5

Four Fundamental Concepts

- *Before We Begin*
- *Fundamental Concept #1: Random Sampling*
- *Fundamental Concept #2: Sampling Error*
- *Fundamental Concept #3: The Sampling Distribution of Sample Means*
- *Fundamental Concept #4: The Central Limit Theorem*
- *Chapter Summary*
- *Some Other Things You Should Know*
- *Key Terms*
- *Chapter Problems*

This chapter deals with four fundamental concepts, some of which have been alluded to before. Everything we've covered up to this point is of little consequence if you gloss over the material in this chapter, so let me urge you to spend some serious time with the material you are getting ready to cover. If you have to read and reread and reread again, do that. The time spent will pay off.

First, we'll deal with the matter of random sampling. From that point, we'll take up the topic of sampling error—an essential notion that underlies the logic of inferential statistics. Then, we'll turn our attention to the idea of a sampling distribution, and more specifically, we'll look at the notion of a sampling distribution of sample means. Finally, we'll turn our attention to the Central Limit Theorem—a fundamental principle that will be important in our first major application of statistical inference.

As you go through the material, you'll likely have to take the time for a dark room moment or two—certainly more than you've had to up to this point. As I said before, don't assume that a dark room moment is beneath your intellectual dignity. Indeed, it may turn out to be a key to success when it comes to understanding the material.

Before We Begin

Let me pose two questions. First, how many times have you heard or read the expression, *random sampling*? Next, what about the expression, *sampling error,* or its cousin, so to speak, the *margin of error*? How many times have you encountered that?

My guess is you've heard phrases such as *random sampling* or *sampling error,* but you may not have a solid understanding of what each expression means. That's fine; it's not often that people have cause to think about such notions. On the other hand, those expressions are tied to some of the fundamental notions and assumptions that accompany statistical inference. From my perspective, it's virtually impossible to grasp the fundamental logic of statistical inference without some understanding of those concepts. Let me repeat: *It's virtually impossible to grasp the fundamental logic of statistical inference without some understanding of those concepts.*

Those concepts—random sampling and sampling error—are two of the concepts covered in this chapter. The other concepts—sampling distribution and the Central Limit Theorem—are no less important. I'm of the opinion that the four concepts, taken together, form the basis of a good amount of statistical inference. Therefore, it's paramount that you develop a firm understanding of each.

That said, I also know that you'll likely wonder why you have to learn each concept. Regrettably, I don't think you're going to like my answer. All I can tell you is that you're very close to entering the world of inferential statistics, and the concepts that you're about to encounter are central to opening the door.

On a positive note, you've more than earned a rest when you get through this chapter. It involves a hefty amount of material—conceptual and theoretical material that requires thinking in an abstract fashion. The chapter also makes reference to concepts that you've previously covered (for example, the standard deviation, populations, and samples). If you have any difficulty recalling what those previously introduced concepts are all about, go back to the earlier chapters to refresh your memory. A solid understanding of those concepts is essential.

Fundamental Concept #1: Random Sampling

Many statistical procedures rest on the assumption that you're working with a sample that was selected in a random fashion. The expression *random sample* is common, but it's also commonly misunderstood. Contrary to popular opinion, a random sample isn't what you get when you simply stand on the sidewalk and interview people who walk by. And a random sample isn't what you get when you use a group of students for research subjects just because they are available or accessible. To assert that you're working with a **random sample** of cases (or cases selected in a random fashion) means that you've met certain selection criteria.

First, a random sample is a sample selected in such a way that every unit or case in the population has an equal chance of being selected. There's a very important point to that requirement—namely, that you have in mind the population to which you intend to generalize. If, for example, you say that you're working with a random sample of registered voters, presumably you have in mind a population of registered voters that exists somewhere. It may be a population of registered voters throughout a city, or county, or state, or the nation. But you do have to have fixed in your mind a larger population in which you have an interest.

The second requirement is that the selection of any single case or unit can in no way affect the selection of any other unit or case. Let's say you devised a sampling plan that was based on your selecting first a Republican and then a Democrat and then a Republican and then a Democrat. If the idea of deliberately alternating back and forth in your selection of Republicans and Democrats is part of your sampling plan, you're not using a random sampling technique. Remember the criterion: The selection of one unit or case in no way affects the selection of another unit or case.

