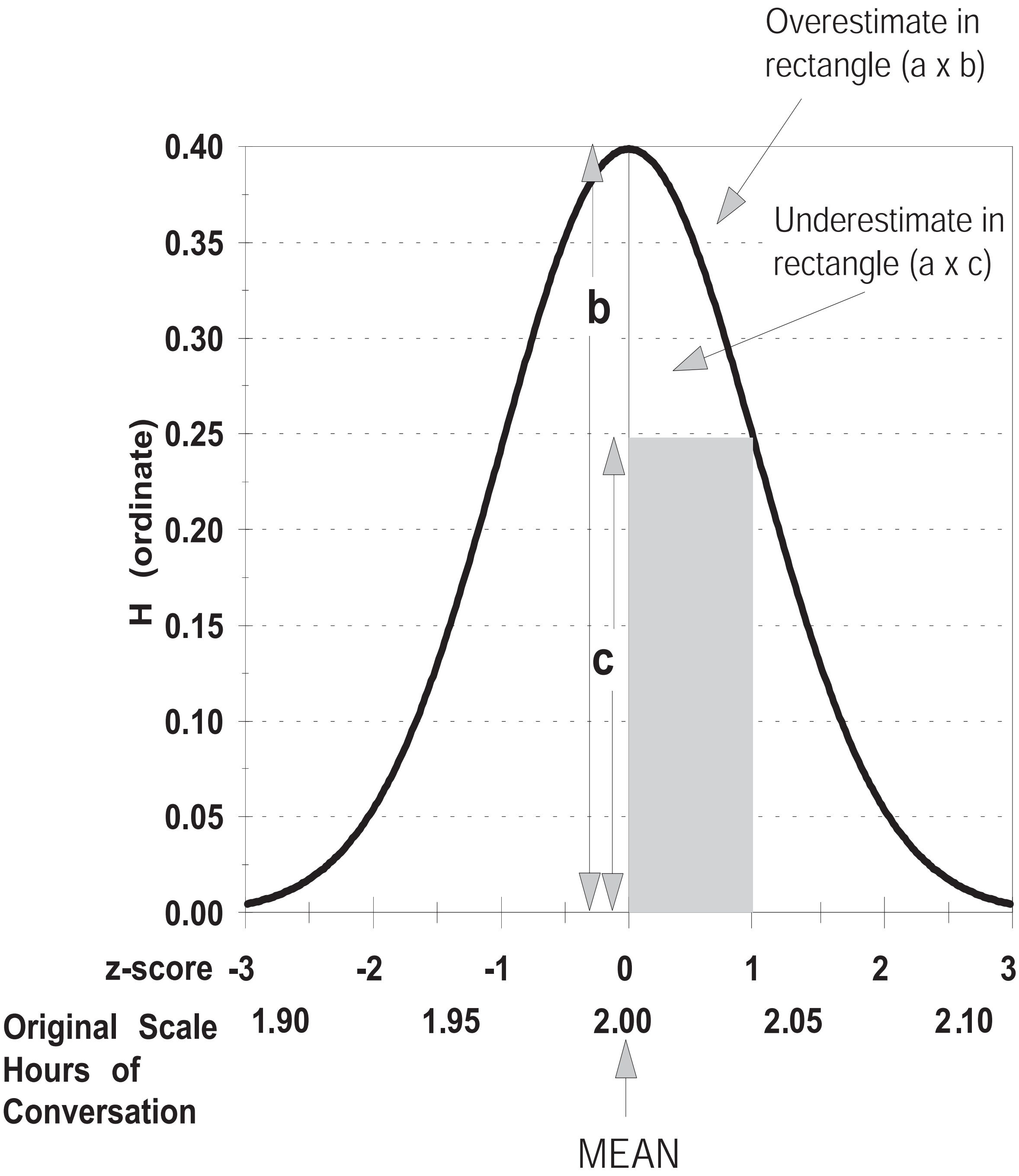
$$H = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

In this equation, the quantity *H* is known as the ordinate (or height) of the curve at any point. This computation involves two mathematical constants (*e*, which equals 2.7183 and *n*, which equals 3.1416) and a quantity called "*z*", which is the standard score as defined and discussed in Chapter 8. A standard score defines any data point (which can be either a data value or a statistic) in terms of the number of standard deviation units (or standard error units) that the point is from the mean. For a given value of *z* (such as the point marked A in Figure 10-2), the height of the normal distribution can then be computed from the above formula.

