calculate sample means for data from a normal population. The mean of the sample means is very close to 205, the population value. In fact, for the theoretical sampling distribution of the means, the value is exactly 205. (Remember, the theoretical distribution of sample means is mathematically derived and tells you precisely what the distribution of the sample means is for all possible samples of a particular size.) In Figure 10.6, the standard deviation of the means, also known as the standard error of the mean, is 7.34.

### Standard Error of the Mean

You saw in Chapter 9 that the standard error of the mean tells you how much sample means from the same population vary. It depends on two things: how large a sample you take (that is, the number of cases used to compute the mean) and how much variability there is in the population. Means based on large numbers of cases vary less than means based on small numbers of cases. Means calculated from populations with little variability vary less than means calculated from populations with large variability.

If you know the population standard deviation (or variance) and the number of cases in the sample, you can calculate the standard error of the mean by dividing the standard deviation by the square root of the number of cases. In this example, the population standard deviation is 35 and the number of cases is 21, so the standard error of the mean is:

$$\frac{35}{\sqrt{21}} = 7.64 \qquad \textbf{Equation 10.1}$$

Note that the value we calculated based on the 500 samples with 21 hypothetical CEO's in each sample was not exactly 7.64, but very close. What we obtained was an *estimate* of the true value. That's because we did not take all possible samples from the population, but restricted our attention to 500.

? *Will the standard error of the mean always be smaller than the standard deviation of the data values?* Yes. It's always the case that the standard error of the mean is smaller than the standard deviation of the data values. That makes sense if you think about it. When you calculate a mean, it falls in between the smallest and largest sample values. It's not as extreme as the actual data values in your sample. Thus the mean has less variability than the original observations. The larger the sample that you take, the more you smooth out the variability of the individual data values when you calculate the mean. ■■■

## Are the Sample Results Unlikely?

Now that you know about the important properties of the sampling distribution of the mean from a normal population, let's return to the cholesterol levels of the CEO's. Figure 10.6 gives you a rough idea of how often you can expect various values for sample means when cholesterol is normally distributed in the population, with a mean of 205 and a standard deviation of 35. It's easy to see that the observed sample value of 193 is not a particularly unusual value.

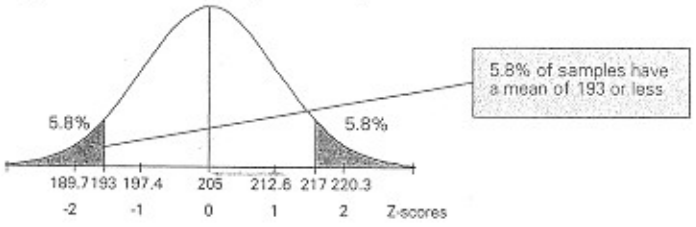
You can use the characteristics of the normal distribution to calculate *exactly* how often you would expect to see, based on 21 cases, a sample mean of 193 (12 less than the population mean) or less, or 217 or greater (12 more than the population mean). You're interested in both large and small cholesterol values, since you don't know in advance whether CEO values will be larger or smaller than those of the general population. It may be that the *foie gras* on Parisian business trips raises their cholesterol levels. Or that exercising in swanky health clubs while the rest of us work decreases their cholesterol levels.

First, you must calculate a standard score for the observed mean. You calculate it in the usual way: subtract the population mean from the observed mean and then divide by the standard deviation. The only trick to remember is that since you're dealing with a distribution of means, you must use the standard deviation of the means (the standard error of the mean), not the standard deviation of the sample values themselves. In our example, the standard score is

$$Z = \frac{193 - 205}{7.64} = -1.57 \qquad \textbf{Equation 10.2}$$

Look at Figure 10.7 for a summary of the situation.

**Figure 10.7 How unlikely is a sample mean of 193?**




You see that the distribution of all possible sample means of 21 cases is normal, with a mean of 205 mg/dL and a standard error of 7.64 mg/dL. The observed sample mean of 193 has a standard score of −1.57. In a normal distribution, 11.6% of the cases have standardized values less than −1.57 or greater than +1.57. Based on this, you don't have enough evidence to conclude that CEO cholesterol levels are different from those in the general population. (The observed significance level of 0.116 is larger than 0.05, the usual criterion for unusual.)

? *Since only 21 out of 200 CEO's responded, shouldn't you be concerned about the results from a survey that has a response rate of 10.5%?* Absolutely. There are many reasons why those who responded to the survey may differ from those who did not. It may be that CEO's with low cholesterol levels are more likely to volunteer this information than CEO's with high cholesterol levels. Or it may be that CEO's who have experienced medical problems are more likely to know, and perhaps to volunteer, their cholesterol levels than CEO's who are healthy. Our analysis was based on the rather shaky assumption that CEO's who responded don't differ from those who didn't. We also made the simplifying assumption that middle-aged males have the same cholesterol distribution as the general population. If this isn't the case we'd have to compare CEO values to those for middle-aged males. (Unfortunately, all results were from middle-aged males.) Our analysis also assumes that the CEO's reported their correct current cholesterol values. Anyone who has read an annual report to shareholders knows that CEO's can cast any kind of data in the best possible light. ■■■



