NORUŠIS, Marija J. 1998. *Guide to Data Analysis*. Upper Saddle River, Prentice Hall.

9 Evaluating Results from Samples

What can you say about a population, based on the results observed in a random sample?

- Are the results you observe in a sample identical to the results you would observe from the entire population?
- What is the sampling distribution of a statistic?
- How is it used to test a hypothesis about the population?
- What factors determine how much sample means vary from sample to sample?
- What is an observed significance level?
- What is the binomial test, and when do you use it?

In previous chapters, you've answered questions like "What percentage of survey respondents are very satisfied with their jobs?" or "What is the relationship between job satisfaction and education?" All you did was describe the results you found in the General Social Survey (GSS). Nothing more.

In this section of the book, you'll begin to look at the problems you face when you want to draw conclusions about a larger number of people or objects than those actually included in your study. You'll learn how to draw conclusions about the population based on the results observed in a sample.

▶ This chapter uses generated computer data in the file *simul.sav*. For information on how to obtain the binomial test results shown in the chapter, see "Binomial Test" on p. 339 in Chapter 17.

From Sample to Population

In the General Social Survey sample, almost 44% of people employed full time rated themselves as being *very satisfied* with their jobs. Unless errors have been made while recording or entering the data, you know this for a fact. Similarly, you know exactly how old the people in the sample are, how much education they have, and so on. You can describe in great detail and with much certainty the results observed in this sample. Unfortunately, that's not really what's of interest. What you really want to do is draw conclusions about the larger group that the people in the GSS represent, the **population**.

The participants in the GSS are a sample from the population of adults in the United States. Based on the results you observe from the participants, you want to draw conclusions about *all* adults in the United States. You want to be able to say, for example, that in the United States, highly paid workers are more satisfied with their jobs than those paid less.

On first thought, that might not seem too complicated. Why not assume that what's true for the sample is also true for the population? That would certainly be simple. But would it always be correct? Do you really believe that, since 43.8% of the full-time workers in your sample are *very satisfied* with their jobs, that's exactly the percentage of *very satisfied* people in the population? Common sense tells you that it's very unlikely that the results you see in a sample are identical to those you would obtain if you made measurements or inquiries of the entire population of interest. If that were the case, one quick poll before an election would eliminate the need to even hold elections.

What's true instead is that different samples give different results, and it's highly unlikely that any one sample will hit the population results on the nose. To see what you can conclude about the population based on a sample, you must consider what results are possible when you select a sample from a population.

A Computer Model

Although we could use mathematical arguments to derive the properties of samples and populations, it's less intimidating and more fun to discover them for yourself. You can use the computer to keep drawing random samples from the same population and see how much the results change from sample to sample. This process is known as a **computer simulation**.

? *What's a random sample?* A random sample gives every member of the population (animal, vegetable, mineral, or whatever) the same chance of being included in the sample. No particular type of creature or thing is systematically excluded from the sample, and no particular type is more likely to be included than any other. Each member is also selected independently; including one particular member doesn't alter the chance of including another.

A sample is **biased** if, for example, rich people have a better chance of being included than poor people, or healthier people are more likely to be selected than sick people. You can't draw correct conclusions about the population based on the results from such a sample. ■ ■ ■

Let's use the computer to solve the following problem. A Noted Physician claims that she has a better treatment for the Disease of Interest. Of 10 patients who received her new treatment, 70% were cured. Extensive literature on the topic indicates that nationwide, only 50% of patients with this disease are cured. Based on the results of her experiment, can you tell if the physician has really made inroads into the treatment of this disease?

Are the Observed Results Unlikely?

To evaluate the physician's claim, you have to ask yourself the question, Are the results she observed (7 out of 10 cures) unlikely if the true population cure rate is 50%? You know that if half of all people with a disease can be cured, that doesn't mean that any time you select 10 patients, exactly 5 will be cured by the treatment. Consider a coin-tossing analogy. You know that if a coin is fair, heads and tails are equally likely. If you flip a fair coin 10 times, however, you don't expect to see exactly 5 heads every 10 flips. Sometimes you get more heads and sometimes, more tails. (Try flipping a coin 10 times and see how many heads—cures—you get. Record your results. Repeat this as many times as you have the patience for and then make a stem-and-leaf plot of the results. You can compare your results with those you'll see in this chapter.)