Why do we need to know this? Because the height of the normal distribution represents the frequency in the sampling distribution with which any mean that is some distance "*z*" from the center of the distribution occurs. We can use the frequency to compute probabilities, as we did in Chapter 9. A mean with a large *z* value is distant from the center of the distribution, and it has a small *H* (i.*e*., frequency). This implies that a mean with a large *z* value has a low probability. The most probable mean (the largest *H* value) appears when $z = 0$. Furthermore, when we compute the height of the curve at two points with different values of *z*, then the two heights will define an area under the curve which will include a certain proportion or percentage of the means in the sampling distribution, which is the same as saying that it defines a certain proportion of the total area under the curve, as can be seen in Figure 10-2.

The Normal Distribution and Areas under the Curve

Determining the magnitude of the proportion of means in a given area under the curve is quite straightforward. It is based, essentially, on the mathematical procedure for determining the area of a rectangle: multiply the length of the rectangle by its width. Remember that the formula for the standard normal distribution allows us to compute *H*, or the height of the curve, for any point *z* on the baseline. Figure 10-2 shows the values for *z*= 0.0 and *z*=1.0, a range in which we might be interested.



When *z*= 0.00, the value of *H* is .3989; when *z*= 1.00, the value of *H* = .2420. The areas of two rectangles can now be computed. The first rectangle is the area defined by the product of these two dimensions: .3989 (or *b*, the height of the curve at *z* =0.0 and the length of the rectangle) and 1.00 (*a*, the difference between *z* = 0.00 and *z* = 1.00, and the width of the rectangle), or (*a* x *b*), which equals .3989. The second rectangle has the dimensions .2420 and 1.00 (or *a* and *c*, the height at *z* = +1.00), giving an area of .2420.

The first rectangle (*a* x *b*) obviously covers more than the area under the curve we are interested in; a certain amount of area in the rectangle is not under the curve. The second rectangle does not cover all the area under the curve: there is a certain amount of area under the curve which is not included in the rectangle. However, it would appear from Figure 10-2 that the overestimate of the first rectangle and the underestimate of the second rectangle might cancel. Taking the average of the two rectangles yields (.3989 + .2420)/2 = .3205, which, as we shall see below, is quite close to the actual figure for the portion of the area under the normal curve between that point where the mean is located and another point one standard deviation from the mean.

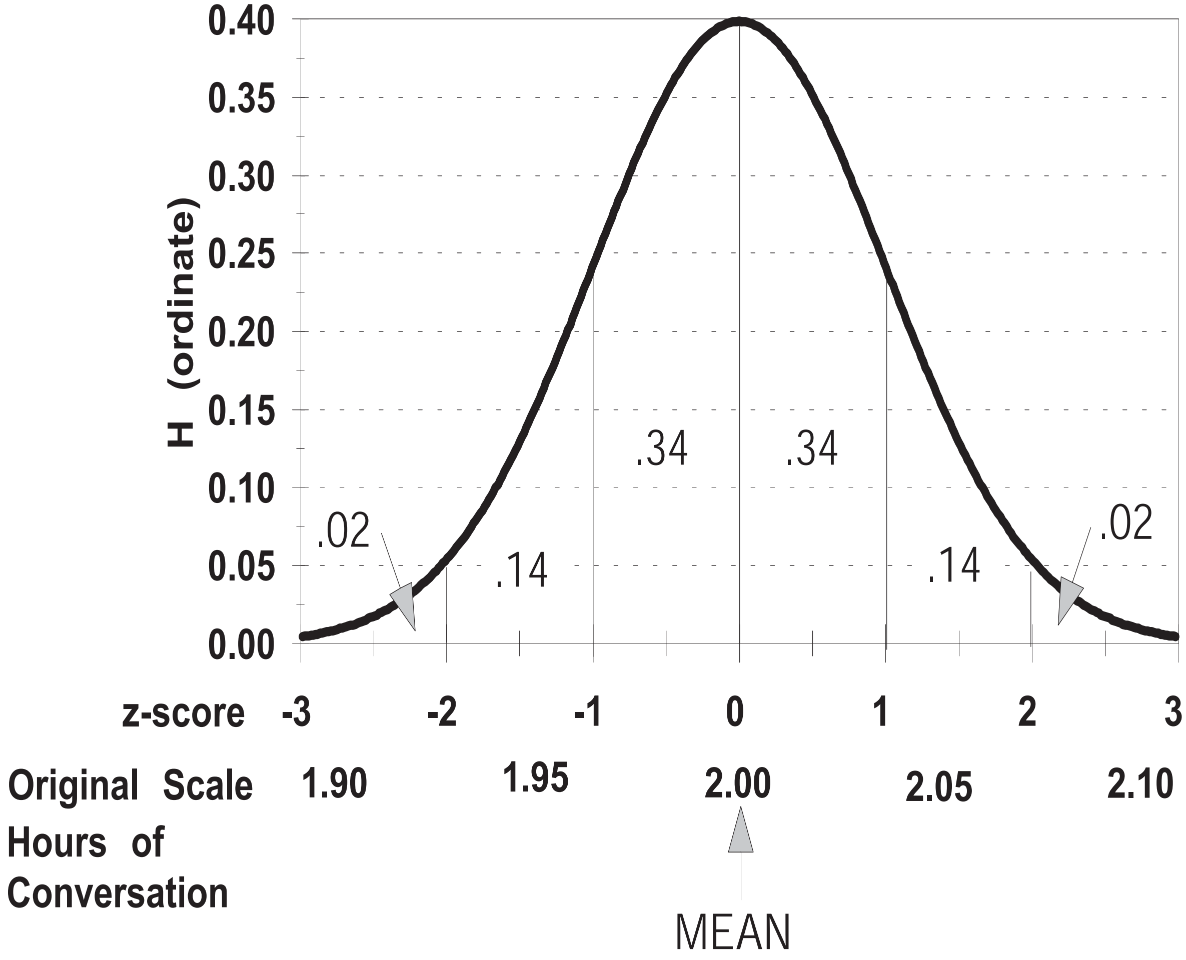
We could also compute the area under the curve between the points *z* = +1.00 and *z* = +2.00. At *z* = +2.00, *H* equals .0540. Following the same averaging procedure as we carried out above, we'd obtain (.2420 + .0540)/2= .1480 as an estimate of the area under the curve between *z* = +1.00 and *z* = +2.00. Since the right half of the distribution contains .50 of the area under the curve, this crude method of determining areas under the curve leaves .50 - (.3205 + .1480)= .0315 of the area beyond *z* = +2.00.