## Testing a Hypothesis

In the previous example, you used statistical methods to test a hypothesis about the population based on results observed in a sample. Here's a summary of the procedure you followed:

1. You wanted to see if the average cholesterol levels of highly paid CEO's differ from those of the general population. You obtained a sample of cholesterol values from 21 such highly paid CEO's.
2. You calculated the average cholesterol value for the 21 CEO's in your sample to be 193 mg/dL.
3. You used the normal distribution with a mean of 205 and a standard error of 7.64 to determine how often you would expect to see average cholesterol values less than 193 or greater than 217, when the population mean is 205.
4. You found that sample means as unusual as the one you observed are expected to occur in about 11.6% of samples from the population, so you didn't have enough evidence to conclude that average cholesterol levels for CEO's are different from those of the population.

## Means from Non-Normal Distributions

You probably weren't too surprised that the distribution of sample means from a normal population is also normal. That makes a certain amount of sense. But it is surprising that the distributions of means shown in Figure 10.1 and Figure 10.2 at the beginning of this chapter also appear to be normal.

Remember that these are not means of a variable that has a normal distribution. The cure variable has only two equally likely values—0 for not cured and 1 for cured. This remarkable finding is explained by what's called the Central Limit Theorem. The **Central Limit Theorem** says that for samples of a sufficiently large size, the distribution of sample means

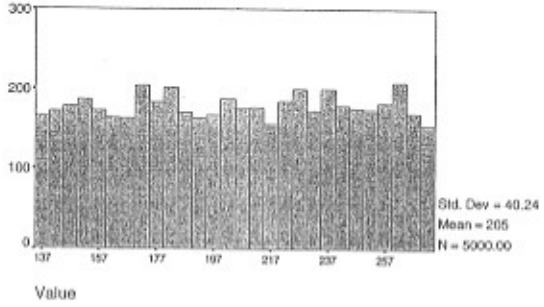
is approximately normal. The original variable can have any kind of distribution. It doesn't have to be bell shaped at all.

? *Sufficiently large size? What does that mean?* How large a sample you need before the distribution of sample means is approximately normal depends on the distribution of the original values of a variable. For a variable that has a distribution not too different from the normal, sample means will have a normal distribution even if they're based on small sample sizes. If the distribution of the variable is very far from normal, larger sample sizes will be needed for the distribution of sample means to be normal. The important point is that the distribution of means gets closer and closer to normal as the sample size gets larger and larger—regardless of what the distribution of the original variable looks like. ■■■

## Means from a Uniform Distribution

As an example of the Central Limit Theorem, let's see what the distribution of sample means looks like if cholesterol values had a uniform distribution in the population. In a uniform distribution, all values of a variable are equally likely. Figure 10.8 shows a histogram of 5000 values from a uniform distribution with a range of 135 to 275. All of the bars representing values from 135 to 275 are of approximately equal length.

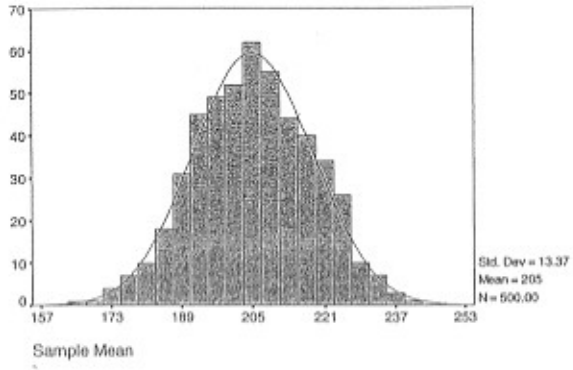
**Figure 10.8 A uniform distribution**




Let's see what happens if we take a sample of 10 cases from the distribution and compute their mean. Figure 10.9 shows the histogram of 500 such sample means.

**Figure 10.9 500 samples of 10 from a uniform distribution**

*To obtain this histogram, open the simul.sav file and select the variable unif10 in the Histograms dialog box.*




What is amazing is that the distribution of sample means looks nothing like the original distribution of values. The distribution of means is approximately normal, even when the distribution of a variable is not, provided that the sample size is large enough. This remarkable fact explains why the normal distribution is so important in data analysis. If the variable you're studying does have a normal distribution, then the distribution of sample means will be normal for samples of any size. The more unlike the normal distribution the distribution of your variable is, the larger the samples have to be for the distribution of means to be approximately normal. You'll be able to use the properties of the normal distribution to test a variety of hypotheses about population means based on the results observed in samples.

## Summary

*What is the normal distribution, and why is it important for data analysis?*

- A normal distribution is bell shaped. It is a symmetric distribution in which the mean, median, and mode all coincide. In the population, many variables, such as height and weight, have distributions that are approximately normal.
- Although normal distributions can have different means and variances, the proportional distribution of the cases about the mean is always the same.
- A standard normal distribution has a mean of 0 and a standard deviation of 1.
- The Central Limit Theorem states that for samples of a sufficiently large size, the distribution of sample means is approximately normal. (That's why the normal distribution is so important for data analysis.)