

Unit 17

The Central Limit Theorem

Objectives:

- To understand and apply the Central Limit Theorem

We shall now make a formal statement of a fact we have observed several times previously concerning the sampling distribution of \bar{x} . A rigorous proof of this theorem involves some very sophisticated mathematics. However, in place of a formal proof we provide a substantial amount of intuitive motivation to support what is known as The Central Limit Theorem.

The Central Limit Theorem. When selecting a simple random sample from a parent population which can be treated as an infinite population with mean μ and standard deviation σ , then the following are true:

- (1) The mean of the sampling distribution of \bar{x} is $\mu_{\bar{x}} = \mu$.
- (2) The standard deviation of the sampling distribution of \bar{x} is $\sigma_{\bar{x}} = \sigma / \sqrt{n}$.
- (3) If the parent population is normally distributed, then the sampling distribution of \bar{x} with samples of size n will be normally distributed for any n ; if the parent population is not normally distributed, then the sampling distribution of \bar{x} can be accurately approximated by a normal distribution for sufficiently large n .

The statement of the Central Limit Theorem we have made here reiterates in a very precise fashion the three facts about the sampling distribution of \bar{x} that we have previously noted on several occasions. Recall that Figure 14-3a displays the distribution of the parent population consisting of six values each painted on one of the six sides of a cube, Figure 14-3b displays the sampling distribution of \bar{x} with simple random samples of size $n = 2$, and Figure 14-4 displays the sampling distribution of \bar{x} with simple random samples of size $n = 6$. The fact that the mean is 8 for the parent population and for each of the sampling distributions of \bar{x} illustrates the first of the three statements ($\mu_{\bar{x}} = \mu$) in the Central Limit Theorem.

The fact that there is less variation in the sampling distributions of \bar{x} than in the parent population, and that the variation in the sampling distributions of \bar{x} decreases as the sample size n increases, illustrates second of the three statements in the Central Limit Theorem. However, this second statement in the Central Limit Theorem does not merely say that the variation in the sampling distributions of \bar{x} decreases as the sample size n increases; the Central Limit Theorem actually gives us a formula for finding the standard deviation of a sampling distribution of \bar{x} . While we have previously noted this decreasing variation with increasing sample size, as illustrated in Figures 14-3a, 14-3b, and 14-4, we have never done any computations involving exactly how much the variation decreases as the sample size increases. If we were to take the trouble to calculate the standard deviation for the parent population, which is σ , and then take the trouble to calculate the standard deviation for the sampling distribution of \bar{x} , which is $\sigma_{\bar{x}}$, we would find that $\sigma_{\bar{x}} = \sigma / \sqrt{n}$, which is what the Central Limit Theorem tells us. (Note: The method for calculating the standard deviation for a population is actually slightly different from the calculation of a sample standard deviation s , but this difference can be treated as negligible for our purposes.) If one were to take the time to do the appropriate calculations, one would find that $\sigma = 4.041$ for the parent population of Figure 14-3a. The Central Limit Theorem tells us that we can simply divide $\sigma = 4.041$ by the square root of the sample size to obtain the standard deviation for any of the sampling distributions of \bar{x} .

The fact that the sampling distribution of \bar{x} with $n = 6$ is bell-shaped, as seen in Figure 14-4, illustrates the last of the three statements in the Central Limit Theorem. The Central Limit Theorem actually tells us that, regardless of what type of distribution the parent population has, we can use a normal distribution to accurately approximate the sampling distribution of \bar{x} with a sufficiently large sample size n . Figure 14-4 seems to suggest that with simple random sample sizes of $n \geq 6$, we can treat the sampling distribution of \bar{x} as if it were a normal distribution. This is a powerful statement because it implies that we may use Table A.2 to obtain probabilities about \bar{x} from a simple random sample of size $n \geq 6$.

By applying the Central Limit Theorem to the sampling distribution of \bar{x} with $n = 6$, we would say that this sampling distribution can be treated as a normal distribution with mean $\mu_{\bar{x}} = \mu = 8$ and standard deviation $\sigma_{\bar{x}} = \sigma / \sqrt{n} = 4.041 / \sqrt{6} = 1.650$. We now pose the following question: What is the probability that the value of \bar{x} obtained from $n = 6$ rolls of the cube will be between 7 and 9? In other words, if someone did not know that $\mu = 8$, what is the probability that \bar{x} , which is used to estimate $\mu_{\bar{x}} = \mu$ is not more than a distance of 1 below or above $\mu = 8$. Since we have access to the actual sampling distribution, as displayed in Figure 14-4, we could obtain the exact value for this probability. The heights of the two bars in Figure 14-4 which give us the desired probability appear to be about 21% and 24%; more precisely, these heights are 20.74% and 23.81%.