Also, we already know that the normal distribution is symmetrical and so half of the area under the curve falls to the left of the Mean, the other half to the right. Consequently, the same proportions for areas under the curve hold for the left hand side of the distribution.

In Figure 10-2, neither the rectangle *a* x *b* nor the rectangle a x *c*, nor even their average, is a completely satisfactory estimate of the area under the curve. It is obvious that a better approximation of the area under the curve between the mean (*z*=0.0) and *z*=1.00 could be obtained by dividing the baseline in tenths (or even smaller parts) of a standard deviation, then computing the area of a larger number of rectangles. The total area could then be obtained by adding together the areas calculated from the larger number of smaller rectangles. Students familiar with elementary calculus will recognize this as the basic process involved in integration of the normal distribution function between two limits. By making the rectangle bases smaller and smaller,, and summing the resulting areas, we will get more and more accurate estimates of the true area between the two *z*-score limits.

Figure 10-3 illustrates the result of the calculus of integrals. It shows the broad outline of the relationship between standard scores and areas under the curve. In a normal distribution approximately .34 (or 34%) of the area under the curve lies between the mean (0.00) and one standard unit in either direction. An additional 14% will be observed between 1.00 and 2.00 standard units. Note that these proportions are quite similar to the values obtained by our rough method of taking the averages of the areas of the various rectangles. This means that between -1.00 and +1.00 standard units, we will find approximately 68% of all the observations in the normal distribution. Between -2.00 and +2.00 standard deviations roughly 96% of the observations in the distribution will occur. That is, an additional 28% (14% + 14%) are observed between 1 and 2 standard units from the mean. The remaining 4% of cases can be found in each tail (2% in the positive tail, and 2% in the negative tail). These represent the cases that are two or more standard deviations away from the mean of the distribution. Also notice that the .34, .14 and .02 in each half of the distribution sum to the .50 probability that we expect.

Earlier in this chapter we wondered about the probability that the true population mean was either higher or lower than the observed sample mean of 2.00. From the sampling distribution we can find the probability that the true population lies between any two values in the sampling distribution by determining the area under the curve between these two values. For instance, there is a .34 probability that the true population mean lies somewhere between 2.00 hours of conversation and 2.05 hours. The symmetry of the distribution allows us to double that proportion to get the probability that the population mean lies somewhere between 1.95 hours and 2.05 hours. This means that we have better than 2:1 odds that the population mean lies within .05 hour (or one standard error) of the sample mean. The probability is .96 that the population mean lies between 1.90 and 2.10 hours of conversation. This example illustrates the way we can use a sampling distribution to quantify our confidence in the results that we get from a random sample.