The third requirement of random sampling is that the cases or units be selected in such a way that all combinations are possible. This requirement is the one that really goes to the heart of inferential statistics, and it's the one you should key in on. The notion that all combinations are possible really means that some combinations may be highly improbable, but all combinations are possible. Indeed, British mathematician and philosopher Bertrand Russell (1955) illustrated the point with a little bit of humor. Describing a venture into a mythical hell while in a fever-induced state of delirium, Russell observed:

> There's a special department in hell for students of probability. In this department there are many typewriters and many monkeys. Every time that a monkey walks on a typewriter, it types by chance one of Shakespeare's sonnets. (p. 30)

Leaving Russell's mythical hell and returning to the more practical world of sampling, here's an illustration to consider. If you rely on a sampling technique that's truly random, and the population of registered voters is fairly evenly split between Republicans and Democrats, you'll probably end up with a sample that is roughly equally split between Republicans and Democrats. Your sample may not reflect the split between Republicans and Democrats in the population with exact precision, but it will probably be fairly close. It's very unlikely that you'll end up with a sample that is 100% Republican or 100% Democrat. Both of those outcomes (all Republicans or all Democrats) are highly improbable, but they are possible—and that's the point. If the sampling technique is truly random, all combinations are possible. For some further examples, see Figure 5-1.

The process of selecting cases or units in a random fashion typically begins with the identification of a **sampling frame,** or physical representation of the population. For example, your ability to make a statement about the population of registered voters in a certain county begins with your identification of a listing of all registered voters in that county. The listing, whether it exists on printed pages or in some electronic format, would constitute your sampling frame. If, on the other hand, you were interested in making a statement about all the students enrolled for six hours or more at a certain university, you would have to begin your sampling by locating some listing of all the students who met the criteria. Presumably you would get such a list from the registrar's office. That list, in turn, would serve as your sampling frame—a representation of your population.

In the case of a simple random sample, every case or unit in the sampling frame would be numbered, and then a *table of random numbers* would be used to select the individual cases for the sample. Most research methods texts have a table of random numbers included as an appendix to the book,