To evaluate the physician's claim, instead of spending the afternoon flipping a coin, you can use the computer to construct a population in which half of the patients are cured and half are not. That's the situation if the physician's claim is not true. Then you can have the computer take a random sample of 10 patients and record the percentage that are cured. Have it repeat this procedure 500 times.

The reason you're doing this is to see what kind of sample results are possible if the new treatment is not different from the standard one. You can then determine whether finding 70% cured in a sample of 10 patients is an unusual finding when the true cure rate is 50%.

A stem-and-leaf plot of the results of the 500 experiments is shown in Figure 9.1.

You can obtain stem-and-leaf plots using the Explore procedure, as described in Chapter 6. Select the variable cured10 in the Explore dialog box.

Figure 9.1 Stem-and-leaf plot of percentage cured for sample size 10

```
Frequency    Stem &  Leaf

     3.00 Extremes    (=<10)
    24.00        2 .  00000000
    70.00        3 .  00000000000000000000000
    98.00        4 .  000000000000000000000000000000000
   114.00        5 .  00000000000000000000000000000000000000
    95.00        6 .  00000000000000000000000000000000
    65.00        7 .  0000000000000000000000
    24.00        8 .  00000000
     6.00 Extremes    (>=90)

 Stem width:     10.00
 Each leaf:      3 case(s)
```

For most samples, the cure rate is close to 50%

From this plot, you can tell approximately how often you would expect to see various outcomes in samples of size 10. The distribution of all possible sample outcomes for a statistic (such as the percentage cured) is called the **sampling distribution** of the statistic.

? *Exactly what is a statistic anyhow?* A **statistic** is some characteristic of a sample. The sample mean and variance are both examples of statistics. The term **parameter** is used to describe the characteristics of the population. For example, the average height of people in your sample is a statistic. If you measured the heights of all people in the population of interest, that would be called a parameter of the population. Parameters are usually designated (by statisticians, at least) with Greek symbols. For example, the mean of a population is called μ (mu), while the mean of a sample is called $\bar{X}$. Similarly, the standard deviation of the population is called σ (sigma), while the value for a sample is called s. Most of the time, population values, or parameters, are not known. You must estimate them based on statistics calculated from samples. ■ ■ ■

The sampling distribution is usually calculated mathematically. In this case, you're using a computer to give you some idea of what it looks like. In Figure 9.1, you see that for most samples, the percentage of cures is close to 50%. In fact, 307 out of the 500 experiments resulted

NORUŠIS, Marija J. 1998. *Guide to Data Analysis*. Upper Saddle River, Prentice Hall.

in cure rates of 40%, 50%, or 60%. The further you move from 50%, in either direction, the fewer samples you see. Although various outcomes are possible, the outcomes are not equally likely. For example, only 6 experiments out of 500 resulted in a cure rate of 90% or greater.

You can calculate descriptive statistics for the data summarized in Figure 9.1. These summary statistics are shown in Figure 9.2. The values range from a minimum of 10% to a maximum of 90%, but the mean is very close to 50%. (In fact, for the mathematically computed sampling distribution, the mean value is exactly 50%, the mean of the population from which the samples are being drawn.) The standard deviation of the percentages, labeled *Std. Deviation* in Figure 9.2, is 16.22%. The standard deviation tells you how much the percentage cured varies in samples of size 10. (The standard deviation of the distribution of all possible values of a statistic is called the **standard error** of the statistic. For example, the standard deviation of all possible values of a sample mean is called the standard error of the mean.)

You can obtain these statistics using the Descriptives procedure, as described in Chapter 4.