The Table of Areas under the Normal Curve

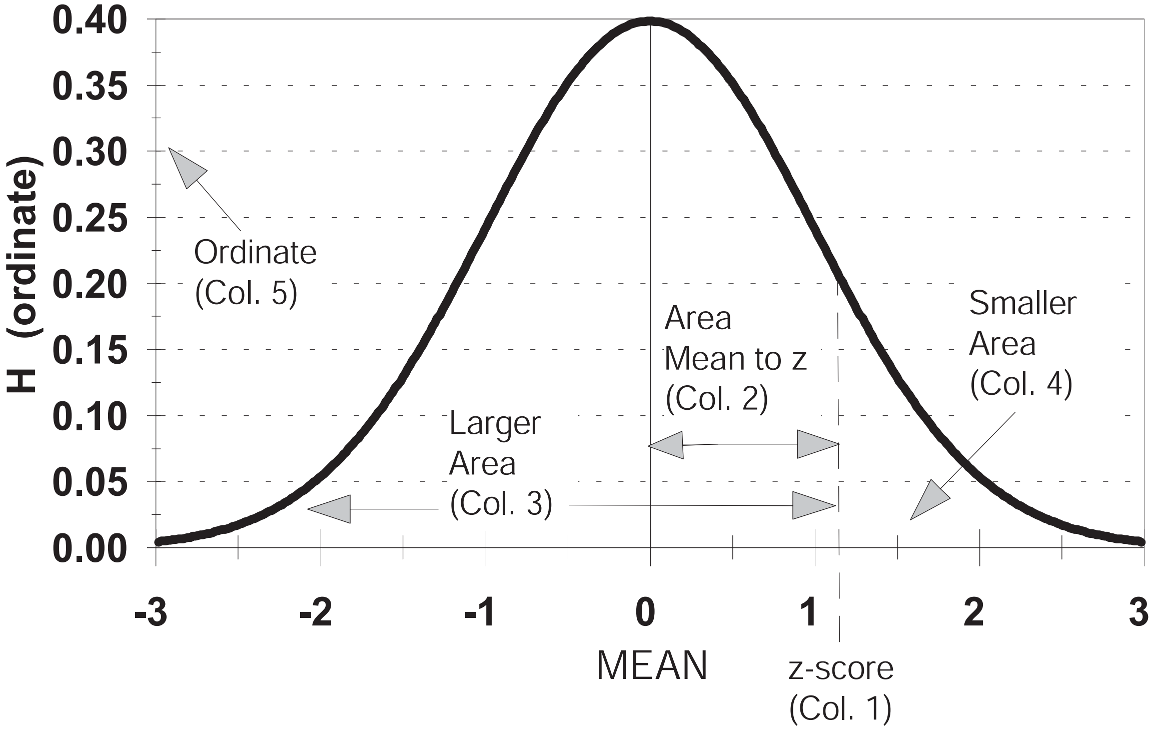
Because of the widespread use of the normal distribution, tables have been prepared which show percentages of area under the normal distribution function very accurately for a wide range of *z*-values. Tabled values for areas under the curve of the normal distribution can be found in Appendix A. We'll reproduce a portion of that table here to illustrate how to interpret it. Use the numbers above the various columns in the Table to refer to areas under the curve.

In row (a) of the table under Figure 10-4, the *z*-score is located exactly on the mean (*z* = 0.0), and therefore the mean-to-*z* area is equal to 0.0. The fact that this *z*-score is located on the mean also implies that the area under the normal curve is divided into two equal parts; thus the "larger" and the "smaller" areas are equal to one another (and also equal to .50). Finally, column 5 gives the height of the curve at that point as .3989.

In row (b) we see a *z*-score of .75, a location three-fourths of a standard unit away from the mean. The area under the curve from the mean to this value of *z* is approximately .27 (or 27%). The larger area is now equal to .50 plus .27, or .77. This means that the smaller portion under the curve is equal to 1.00 - .77 = .23.

Rows (*c*) and (d) illustrate what happens when the *z*-score becomes larger, that is, when we move farther away from the mean of the normal distribution. The area between the mean and the *z*-value becomes larger, the "larger" portion becomes larger, the "smaller" portion becomes smaller, and the height of the curve decreases. For instance, row (d) indicates that when *z* = 2.50, the values in the "larger" portion of the distribution (i.*e*., those with lower *z*-values) constitute 99.38% of all cases in the distribution. Cases that have values that are 2.50 standard units or further from the mean constitute only the remaining .62% of all the cases in the distribution.

Since the normal distribution is symmetrical, the results for negative *z*-scores are the same as those of the positive *z*-values. In order to use the tables for negative *z*-scores, you need only look up the corresponding positive *z*-score. With negative *z*-scores the "smaller area" under the curve will



Column	(1)	(2)	(3)	(4)	(5)
		Area Mean	Larger	Smaller	H
	z-score	to z	Area	Area	(ordinate)
Row					
(a)	0.00	.0000	.5000	.5000	.3989
	. .				
(b)	0.75	.2734	.7734	.2266	.3011
	. .				
(c)	1.65	.4505	.9505	.0495	.1023
	. .				
(d)	2.50	.4938	.9938	.0062	.0175

be located in the left-hand side of the curve; the “larger area” will be on the right. The “mean-to-*z*” and “ordinate” values are the same for negative *z*-scores.