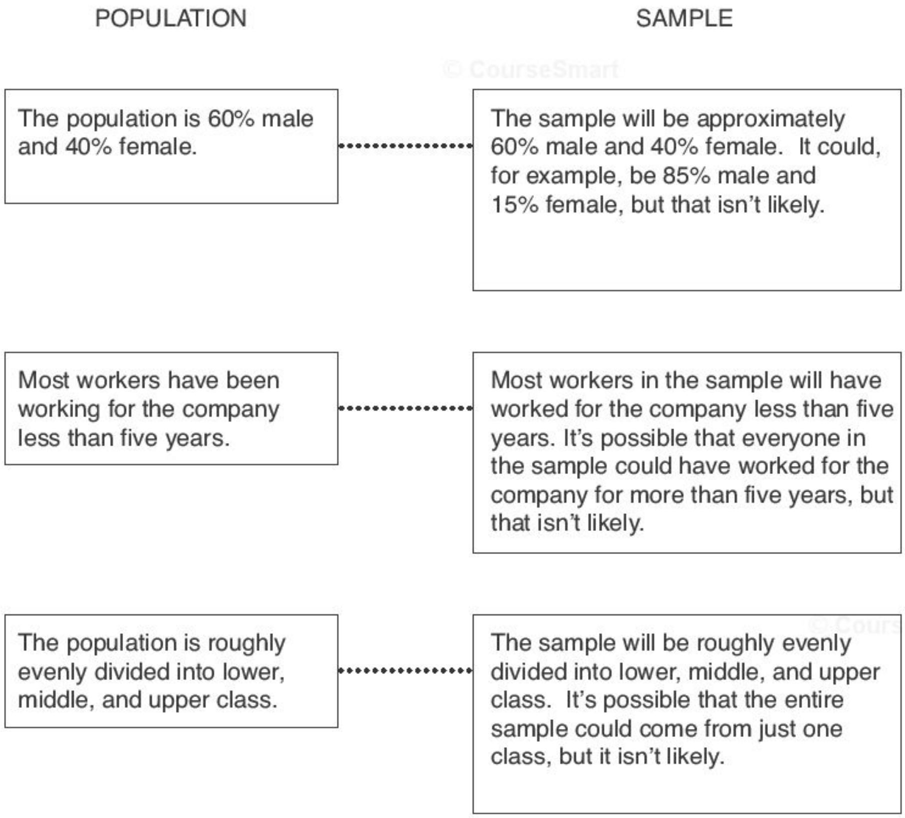


Figure 5-1 Relationship Between a Population and a Random Sample

and a quick read of the material on random sampling will provide you with a step-by-step procedure for selecting a simple random sample. In fact, most research methods texts include information on a variety of sampling designs—everything from systematic random sampling to stratified random sampling. For our purposes, though, you should simply have fixed in your mind what the term *random sampling* is all about and what is necessary if you're going to assert that you're working with a sample selected in a random fashion.

Fundamental Concept #2: Sampling Error

Assuming you've now got a grasp of what is meant by the concept of random sampling, it's time to turn to the concept of *sampling error*—something that was mentioned earlier, but only briefly. Now it's time to take a closer look. To illustrate the concept, we'll start with a simple example.

Let's say you're working as a university administrator, and you've been asked to provide an estimate of the average age of the students who are enrolled for six hours or more. Let's also say that, for all the reasons we've discussed before (factors such as time and cost), you've decided to rely on a sample to make your estimate—a random sample of 200 students from a population of 25,000 students (all enrolled for at least six hours of coursework).

The entire population of students probably includes a considerable range of ages. Some students might be extremely young—students who skipped a few years in high school because they were exceptionally bright. There may not be many students like that in the population, but there could be a noticeable number. By the same token, there might be a small but noticeable number of very old students—retirees who decided to return to school. Like the very young students, the older students would represent an extreme portion of the distribution.

The idea of sampling error comes into play with the recognition that an *infinite* number of samples are possible. You could take one sample, then another, then another (see Figure 5-2). You could continue the process time and

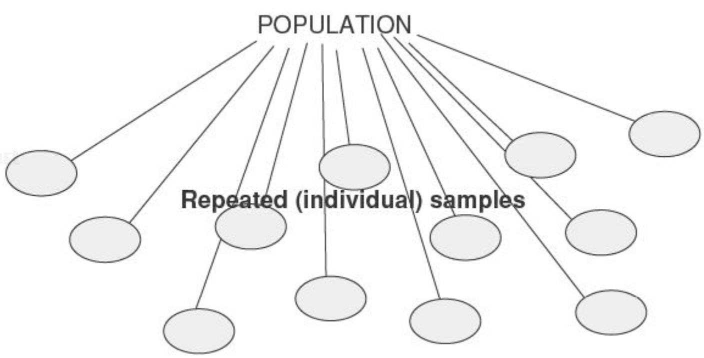


Figure 5-2 Representation of Repeated Samples from the Same Population

time again. You might not want to do something like that, but you could. And that's the point: An infinite number of samples are possible. This point is extremely important, so let me suggest here that you spend a dark room moment or two on it. Just think about the notion of taking sample after sample after sample from the same population. As ridiculous as that may seem, think about what the process would entail.

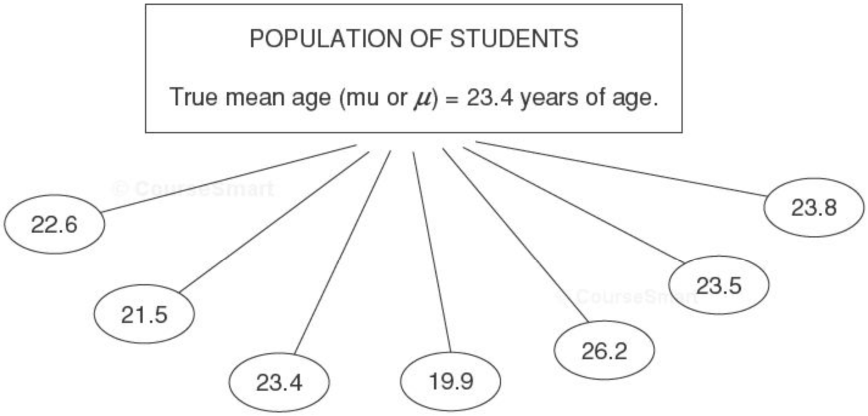
Assuming you've given some thought to the notion that an infinite number of samples is possible, let's now consider the real world. In reality, you'll have just one that you are working with. An infinite number of samples are possible, but you'll be working with only one of those samples. When it comes time to collect some information and carry out some calculation, you may think you're working with the best sample in the world (whatever that means) and that it is somehow a very special sample, but it really isn't special at all. In reality, you're working with one sample—just one out of an infinite number of samples that are possible—and your sample may or may not be an accurate reflection of the population from which it was taken.