Consequently, the probability that the sample mean \bar{x} obtained from a simple random sample of size $n = 6$ will be within a distance of 1 below or above $\mu = 8$ is $20.74\% + 23.81\% = 44.55\%$.

Let us now obtain this same probability by using the fact that the sampling distribution can be treated as a normal distribution with mean $\mu_{\bar{x}} = \mu = 8$ and standard deviation $\sigma_{\bar{x}} = \sigma / \sqrt{n} = 4.041 / \sqrt{6} = 1.650$. To find this probability, we first use $\mu_{\bar{x}} = 8$ and $\sigma_{\bar{x}} = 1.650$ to find the z -score of 7 to be

$$\frac{7 - 8}{1.650} = -1.37 ,$$

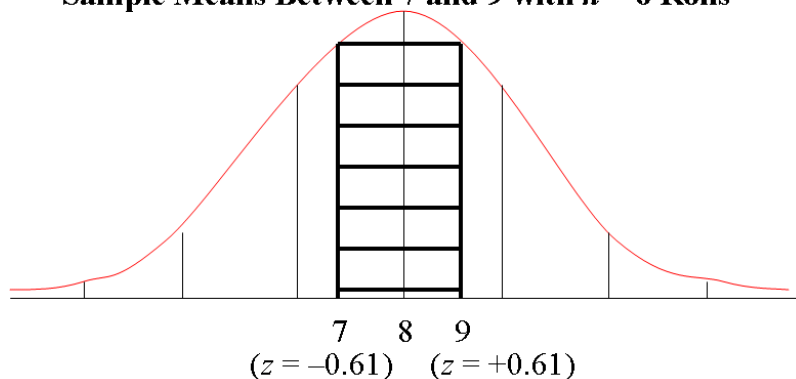
and to find the z -score of 9 to be

$$\frac{9 - 8}{1.65} = +0.61 .$$

The probability that \bar{x} from a random sample of $n = 6$ rolls will be between 7 and 9 is the proportion of shaded area in Figure 17-1. The unshaded area below 7 in Figure 17-1 is a mirror image of the shaded area in the figure at the top of Table A.2; also, the unshaded area above 9 in Figure 17-1 corresponds exactly to the shaded area in the figure at the top of Table A.2. Since a normal distribution is symmetric, we can find the total unshaded area by doubling the entry of Table A.2 in the row labeled 0.6 and the column labeled 0.01; we then obtain the desired area by subtracting this unshaded area from 1. The probability that \bar{x} from a random sample of $n = 6$ rolls will be between 7 and 9 is $1 - (0.2709 + 0.2709) = 0.4582$ (or 45.82%).

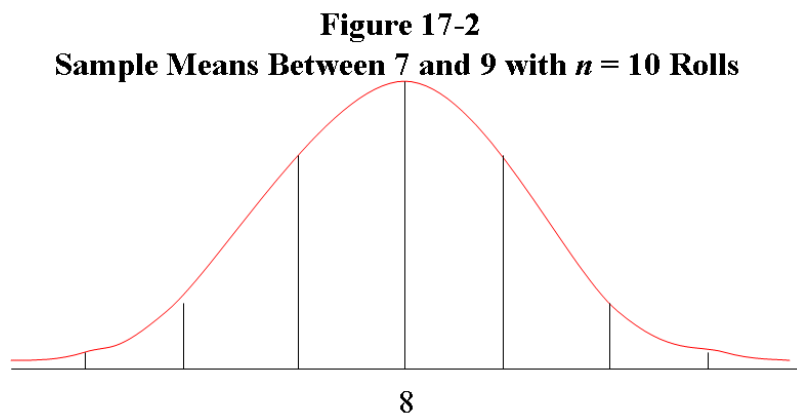
The probability 45.82%, obtained using the normal distribution, is close to the probability 44.55%, obtained directly from the sampling distribution (Figure 14-4). Since these probabilities concerned simple random samples of size $n = 6$, we would then expect probability calculations concerning simple random samples of size $n > 6$ to be even more accurate. As another illustration, we shall have you find the probability that the sample mean \bar{x} from a random sample of $n = 10$ rolls will be between 7 and 9. Find the z -score for each of 7 and 9, and label the values 7 and 9 on the horizontal axis in Figure 17-2. Then, shade the desired area under the