The Sampling Distribution of Means and the Sample Size

We know from the examples in Chapters 5 and 9 that increasing the sample size will decrease the sampling error. By using sampling distributions based on the normal curve, constructed from a single sample with the mean and variance estimates provided by the Central Limit Theorem, we can see the effects of sample size even more clearly. Figure 10-5 illustrates sampling distributions constructed from samples of different size which were drawn from a population with a mean of 16 and a variance of 25.

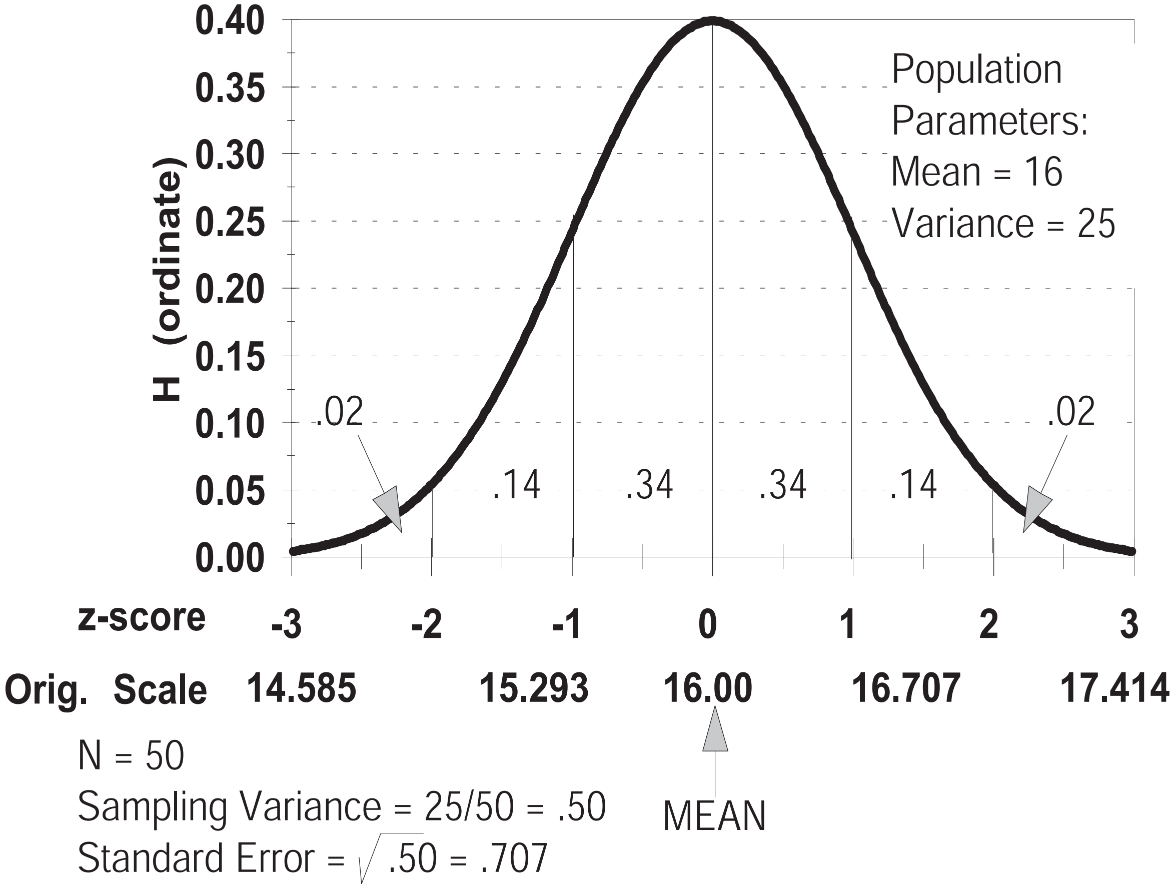
First, a general observation. Figure 10-5 shows us how the normal distribution can be used to associate probabilities with certain sample means. For instance, the first sampling distribution shows that when we take samples of *N* = 50 from this population we would expect 96% of all the sample means to fall between 14.585 and 17.41 inclusive. Only 2% of the sample means would be expected to be larger than 17.41 (or smaller than 14.585). In other words, applying the Normal distribution to the Sampling distribution allows us to distinguish between "likely" and "unlikely" sample means when sampling from a given population. The ability to do this is of key importance in hypothesis testing and this topic will be taken up again later.