Figure 9.2 Descriptive statistics for samples of size 10

	N	Minimum	Maximum	Mean	Std. Deviation
CURED10	500	10.00	90.00	50.0200	16.2212
Valid N (listwise)	500				

? *What's the difference between a standard deviation and a standard error?* Standard deviation refers to the variability of the observations in a sample. The term standard error is used when you are talking about the variability of a statistic. For example, if you have a sample of 10 systolic blood pressures, you can calculate their mean, variance, and standard deviation in the usual way. From the standard deviation of the 10 blood pressure measurements, you can also estimate how much *average* blood pressures calculated from samples of 10 people vary. That's the standard error of the mean for samples of this size. Figure 9.2 contains descriptive statistics for 500 means for samples of size 10. The standard deviation of these 500 means is an estimate of the standard error of the mean for samples of size 10. ■■■

Using Figure 9.1 as a guideline, you can estimate whether the physician's results are unusual if the true cure rate is 50%. You see that 96 out of 500 simulated experiments (19.2%) resulted in cure rates of 70% or more. That indicates that even if the new treatment is no better than the standard, you would expect to see cure rates at least as large as those observed by the physician almost 1 out of 5 times you repeated the experiment. (In fact, it is possible to calculate mathematically that the probability of obtaining 7 or more cures in a sample of 10 is close to 17% when the true cure rate is 50%.)

Of course, it's always possible that the new treatment is really *less* effective than the usual treatment. So if you want to test the hypothesis that the new treatment is not different from the standard treatment, you must evaluate the probability of results as extreme as the one observed in either direction—increasing or decreasing the cure rate. You can estimate from Figure 9.1 that the probability of 30% or fewer cures and the probability of 70% or more cures is $(96 + 97)/500 = 38.6\%$.

Based on this, you have little reason to believe that the physician is really onto something. Her results are certainly not incompatible with samples selected from a population in which the true cure rate is 50%.

? *Why look at cure rates of 70% or more* and *cure rates of 30% or less?* Consider the following analogy. Your friend gives you a coin and claims that it is not fair. That is, heads and tails are not equally likely. Your friend wants your opinion. What outcomes will make you suspicious of the coin? Obviously, too many or too few heads (or tails) will cause you to be suspicious. You have to consider both possibilities if you don't know whether the coin is biased in favor of heads or tails. On the other hand, if you know that the coin would be rigged only in favor of heads, because that's what the coin's owner always bets on, you can ignore the possibility of getting too few heads.

Returning to the Noted Physician example, you are interested in both possibilities—too few and too many cures. That's because it's possible that the new treatment may work worse than the standard, and you want to know that. If there is a reason why the new treatment can't be worse—for example, if it involves adding meditation to the standard treatment—you can restrict your attention to cure rates at least as large as the one observed. ■■■

NORUŠIS, Marija J. 1998. *Guide to Data Analysis*. Upper Saddle River, Prentice Hall.

The Effect of Sample Size

As you saw above, when the true cure rate is 50%, there's a good chance that anywhere from 3 to 7 patients could be cured in a sample of 10. Most of the outcomes that can occur would not be considered unusual, because they could reasonably occur if the true cure rate is 50%. What's more, if the new treatment results in a cure rate of 60% or 70%, you probably would not detect the improvement, since many sample rates that are compatible with true rates of 60 or 70% are also compatible with the 50% rate. That means that based on a sample of only 10 patients, it's very difficult to evaluate a new treatment.

? *Can you ever tell from a sample of just 10 patients that a new treatment is better?* Yes. Since the existence of one little green man could convince you that there's life on Mars, similarly, 10 cures of a previously incurable disease could convince you that it's worth pursuing your treatment. It all depends on how unlikely your results are. ■■■

To see what effect sample size has on your ability to evaluate the physician's claim, consider what happens if you take samples of 40 patients, instead of just 10, from the same population with a cure rate of 50%. The results of this computer experiment are shown in Figure 9.3. (Note that each stem in the plot is now divided into two rows.) When you compare Figure 9.3 with Figure 9.1, you see that the values are much closer to 50% than before. Values greater than 60% or less than 40% are now noticeably less likely. These rates were not particularly unusual when you had samples of 10 patients. Based on Figure 9.3, you would estimate your chance of finding a sample rate of 70% or more or 30% or less when the true rate is 50% to be about 3 in 500, 0.6%. That means that only about 1 in 200 times would such a cure rate occur if the new treatment doesn't differ from the standard treatment.