What if, just by chance, you ended up with a sample that was somewhat overloaded with extremely young students? As you're probably aware, the chance of something like that happening may be small, but it's possible. In fact, you could end up with, let's say, 150 of the 200 cases somehow coming from the portion of the population distribution that contained the really young students. As I said, the chances are slim, but the possibility is there. By the same token, it's possible that you could end up with a random sample that was overloaded with extremely old students. Likely? No. Possible? Yes.

If your sample had an extreme overrepresentation of really young students, the sample mean age would be pulled down (the effect of extremely low values in the distribution). As a result, mean age for the sample wouldn't be a true reflection of the mean of the population (μ). Had you selected a sample that happened to have an overrepresentation of much older students, the mean of your sample would be higher than the true mean of the population. Once again, there would be a difference between your sample mean and the true mean of the population—just by chance.

You're probably starting to get the point of all this, but it's important that you understand the concept of sampling error at a level that's almost intuitive. For this reason, let me suggest that you take a serious look at the example shown in Figure 5-3. It illustrates what you might get in the way of several different sample means from one population. Even if you think you understand all of this, let me suggest that you pay attention to the specifics of the example. It takes very little effort, but it can help you understand the point in a way that will stay with you forever.

If you're starting to have a little conversation with yourself—if you're telling yourself, OK, I get it; this makes sense; of course I'd expect to see some difference—then you're on the right track as far as understanding one of the central concepts involved in statistical inference. What you've just dealt with is the concept of **sampling error**—the difference between a sample statistic and a population parameter that's just due to chance.



Seven samples and seven different sample means

One sample mean equals the mean of the population, but the other sample means are slightly higher or lower than the true population mean (mu or μ).

Figure 5-3 Illustration of Sampling Error

The difference could relate to a mean or a range or any other statistic. For example, a difference between the mean of the sample and the mean of the population (mu) that is just due to chance would amount to sampling error (of the mean). A chance difference between the range of the sample and the range of the population would also amount to sampling error (of the range). In both cases, we would categorize the difference as sampling error—the difference between a sample statistic and a population parameter that is due to chance.

You could be dealing with a lot of sampling error (particularly if you, by chance, came up with a rather extreme sample), or you could be dealing with only a small amount of it (if you came up with a highly representative sample). How statisticians deal with all of that is a topic for discussion down the road. For the moment, though, let's move forward to the next concept.

Fundamental Concept #3: The Sampling Distribution of Sample Means

To begin our discussion of this concept, I'll ask you to return to our earlier example. Imagine for a moment that you're taking sample after sample after sample from the population of students. The fact that nobody except a statistician is apt to do something like that shouldn't concern you. Just imagine for a moment that you're going through the exercise—taking sample after sample after sample. Let's say that each time you take a sample you select 50 students.

Now imagine that each time you select a sample, you ask students their age and record the information. You could easily calculate the mean age of each sample—right? Of course you could. As you learned in the previous section, though, the mean age of any one of those samples is likely to be slightly different from the population mean, just by chance (or due to sampling error). Let's say you went through the process 1000 times—each time selecting 50 students, collecting information on the students' ages, and calculating the mean age for that sample. If you recorded the mean for each of the 1000 samples, you would then have what is known as a **sampling distribution of sample means.**

At this point, let me suggest that you go no further unless you're absolutely certain you have that last notion firmly fixed in your mind. Here it is again: You could take sample after sample, selecting 50 students each time. You could repeat this process until you had selected 1000 samples. If you calculated the mean of each sample, you would then have a distribution of 1000 sample means. This distribution would be known as a sampling distribution of sample means.

There's no doubt about it, that phrase is a mouthful. So let's take it apart, element by element.