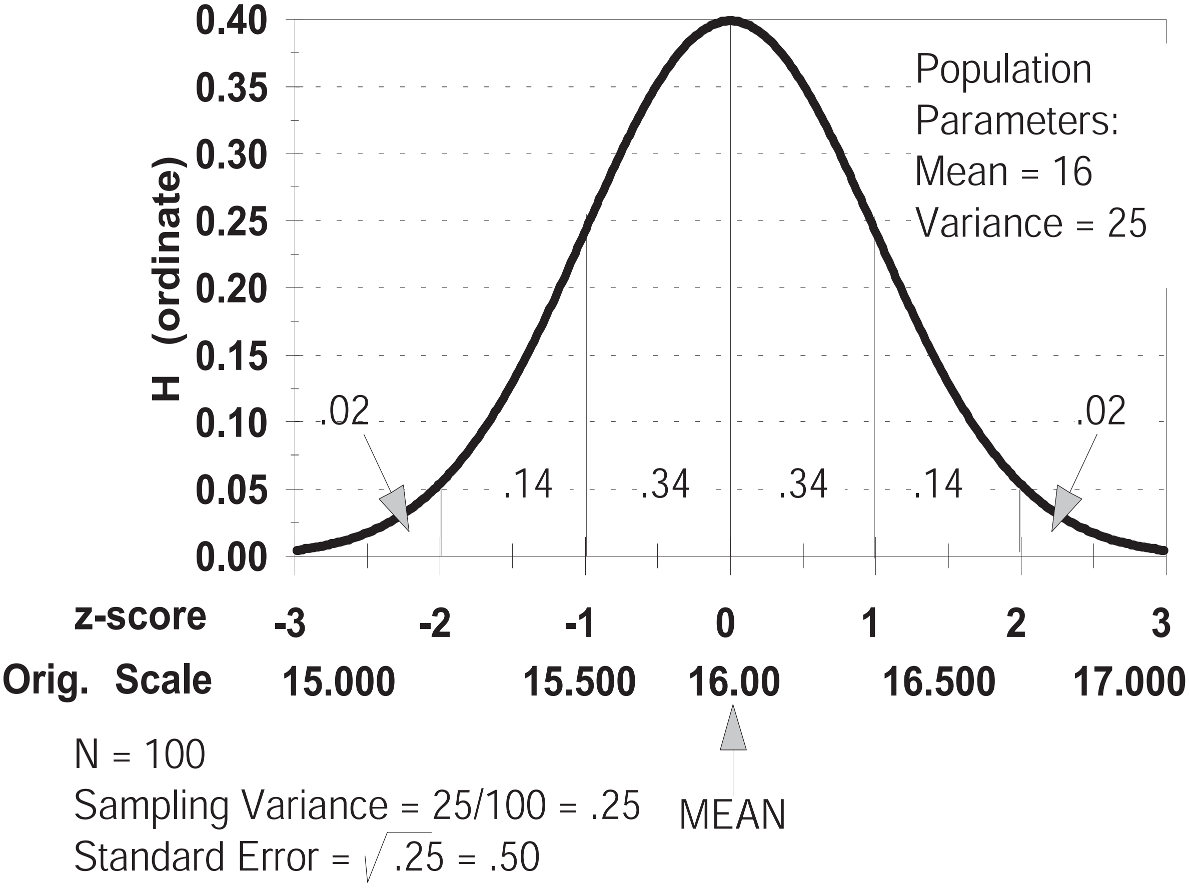
There are several items in Figure 10-5 you should pay close attention to. First, all sampling distributions have the same mean, since they are centered at the population mean. Further, since all are based on a normal curve, the *z*-values for each band of probabilities are the same. But these are the only things the three distributions have in common. Although each distribution has the same central tendency and is symmetrical, the dispersion differs. This can be seen by looking at the distribution as a function of the expected sample means, rather than the standard scores. The same ranges of *z*-values are associated with an increasingly narrower range of sample means. The sample means become more tightly clustered around the population mean when the sample size increases.

Thus the general observation that we can draw from Figure 10-5 is that increases in *N* are associated with decreases in sampling variance and standard error. That is, the sample means that we will obtain will be better estimates of the true value of the population mean as we increase the number of observations.

It is important to note, however, that when we plot the sample sizes and the values of the standard error, as in Figure 10-6, we observe the relationship between sample size and the size of the standard error to be non-linear. That is, for each identical increase in sample size we obtain a smaller reduction in the standard error.