Figure 17-1
Sample Means Between 7 and 9 with $n = 6$ Rolls



normal curve, and use Table A.2 to obtain the desired probability. (You should find that the z -scores are -0.78 and $+0.78$, and that the probability that the probability that \bar{x} from a random sample of $n = 10$ rolls will be between 7 and 9 is 0.5646 (or 56.46%).)

It should certainly come as no surprise that the probability that \bar{x} will be between 7 and 9 increases as the sample size increases. We should expect the probability of obtaining a sample mean close to μ to increase with larger simple random sample sizes. One could ask how large the simple random sample size must be in order that the probability of obtaining \bar{x} between 7 and 9 is very large, say around 95%. The answer to this particular question is that the simple



random sample size must be at least $n = 63$ rolls. In general, to find the simple random sample size needed to insure that the probability of obtaining a sample mean no more than d away from μ is at least 95%, one can solve the equation $d = 1.96 / \sqrt{n}$ for n ; the value 1.96 is in the equation because the z -scores between which lies 95% of the area under a standard normal density curve are -1.96 and $+1.96$.

How large does a simple random sample size n have to be in order for us to be able use a normal distribution to obtain reasonably accurate probabilities involving \bar{x} ? In other words, how large does the simple random sample size n have to be in order to treat the sampling distribution of \bar{x} as a normal distribution? There is no single answer to this question. The answer depends on how different the distribution of the parent population is from a normal distribution. With the parent population consisting of the six values 0, 6, 9, 10, 11, and 12, whose distribution is displayed in Figure 14-3a, sampling distributions of \bar{x} with simple random samples of size $n > 6$ appear to be reasonably close to normal distributions. When the distribution of the parent population is extremely skewed (such as in Figures 15-1e or 15-1f) or has other characteristics that make it very different from a normal distribution, a very large sample size n may be necessary before the sampling distribution of \bar{x} can be treated as a normal distribution. On the other hand, if the distribution of the parent population is very close to being a normal distribution, or actually is a normal distribution, then the sampling distribution of \bar{x} can be treated as a normal distribution for any simple random sample size.

Recall that each of Figures 16-1 through 16-7 displays the density curve for the population of weights of oranges from a particular grove where the weights have a normal distribution with mean $\mu = 7.81$ oz. and standard deviation $\sigma = 1.32$ oz. Since this parent population has a normal distribution, we may assume that the sampling distribution of \bar{x} is a normal distribution for any random sample size n . Let us now pretend that we do not know that the mean orange weight for the population is $\mu = 7.81$ oz. (even though we really do know), and it is our intention to estimate this population mean with the mean from a simple random sample of orange weights. Further suppose that we are interested in the probability that our estimate is within 0.2 ounces of the population mean, that is, between 7.61 oz. and 8.01 oz. First, we shall consider basing our estimate on $n = 1$ randomly selected orange weight (which of course is the smallest possible sample size we could ever have!), then on $n = 9$ randomly selected orange weights, and finally on $n = 25$ randomly selected orange weights. We, of course, expect the probability that \bar{x} is within 0.2 ounces of the population mean to increase as n increases.

We begin by considering the probability that one randomly selected orange weight is within 0.2 ounces of the population mean. To find this probability, we first use $\mu = 7.81$ and $\sigma = 1.32$ to find the z -score of 6 oz. to be