The result of your exercise would be a distribution, just like any other distribution (of income, weight, height, or any other variable). Only in this case, it would be a distribution of means taken from different samples—hence the expression *distribution of sample means.* You could just as easily have a distribution of sample ranges. All you would have to do is take sample after sample after sample, record the range of each sample, and report those ranges in a distribution. Typically, though, statisticians deal with the concept of a sampling distribution of sample means, rather than a sampling distribution of sample ranges.

The expression *sampling distribution* simply means a distribution that is the result of repeated sampling. Once again, it is a rather abstract concept, and very few people would ever bother to construct a sampling distribution of anything. But here's the point: You could construct a sampling distribution if you wanted to. As a matter of fact, you could very easily construct a sampling distribution of sample means. All it would take is a little bit of time. Once you did that, you could very easily develop a graph or plot of the sampling distribution of sample means. And that brings us to the last of the fundamental concepts.

Fundamental Concept #4: The Central Limit Theorem

Imagine for a moment that you had actually constructed the sampling distribution of sample means described in the previous example. In other words, you went to the trouble of taking 1000 different samples with 50 subjects in each sample. For each sample, you calculated and recorded a mean age, and you eventually put all the mean ages into a distribution.

Now imagine that you developed a graph or plot of all of those means, producing a curve. Do you have any idea what that curve might look like? Before you answer, think about the question for just a moment. Think about how you would produce the graph or curve, and what sort of values you would be plotting. Just to help you along in your thinking, consider the following:

1. You're taking sample after sample after sample (until you have 1000 samples).
2. Each time you take a sample, you calculate a mean age for the sample (a sample mean based on 50 cases).
3. Because of sampling error, your sample mean is likely to differ from the true mean of the population.
4. Sometimes your sample mean will be less than the true mean of the population.
5. Sometimes your sample mean will be greater than the true mean of the population.

By now you should be getting a picture in your mind of all these sample means (or sample mean values)—some higher than others, some lower than others, a few really high values, a few really low values, and so forth and so on. If you're getting the idea that the distribution of sample means would graph as a normal curve, you're on the right track. Now take a look at Figure 5-4. How do we know that a sampling distribution of sample means would look like a normal curve? We know it because it's been demonstrated. The idea has been tested; the idea holds up.

As it turns out, statisticians know quite a bit about what would happen if you set out to construct a sampling distribution of sample means. What's more, they know quite a bit about how the sampling distribution of sample means would be related to the population from which the samples were drawn. As a matter of fact, this relationship—the relationship between the sampling distribution of sample means and the population from which the samples were drawn—has a name. It is known as the Central Limit Theorem. Before we deal with the Central Limit Theorem and what it says, though, let me make three more points about a sampling distribution of sample means.

First, any sampling distribution of sample means will have a mean of its own—right? To convince yourself of that, just imagine a plot or graph of all the different means you would get if you took 1000 samples and plotted the means from those 1000 samples. The plot or graph would represent an underlying distribution, and that distribution (like any distribution) would have a mean. In the case we're discussing, it would be the mean of a sampling distribution of sample means.

Second, that distribution (the sampling distribution of sample means) would, like any distribution, have a standard deviation—right? Remember: The