When *N* = 50, the standard error was equal to .707. Doubling *N* to 100 (an increment of 50) brings the standard error down to .50, for a reduction of .207. An additional increment of 50 in *N* (to a total of 150) brings the standard error to .408, for a reduction of .092. The reduction in the standard error for the second increment is less than half of that of the first increment. Were we to increase *N* to 200, the new standard error would be .354, a reduction of .054, which is only one-fourth of the





reduction in the standard error obtained from the first increment of 50 in the sample size. Since subsequent increases in sample size will yield increasingly smaller reductions in sampling error, it clearly is not worth our while to continue increasing sample size indefinitely. In our current example, for instance, there doesn't seem to be much point in increasing the sample size beyond 150.

Another point about the relationship between sample size and sampling error is less evident, but just as important. Note that none of our discussion about sampling distributions and sampling error has involved the size of the population. Only the size of the sample has been shown to have an effect on sampling error. This seems counter-intuitive, but it's correct: sampling error in a true random sample is a function only of sample size. It doesn't matter if we sample from a population with 1000 members or with 100,000,000. The standard error depends only on the number of observations that we actually make.

Often communication or public opinion research is criticized for generalizing to huge populations on the basis of sample sizes that are small, relative to the total size of the population. For example, people often ask "How can the TV ratings services describe what's going on in 100,000,000 households when they only collect information in 1200 homes?" The answer to this question is contained in this chapter: the error in measurement due to sampling is determined solely by the sample *N* of 1200, and the total population size is irrelevant. A sample of 1200 in a single city with only 10,000 houses will be just as accurate (or inaccurate) as a national sample of the same size.

Summary

In this chapter, we demonstrated how the Central Limit Theorem can eliminate the need to construct a sampling distribution by examining all possible samples that might be drawn from a population. The Central Limit Theorem allows us to determine the sampling distribution by using the parameter mean and variance values or estimates of these obtained from a single sample. In particular, the theorem tells us that the estimated sampling distribution has a mean which is the

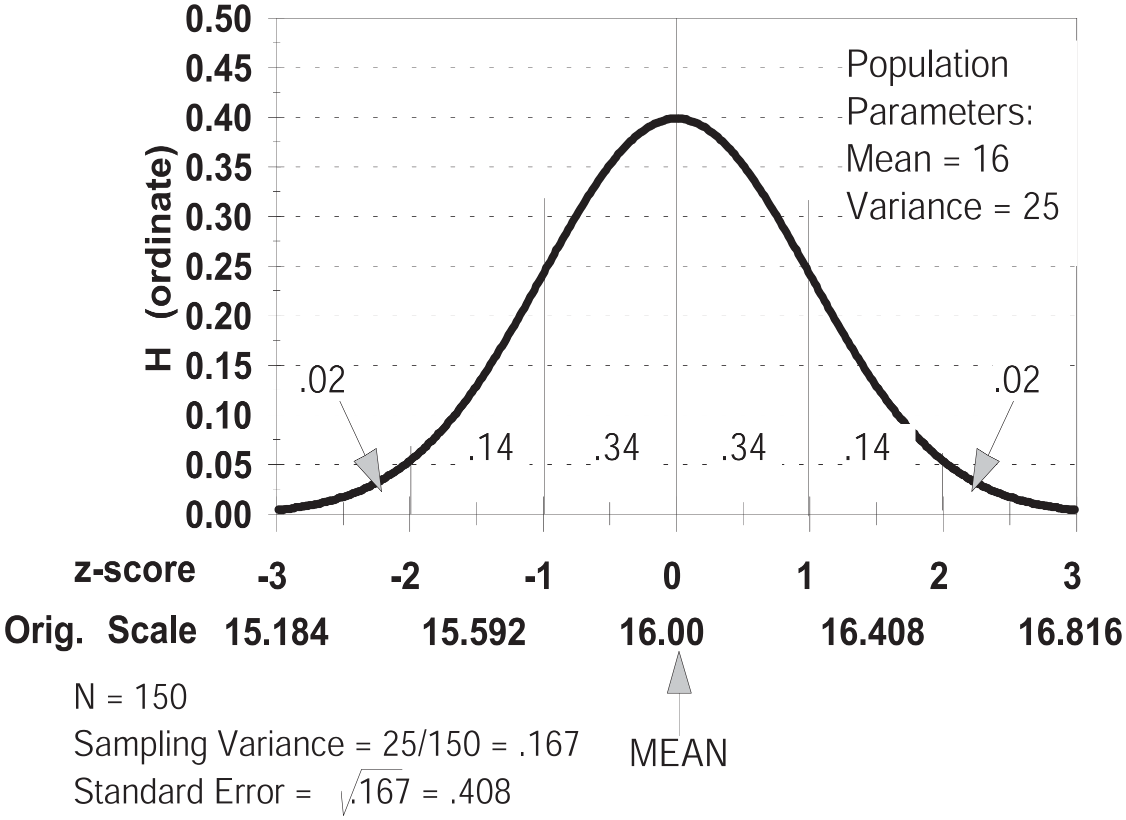
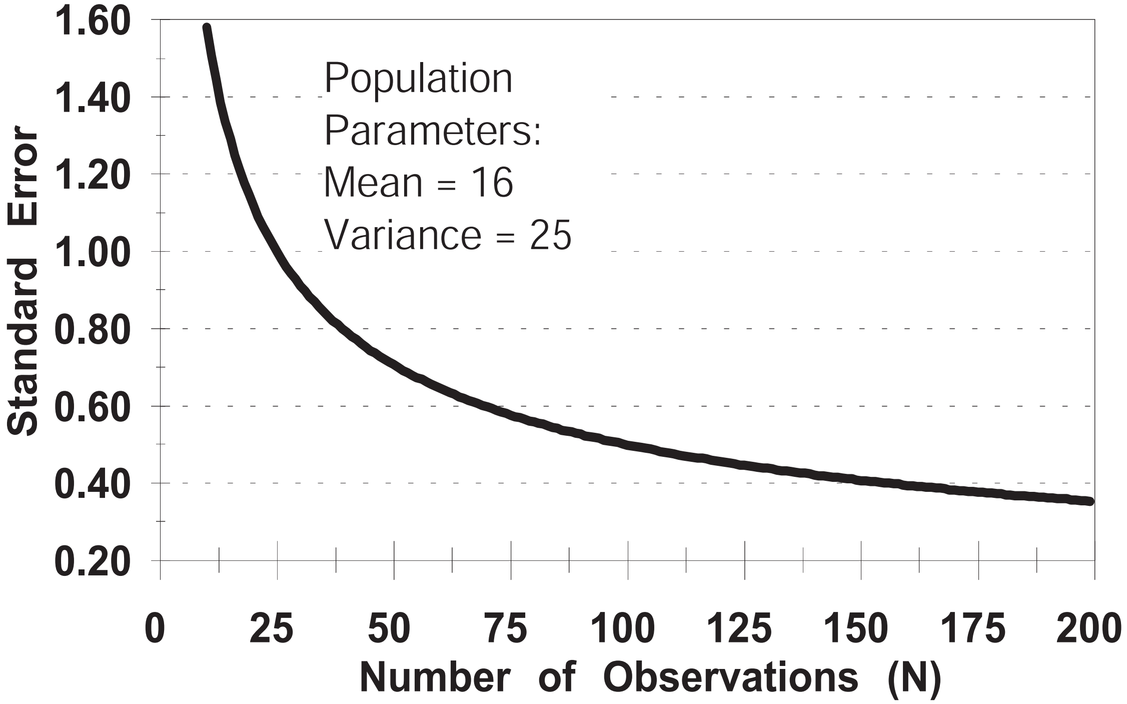