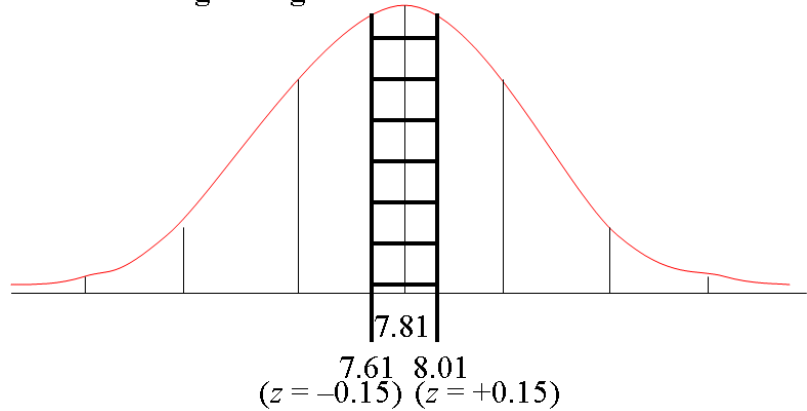
$$\frac{7.61 - 7.81}{1.32} = -0.15 ,$$

and to find the z -score of 8 oz. to be

$$\frac{8.01 - 7.81}{1.32} = - + 0.15 .$$

The proportion of shaded area in Figure 17-3 is the desired probability. The unshaded area below 7.61 in Figure 17-3 is a mirror image of the shaded area in the figure at the top of Table A.2; also, the unshaded area above 8.01 in Figure 17-3 corresponds exactly to the shaded area in the figure at the top of Table A.2. Since a normal distribution is symmetric, we can find the total unshaded area by doubling the entry of Table A.2 in the row labeled 0.1 and the column labeled 0.05; we then obtain the desired area by subtracting this unshaded area from 1. The probability that one randomly selected orange weight will be within 0.2 ounces of the population mean is $1 - (0.4404 + 0.4404) = 0.1192$ (or 11.92%).

Figure 17-3
Orange Weights Between 6 and 8 oz.



Next, we consider the probability that the sample mean \bar{x} for a simple random sample of $n = 9$ orange weights is within 0.2 ounces of the population mean. To find this probability, we first note that $\mu_{\bar{x}} = \mu = 7.81$ and standard deviation $\sigma_{\bar{x}} = \sigma / \sqrt{n} = 1.32 / \sqrt{9} = 0.44$. We then find the z-score of 7.61 oz. to be

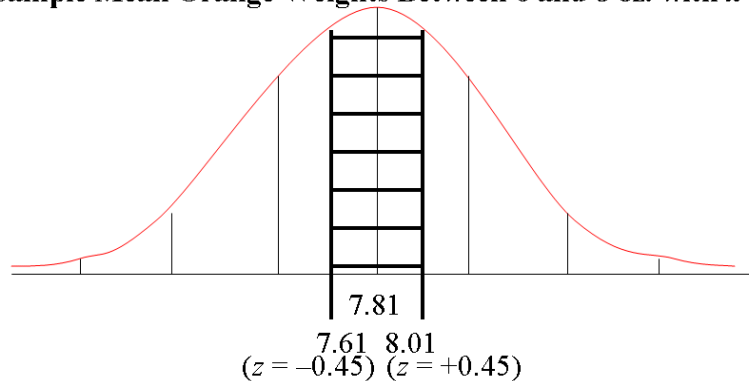
$$\frac{7.61 - 7.81}{0.44} = -0.45 ,$$

and find the z-score of 8.01 oz. to be

$$\frac{8.01 - 7.81}{0.44} = +0.45 .$$

The proportion of shaded area in Figure 17-4 is the desired probability. The unshaded area below 7.61 in Figure 17-4 is a mirror image of the shaded area in the figure at the top of Table A.2; also, the unshaded area above 8.01 in Figure 17-4 corresponds exactly to the shaded area in the figure at the top of Table A.2. Since a normal distribution is symmetric, we can find the total unshaded area by doubling the entry of Table A.2 in the row labeled 0.4 and the column labeled 0.05; we then obtain the desired area by subtracting this unshaded area from 1. The probability that \bar{x} from a simple random sample of $n = 9$ orange weights will be within 0.2 ounces of the population mean is $1 - (0.3264 + 0.3264) = 0.3472$ (or 34.72%).

Figure 17-4
Sample Mean Orange Weights Between 6 and 8 oz. with $n=9$



Notice that the shaded area in Figure 17-4 is larger than the shaded area in Figure 17-3, even though in both figures the shaded area represents the area under a normal density curve between 7.61 oz. and 8.01 oz. It is important to realize that the two density curves are not drawn to the same scale. Since Figure 17-3 represents the normal density curve for the parent population, and Figure 17-4 represents the normal density curve for the sampling distribution of \bar{x} with $n = 9$, there is actually much less variation in the density curve of Figure 17-4 than in the density curve of Figure 17-3. This would be apparent if we were to draw the two density curves to the same scale.