In summary, when you have samples of 40 cases, an observed rate of 70% or more, or 30% or less, is possible, but not very likely when the true population rate is 50%. If the physician sees the same cure rate of 70% based on a sample of 40 patients, you would be more likely to believe that perhaps she's onto something. Her results really would be unusual when the true cure rate is 50%.

? *Just how unusual does "unusual" need to be?* The rule of thumb that is usually used to characterize results as unusual is a probability of 5% or less. That is, if results as extreme or more extreme than those observed are expected to occur in 5 (or fewer) samples out of 100, the results are considered unusual, or statistically significant. ■■■

To obtain this stem-and-leaf plot, select the variable cured40 in the Explore dialog box.

Figure 9.3 Stem-and-leaf plot of percentage cured for sample size 40

```
Frequency    Stem &  Leaf
     7.00       3 .  22&
    28.00       3 .  555777777
    72.00       4 .  000000002222222222222222
   114.00       4 .  55555555555555555577777777777777777777
   128.00       5 .  0000000000000000000222222222222222222222222
    93.00       5 .  5555555555555555577777777777777
    47.00       6 .  000000000002222
     8.00       6 .  55&
     3.00       7 .  0

 Stem width:     10.00
 Each leaf:       3 case(s)

 & denotes fractional leaves.
```

Cure rates cluster more tightly around 50% than in Figure 9.1

Larger samples improve your chances of detecting a difference in the cure rates (if in fact there is one) because there is less variability in the possible outcomes. Consider Figure 9.4, which contains descriptive statistics for the distribution shown in Figure 9.3. The mean value is again close to 50%. The standard deviation, however, is much smaller than for samples of size 10. It is now 7.29%, compared to the standard deviation of 16.22% in Figure 9.2. There's a pattern in the way that sample size affects the variance of the sampling distribution of means. If you increase the sample size by a factor of four, the variance decreases by a factor of four. Since the standard deviation is the square root of the variance, it decreases by a factor of two.

Figure 9.4 Descriptive statistics for samples of size 40

	N	Minimum	Maximum	Mean	Std. Deviation
CURED40	500	30.00	70.00	49.8350	7.2864
Valid N (listwise)	500				

The standard error of the mean is much smaller than for samples of size 10

NORUŠIS, Marija J. 1998. *Guide to Data Analysis*. Upper Saddle River, Prentice Hall.

The Binomial Test

In the previous example, you estimated the probability of various outcomes of an experiment from a stem-and-leaf plot obtained by repeated samples from the same population. The reason for doing it this way is to show you that when you take a sample from a population, the value you calculate for a statistic such as the mean is one of many possible values you can obtain. The possible values have a distribution—the sampling distribution of the statistic. Results vary from sample to sample, and you must take this variability into account when drawing conclusions about the population based on results observed from a sample.

Fortunately, in most situations, you don't personally need to determine the possible outcomes and their likelihoods by performing computer experiments. These can be mathematically calculated for you by SPSS. For example, you can use the **binomial test** to determine whether an observed cure rate is unlikely if the true rate is 50%. Your goal is to compare your experiment's success rate to a standard or usual rate. You observe the outcome of interest for a sample of subjects or objects.

To use the binomial test, your experiment or study must have only two possible outcomes, such as cured/not cured, pass/fail, buy/not buy, defective/not defective, and so on. All of the observations must be independent, and the probability of success must be the same for each member of the sample population.