FIGURE 10-5C **Sampling Distribution of the Mean (*N* = 150)**



same as the population mean, or our best estimate, a sample mean, and that the variance of the sampling distribution can be determined from either the parameter or its estimate divided by the number of observations in the sample.

The third contribution of the Central Limit Theorem is that it tells us that the shape of the sampling distribution can be approximated by a normal curve. The normal curve can be described as a function of standard, or *z*, scores. Using the normal curve allows us to describe sampling error in terms of probabilities, which are represented as areas under the curve. Standard tables which describe the areas under the normal curve have been constructed to aid in making probability statements. We can use these tables to see what the likelihood is that samples with means a certain distance from the true population mean will be observed.

When the sample size is increased, we see that the range of values that correspond to a fixed probability level under the normal curve decreases, indicating a corresponding decrease in sampling error. The decrease in sampling error is a simple function of sample size, and is not related to the size of the population. Furthermore, increases in sample size and decreases in sampling error are related by a law of diminishing returns, making it inefficient to increase samples beyond a certain size.

With the tools provided by the Central Limit Theorem, we can proceed to the next task in communication research: testing hypotheses about the relationships between our theoretical constructs. In order to do this we will need to use the sampling distributions and the probability statements that they give us.

References and Additional Readings

Annenberg/CPB (1989). *Against all odds*. [Videotape]. Santa Barbara, CA: Intellimation.

Hays, W.L. (1981). *Statistics* (3rd. Ed.). New York: Holt, Rinehart & Winston. (Chapter 6, “Normal Population and Sampling Distributions”).

Kerlinger, F.*N*. (1986). *Foundations of behaviorial research* (3rd ed.) New York: Holt, Rinehart and Winston. (Chapter 12, “Testing Hypotheses and the Standard Error”).

Moore, D.S. & McCabe, G.P. (1989). *Introduction to the practice of statistics*. New York: Freeman. (Chapter 6, “From Probability to Inference”).