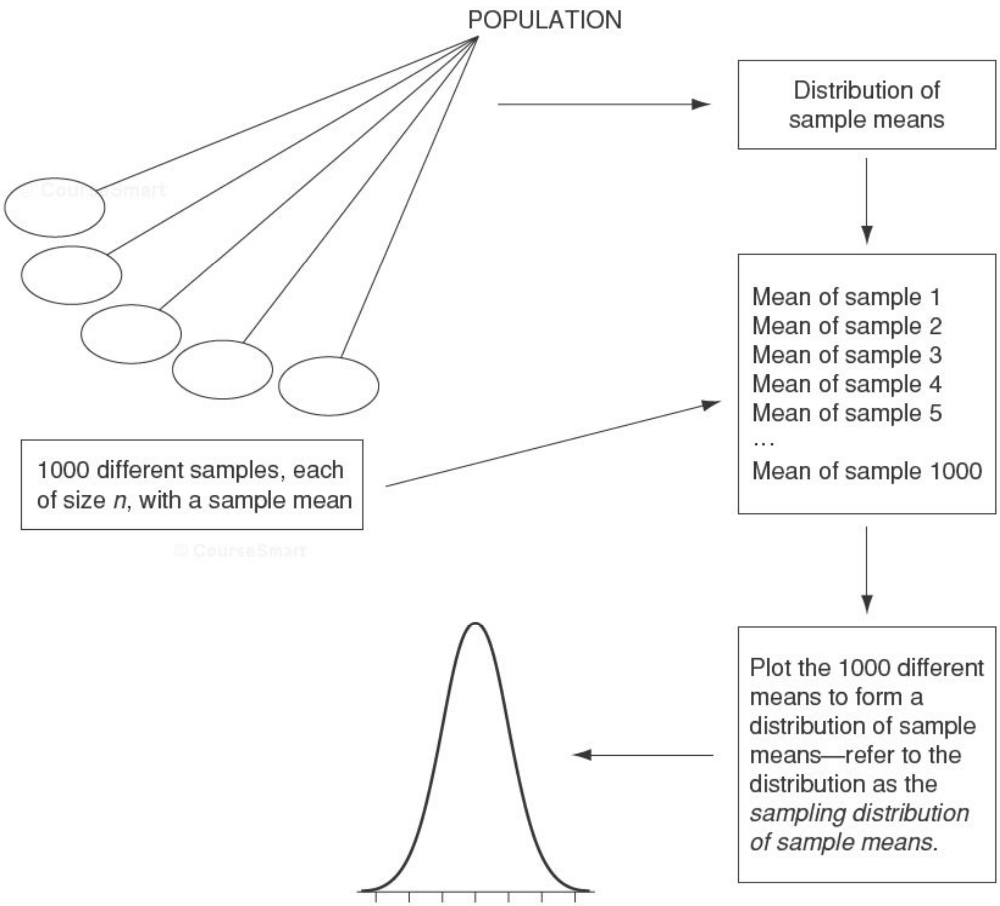


Figure 5-4 Constructing a Sampling Distribution of Sample Means

sampling distribution of sample means is, in a sense, just another distribution. All distributions have a standard deviation. In this case, we're considering a sampling distribution of sample means. It is no different. It would have a standard deviation.

Third, statisticians have a special term for the standard deviation of a sampling distribution of sample means. They refer to it as the **standard error of the mean.** That term or phrase, *standard error of the mean,* actually makes a lot of sense if you take a moment or two to think about it. It makes sense, in part, because a sampling distribution of sample means is actually a distribution of sampling error. The sampling distribution is based on a lot of means, and many of those means will actually vary from the true mean of the population. As you learned before, we refer to that chance difference between a sample mean and a population mean as *sampling error*—hence the term *error* in the expression *standard error of the mean.* Instead of saying *standard deviation of a sampling distribution of sample means,* statisticians use the expression *standard error of the mean.*

With all of that as background, let's now have a look at the Central Limit Theorem and what it tells us. First I'll present the theorem; then I'll translate.

Here is the **Central Limit Theorem:**

> If repeated random samples of size *n* are taken from a population with a mean or mu (μ) and a standard deviation (σ), the sampling distribution of sample means will have a mean equal to mu (μ) and a standard error equal to $\frac{\sigma}{\sqrt{n}}$. Moreover, as *n* increases, the sampling distribution will approach a normal distribution.

Now comes the translation: Imagine a population, and give some thought to the fact that this population will have a mean (mu or μ) and a standard deviation (σ). Now imagine a sampling distribution of sample means constructed from that population—a distribution of sample means, based on random sample after random sample after random sample, taken from the same population. That sampling distribution will have a mean, and it will equal the mean of the population (mu or μ). The sampling distribution of sample means will also have a standard deviation—something we refer to as the standard error of the mean. The standard error of the mean (the standard deviation of the sampling distribution of sample means) will be equal to the standard deviation of the population (σ) divided by the square root of *n* (where *n* is the number of cases in each sample).

In other words, a sampling distribution of sample means will eventually look like a normal curve (see Figure 5-5). Besides that, there's a very definite and predictable relationship between a population and a sampling distribution of sample means based on repeated samples from that population. We know that the relationship between the two is predictable because mathematicians have demonstrated that it is predictable.