? *What do you mean by independent?* For observations to be independent, one subject's response can't influence that of another. For example, if students collaborate on an exam, their scores are not independent. One student's results influence those of another. If you make multiple observations on the same subject, the observations are similarly not independent. Curing the same patient from 10 bouts of a disease is not equivalent to curing 10 patients from 1 bout. The 10 observations from a single patient are not independent. ■■■

Figure 9.5 shows the results of the binomial test for the 10-subject experiment. You see that there are 10 cases, 7 of which are coded 1, indicating a cure, and 3 of which are coded 0, indicating no cure. The population value that you want to test against (0.5) is labeled *Test Prop.* The proportion of successes in the sample, 0.7, is labeled *Observed Prop.* The probability of obtaining results as extreme or more extreme than the ones you observe in your sample, when the true probability of a cure is 0.5, is labeled *Exact Sig. (2-tailed)*.

For instructions on how to obtain a binomial test, see "Binomial Test" on p. 339 in Chapter 17.

The observed significance level tells you that the probability of obtaining a cure rate of 70% or greater or 30% or less, when the true cure rate is 50%, is 0.34. (Note how close this exact probability is to your estimated probability of 0.386 from Figure 9.1.) Since the observed significance level is larger than 0.05, the usual frame of reference, you don't have enough evidence to believe that the physician has achieved a cure rate different from 50%. The sample with an observed cure rate of 70% is not particularly unusual if the true population cure rate is 50%. In fact, more than 34% of samples from this population are as unusual as the one sample that the physician observed.

Figure 9.5 Binomial test: Sample size 10

		Category	N	Observed Prop.	Test Prop.	Exact Sig. (2-tailed)
CURE	Group 1	1	7	.70	.50	.344
	Group 2	0	3	.30		
	Total		10	1.00		

There is a 34% chance of observing a cure rate as extreme as 70% when the true rate is 50%.

The results from the 40-patient experiment are shown in Figure 9.6. There are now 28 cases with the response of 1, and 12 cases with the response of 0, giving the same observed proportion of 0.70. The test proportion is unchanged at 0.50. The observed significance level is 0.018. That means that, with samples of size 40, you would expect to see samples as unusual as the one observed less than 2% of the time. (Again, this value is reasonably close to the empirical estimate of 0.6% from Figure 9.3.) If the physician finds a 70% cure rate based on 40 patients, you're much more likely to believe that the physician is doing better than the usual 50%.

NORUŠIS, Marija J. 1998. *Guide to Data Analysis*. Upper Saddle River, Prentice Hall.

Figure 9.6 Binomial test: Sample size 40

		Category	N	Observed Prop.	Test Prop.	Asymp. Sig. (2-tailed)
CURE	Group 1	1	28	.70	.50	.018[1]
	Group 2	0	12	.30		
	Total		40	1.00		

1. Based on Z Approximation.

Probability of results this extreme decreases to less than 2%

? *Would you embrace her cure based only on these results?* Of course not. A statistical analysis is useless if a study is poorly designed. Here are some important concerns: How were patients selected for inclusion in her study? Is there something about them that would make them more likely to be cured than those in the population at large? Were there objective criteria for establishing a cure, or was it a subjective judgment? Did the evaluator and/or the patient know that a new drug was being used?

The correct way to conduct an evaluation of a new treatment is to allocate patients randomly to two treatment groups. One receives the standard treatment, and the other receives the new one. Ideally, neither the patient nor the physician knows which treatment the patient is receiving. Evaluation is done, based on well-established criteria, by physicians who are unaware of which patients received which treatment. These precautions help to ensure that the results of the study measure what they were intended to measure. ■■■

Summary

What can you say about a population, based on the results observed in a random sample?

- When you take a sample from a population, you won't get the same results as you would if you had data for the entire population.
- The sampling distribution of a statistic tells you, for a particular sample size, about the distribution of all possible sample values of that statistic.
- From the sampling distribution of a statistic, you can tell if observed sample results are unusual under particular circumstances.
- As the sample size increases, the variability of statistics calculated from the sample decreases.
- The observed significance level is the probability of observing a sample difference at least as the large as the one observed, when there is no difference in the population.
- A binomial test is used to test the hypothesis that a variable comes from a binomial population with a specified probability of an event occurring. The variable can have only two values.