Finally, we consider the probability that the sample mean \bar{x} for a simple random sample of $n = 25$ orange weights is within 0.2 ounces of the population mean. To find this probability, we first note that

$\mu_{\bar{x}} = \mu = 7.81$ and standard deviation $\sigma_{\bar{x}} = \sigma / \sqrt{n} = 1.32 / \sqrt{25} = 0.264$. We then find the z-score of 7.61 oz. to be

$$\frac{7.61 - 7.81}{0.264} = -0.76 ,$$

and find the z-score of 8.01 oz. to be

$$\frac{8.01 - 7.81}{0.264} = +0.76 .$$

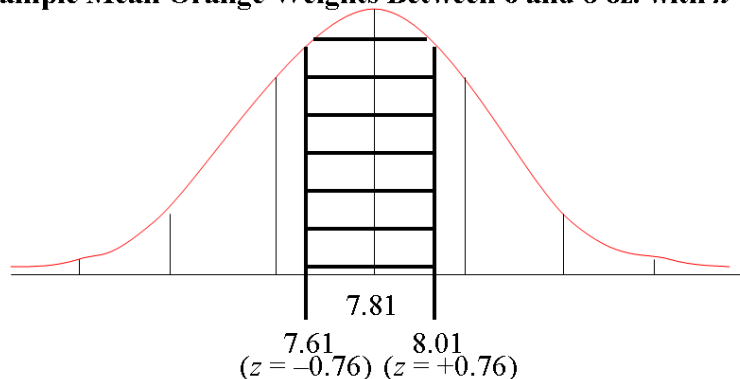
The proportion of shaded area in Figure 17-5 is the desired probability. The unshaded area below 7.61 in Figure 17-5 is a mirror image of the shaded area in the figure at the top of Table A.2; also, the unshaded area above 8.01 in Figure

17-5 corresponds exactly to the shaded area in the figure at the top of Table A.2. Since a normal distribution is symmetric, we can find the total unshaded area by doubling the entry of Table A.2 in the row labeled 0.7 and the column labeled 0.06; we then obtain the desired area by subtracting this unshaded area from 1. The probability that \bar{x} from a simple random sample of $n = 25$ orange weights will be within 0.2 ounces of the population mean is $1 - (0.2236 + 0.2236) = 0.5528$ (or 55.28%).

Notice that the shaded area in Figure 17-5 is larger than the shaded area in each of Figures 17-3 and 17-4, even though in all figures the shaded area represents the area under a normal density curve between 7.61 oz. and 8.01 oz. Once again, it is important to realize that the three density curves are not drawn to the same scale. Since Figure 17-3 represents the normal density curve for the parent population while Figures 17-4 and 17-5 represent the normal density curves for the sampling distributions of \bar{x} with $n = 9$ and $n = 25$ respectively, there is actually much less variation in the density curve of Figure 17-5 than in the density curve of Figure 17-4, just as there is actually much less variation in the density curve of Figure 17-4 than in the density curve of Figure 17-3. This would be apparent if we were to draw the density curves to the same scale.

As expected, the probability that \bar{x} from a simple random sample of orange weights will be within 0.2 ounces of the population mean increases as the sample size n increases. In other words, our chances that \bar{x} will estimate the population mean to within 0.2 oz. increase as our simple random sample size increases. Of course, this is the same as saying that our chances that \bar{x} will be more than 0.2 oz. away from the population mean decrease as our simple random sample size increases.

Figure 17-5
Sample Mean Orange Weights Between 6 and 8 oz. with $n=25$



Self-Test Problem 17-1. In Self-Test Problem 16-1, we were told that the right-hand grip strength for men between the ages of 20 and 40 is normally distributed with mean 86.3 lbs. and standard deviation 7.8 lbs.