It isn't the case that the mean of a sampling distribution of sample means will eventually be fairly close to or approximate the mean of the population (mu or μ). Instead, the mean of the sampling distribution of sample means will *equal* the mean of the population (mu or μ).

By the same token, it isn't the case that the standard deviation of the sampling distribution of sample means (the standard error) will sort of be related to the standard deviation of the population. Rather, the standard error will *equal* the population standard deviation (σ) divided by the square root of *n* (or the number of cases in the sample).

In the next chapter, we'll make some direct application of all of this material—but it won't do you any good to race ahead to the next chapter.

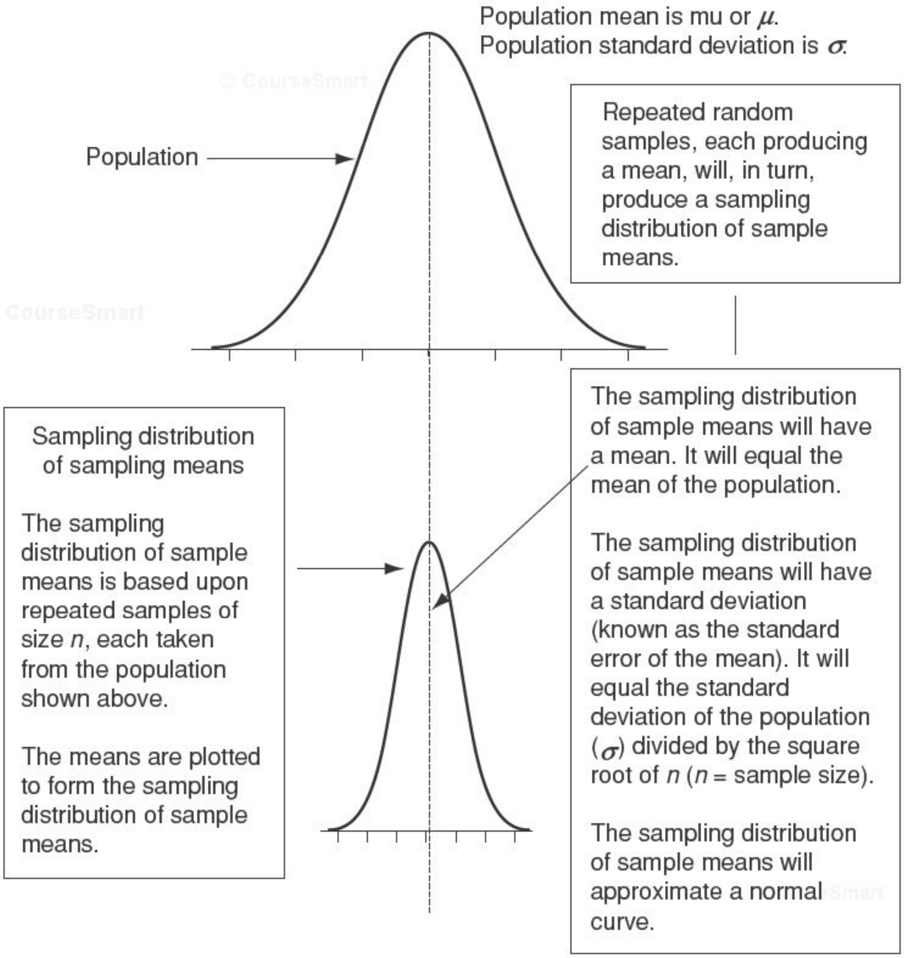


Figure 5-5 The Central Limit Theorem

Racing ahead without thoroughly understanding what we've just covered will only set you back in the long run. In fact, racing ahead will probably cause you to hit what I call the "brick wall of misunderstanding"—an experience that makes it impossible to understand all that lies ahead.

In my view, there's only one way to get over, under, around, or through the brick wall of misunderstanding, and that's to focus on the fundamental concepts until you finally understand each one of them. It won't do to tell yourself you understand when you don't. Instead, reread this entire chapter, if you have to. Read it and reread it until you understand the material at a near-intuitive level. Once you've done that, you'll be in a position to more forward.

Chapter Summary

At this point, you deserve a break. You've just been through some rather abstract and theoretical territory. If you found the material a little tough to digest at the outset, that's normal. The material is new by all reasonable standards—new concepts, new ideas, and new ways of looking at the world. New material? You bet. Difficult material? Not really. It's all a matter of thinking about each element until you have a solid understanding. As to what you just covered, it was significant.

For example, you were introduced to a technical definition of random sampling, in a way that emphasized what a random sample is and is not. You also learned that the assumption of a random sample is central to many statistical applications. Equally important, you were introduced in some detail to the concept of sampling error. Ideally, you learned that it is sampling error that prevents a direct leap from sample statistics to population parameters. Beyond all of that, you were introduced to the concept of a sampling distribution of sample means and the Central Limit Theorem. In the process, you found your way into the heart of statistical inference (at least as it relates to certain applications). A lot of material, indeed.

As we close out this chapter, let me underscore how beneficial a dark room moment might be for understanding some of the concepts that you just covered. These concepts deserve your full attention, and that's what a dark room moment is all about—a chance to bring your full attention to the question at hand.

Some Other Things You Should Know

Normally, I use this section of each chapter to point you in the direction of relevant topics left unexplored in the interest of a succinct presentation. The chapter you just read justifies a departure from that approach. Instead of pointing you to unexplored topics or directing you to additional resources, I'm going to let you in on a little secret. Here it is.

The material you just covered is, for many students, the source of the brick wall. It's the collection of concepts that ultimately separate the women from the girls and the men from the boys. My experience in teaching statistics tells me that many students say they "get it" when, in fact, they don't. The issue, of course, isn't what the students tell me; it's what they tell themselves.

The four fundamental concepts presented in this chapter will eventually be linked for you in the form of practical applications. But the logic of those applications always comes back to the fundamental concepts, and that's why they are so essential.

There's no question that some of the concepts are highly abstract. Indeed, it is this collection of concepts that always come to my mind when I stress the importance of taking time out for a dark room moment. Much material remains to be covered, so don't hamper your learning by going forward unprepared. If you need to take time out for a few dark room moments, now is the time to do it. Shore up the moments with a second or third read of the material, if necessary.

Key Terms

Central Limit Theorem
random sample
sampling distribution of sample means
sampling error
sampling frame
standard error of the mean

Chapter Problems

Fill in the blanks, calculate the requested values, or otherwise supply the correct answer.

General Thought Questions

1. In a random sample, every unit in the population has a(n) __________ chance of being selected.
2. In a random sample, the selection of any one unit __________ affect the selection of any other unit.
3. In a random sample, __________ combinations are possible.
4. When selecting a sample, the physical representation of the population is known as the __________.
5. A representative sample is one in which important characteristics in the population are mirrored in the __________.

11. According to the Central Limit Theorem, and given a sampling distribution of sample means, the standard error of the mean will equal the __________ of the population divided by the __________ of the sample size.

12. The shape of a sampling distribution of sample means will approach the shape of a __________ curve.

Application Questions/Problems

1. A population has a mean (μ) of 24.12 and a standard deviation (σ) of 4. Assume that a sampling distribution of sample means has been constructed, based on repeated samples of n = 100 from this population.
- **a.** What would be the value of the mean of the sampling distribution?
- **b.** What would be the value of the standard error of the mean?

2. A population has a mean (μ) of 30 and a standard deviation (σ) of 6. Assume that a sampling distribution of sample means has been constructed, based on repeated samples of n = 225 from this population.
- **a.** What would be the value of the mean of the sampling distribution?
- **b.** What would be the value of the standard error of the mean?

3. A population has a mean (μ) of 120 and a standard deviation (σ) of 30. Assume that a sampling distribution of sample means has been constructed, based on repeated samples of n = 100 from this population.
- **a.** What would be the value of the mean of the sampling distribution?
- **b.** What would be the value of the standard error of the mean?

4. A population has a mean (μ) of 615 and a standard deviation (σ) of 90. Assume that a sampling distribution of sample means has been constructed, based on repeated samples of n = 400 from this population.
- **a.** What would be the value of the mean of the sampling distribution?
- **b.** What would be the value of the standard error of the mean?

5. A population has a mean (μ) of 55 and a standard deviation (σ) of 17. Assume that a sampling distribution of sample means has been constructed, based on repeated samples of n = 100 from this population.
- **a.** What would be the value of the mean of the sampling distribution?
- **b.** What would be the value of the standard error of the mean?