- (a) Draw a sketch illustrating the probability that the right-hand grip strength for one randomly selected male is within 4 lbs. of the population mean, and find this probability.
- (b) Draw a sketch illustrating the probability that the mean right-hand grip strength for a simple random sample of 4 males is within 4 lbs. of the population mean, and find this probability.
- (c) Draw a sketch illustrating the probability that the mean right-hand grip strength for a simple random sample of 9 males is within 4 lbs. of the population mean, and find this probability.
- (d) Draw a sketch illustrating the probability that the mean right-hand grip strength for a simple random sample of 25 males is within 4 lbs. of the population mean, and find this probability.
- (e) Draw a sketch illustrating the probability that the right-hand grip strength for one randomly selected male is within 2 lbs. of the population mean, and find this probability.
- (f) Draw a sketch illustrating the probability that the mean right-hand grip strength for a simple random sample of 4 males is within 2 lbs. of the population mean, and find this probability.
- (g) Draw a sketch illustrating the probability that the mean right-hand grip strength for a simple random sample of 9 males is within 2 lbs. of the population mean, and find this probability.
- (h) Draw a sketch illustrating the probability that the mean right-hand grip strength for a simple random sample of 25 males is within 2 lbs. of the population mean, and find this probability.
- (i) Draw a sketch illustrating the probability that the right-hand grip strength for one randomly selected male is more than 1.5 lbs. away from (below or above) the population mean, and find this probability.
- (j) Draw a sketch illustrating the probability that the mean right-hand grip strength for a simple random sample of 10 males is more than 1.5 lbs. away from (below or above) the population mean, and find this probability.
- (k) Draw a sketch illustrating the probability that the mean right-hand grip strength for a simple random sample of 300 males is more than 1.5 lbs. away from (below or above) the population mean, and find this probability.
- (l) Should a male with a right-hand grip strength of 87.8 lbs. be considered extremely unusual? Why or why not?
- (m) Should a simple random sample of 10 males with a mean right-hand grip strength of 87.8 lbs. be considered extremely unusual? Why or why not?
- (n) Should a simple random sample of 300 males with a mean right-hand grip strength of 87.8 lbs. be considered extremely unusual? Why or why not?

Self-Test Problem 17-2. The mean yearly income per household in a particular region is to be estimated by selecting a simple random sample of households and obtaining the sample mean yearly income per household. It is believed that the distribution of yearly incomes is positively skewed and that the standard deviation in the region is about \$8000.

- (a) Why can we not expect to use the normal distribution to accurately find the probability that the yearly income for one randomly selected household is within \$1200 of the population mean?
- (b) Why might we reasonably expect to use the normal distribution to accurately find the probability that the mean yearly income for a simple random sample of n households is within \$1200 of the population mean, if say $n \geq 25$?
- (c) Draw a sketch illustrating the probability that the mean yearly income for a simple random sample of 25 households is within \$1200 of the population mean, and find this probability.
- (d) Draw a sketch illustrating the probability that the mean yearly income for a simple random sample of 100 households is within \$1200 of the population mean, and find this probability.
- (e) Draw a sketch illustrating the probability that the mean yearly income for a simple random sample of 400 households is within \$1200 of the population mean, and find this probability.
- (f) Is it very unlikely to select a simple random sample of 25 households with a sample mean income per household more than \$1200 away from (below or above) the population mean? Why or why not?
- (g) Is it very unlikely to select a simple random sample of 400 households with a sample mean income per household more than \$1200 away from (below or above) the population mean? Why or why not?

Strictly speaking, the Central Limit Theorem concerns a parent population which can be treated as an infinite population. In a very large number of practical situations, this is a reasonable assumption; however, there are special cases where it is not reasonable to treat the parent population as infinite. Suppose, for instance, that $n = 5$ accounts are to be selected from a population of 25 accounts in order to perform a very time consuming audit. A parent population consisting of only 25 items certainly cannot be treated as infinite.

But wait! You may be thinking to yourself that our parent population corresponding to the six-sided cube was certainly not infinite. However, there is a big difference between rolling the cube and selecting $n = 5$ out of the 25 accounts. The difference is that when rolling the cube, it is certainly possible that the same side will be facing up repeatedly, which in effect says that we may select the same side over and over again, but when selecting $n = 5$ out of the 25 accounts for auditing, it does not make sense to select the same account multiple times. The fact that the same side of the cube can occur multiple times is what makes the corresponding parent population infinite; the fact that we do not want to select an account multiple times is what makes the corresponding parent population finite.

Allowing items in a population to be selected multiple times is called *sampling with replacement*; not allowing any item in a population to be selected more than once is called *sampling without replacement*. Sampling with replacement always implies that we may consider our parent population to be infinite. When sampling without replacement from a population containing N items, we may treat the parent population as infinite, if the sample size n is only a very small fraction of the population size N . This is often true in many practical situations, but this is not true in the case where $n = 5$ accounts are to be sampled from $N = 25$ accounts. In situations where we cannot treat the parent population as infinite, the Central Limit Theorem can still be applied with the following modification: in place of $\sigma_{\bar{x}} = \sigma / \sqrt{n}$, we would have $\sigma_{\bar{x}} = c(\sigma / \sqrt{n})$ where $c = \sqrt{(N - n)/(N - 1)}$. We call c the *finite population correction factor*.

Since the finite population correction factor is needed only in special cases, we shall always assume (unless otherwise stated) that a parent population can be treated as infinite. We have stated the finite population correction factor here mostly for the sake of being informative and to be complete in our discussion of the Central Limit Theorem, not because we shall need to make heavy use of it.

Answers to Self-Test Problems

- 17-1** (a) 0.3900 or 39.00% (b) 0.6970 or 69.70% (c) 0.8764 or 87.64% (d) 0.9896 or 98.96% (e) 0.2052 or 20.52% (f) 0.3900 or 39.00% (g) 0.5588 or 55.88% (h) 0.7794 or 77.94% (i) 0.8494 or 84.94% (j) 0.5418 or 54.18% (k) practically 0 or 0% (l) Since about 42.47% of all males have a right-hand grip strength farther above the population mean than 87.8 lbs. ($z = +0.19$), we cannot consider this extremely unusual. (m) Since 27.09% of all simple random samples of 10 males have a mean right-hand grip strength farther above the population mean than 87.8 lbs. ($z = +0.61$), we cannot consider this extremely unusual. (n) Since less than 0.1% of all simple random samples of 300 males have a mean right-hand grip strength farther above the population mean than 87.8 lbs. ($z = +3.33$), we should consider this extremely unusual.
- 17-2** (a) Since the distribution of yearly incomes is positively skewed in the parent population, the normal distribution will probably not provide a good approximation. (b) If we are willing to assume that $n = 25$ is sufficiently large, then we may reasonably treat the sampling distribution of \bar{x} as a normal distribution. (c) 0.5468 or 54.68% (d) 0.8664 or 86.64% (e) 0.9974 or 99.74% (f) Since 45.32% of all simple random samples of 25 households have a mean income more than \$1200 away from the population mean, we cannot consider this very unlikely. (g) Since only 0.26% of all simple random samples of 400 households have a mean income more than \$1200 away from the population mean, we should consider this very unlikely.

Summary

The Central Limit Theorem says that when selecting a simple random sample from a parent population which can be treated as an infinite population with mean μ and standard deviation σ , then the following are true:

- (1) The mean of the sampling distribution of \bar{x} is $\mu_{\bar{x}} = \mu$.
- (2) The standard deviation of the sampling distribution of \bar{x} is $\sigma_{\bar{x}} = \sigma / \sqrt{n}$.
- (3) If the parent population is normally distributed, then the sampling distribution of \bar{x} with samples of size n will be normally distributed for any n ; if the parent population is not normally distributed, then the sampling distribution of \bar{x} can be accurately approximated by a normal distribution for sufficiently large n .

How large the simple random sample size n has to be in order to treat the sampling distribution of \bar{x} as a normal distribution depends on how different the distribution of the parent population is from a normal distribution. When the distribution of the parent population is extremely skewed or has other characteristics that make it very different from a normal distribution, a very large sample size n may be necessary before the sampling distribution of \bar{x} can be treated as a normal distribution; if the distribution of the parent population is very close to being a normal distribution, or actually is a normal distribution, then the sampling distribution of \bar{x} can be treated as a normal distribution for any simple random sample size.

Allowing items in a population to be selected multiple times is called *sampling with replacement*; not allowing any item in a population to be selected more than once is called *sampling without replacement*. Sampling with replacement always implies that we may consider our parent population to be infinite. When sampling without replacement from a population containing N items, we may treat the parent population as infinite if the sample size n is only a very small fraction of the population size N , and this is often the case. In situations where we cannot treat the parent population as infinite, the Central Limit Theorem can still be applied with the following modification: in place of $\sigma_{\bar{x}} = \sigma / \sqrt{n}$, we would have $\sigma_{\bar{x}} = c(\sigma / \sqrt{n})$ where $c = \sqrt{(N - n)/(N - 1)}$. We call c the *finite population correction factor*.