CHAPTER 4

Sampling Distributions and Hypothesis Testing

Objectives

To lay the groundwork for the procedures discussed in this book by examining the general theory of hypothesis testing and describing specific concepts as they apply to all hypothesis tests.

Contents

IN CHAPTER 2, we examined a number of different statistics and saw how they might be used to describe a set of data or to represent the frequency of the occurrence of some event. Although the description of the data is important and fundamental to any analysis, it is not sufficient to answer many of the most interesting problems we encounter. In a typical experiment, we might treat one group of people in a special way and want to see whether their scores differ from the scores of people in general. Or we might offer a treatment to one group but not to a control group and want to compare the means of the two groups on some variable. Descriptive statistics will not tell us, for example, whether the difference between a sample mean and a hypothetical population mean, or the difference between two obtained sample means, is small enough to be explained by chance alone or whether it represents a true difference that might be attributable to the effect of our experimental treatment(s).

Statisticians frequently use phrases such as "variability due to chance" or "sampling error" and assume that you know what they mean. Probably you do, but if you do not, you are headed for confusion in the remainder of this book unless we spend a minute clarifying the meaning of these terms. We will begin with a simple example.

In Chapter 3, we considered the distribution of Total Behavior Problem scores from the Achenbach Youth Self-Report form. Total Behavior Problem scores are normally distributed in the population (i.e., the complete population of such scores is approximately normally distributed) with a population mean (μ) of 50 and a population standard deviation (σ) of 10. We know that different children show different levels of problem behaviors and therefore have different scores. We also know that if we took a sample of children, their sample mean would probably not equal exactly 50. One sample of children might have a mean of 49, but a second sample might have a mean of 52.3. The actual sample means would depend on the particular children who happened to be included in the sample. This expected variability from sample to sample is what is meant when we speak of "variability due to chance." The phrase refers to the fact that statistics (in this case, means) obtained from samples naturally vary from one sample to another.

sampling error

Along the same lines, the term **sampling error** often is used in this context as a synonym for variability due to chance. It indicates that the numerical value of a sample statistic probably will be in error (i.e., will deviate from the parameter it is estimating) as a result of the particular observations that happened to be included in the sample. In this context, "error" does not imply carelessness or mistakes. In the case of behavior problems, one random sample might just happen to include an unusually obnoxious child, whereas another sample might happen to include an unusual number of relatively well-behaved children.

4.1 Two Simple Examples Involving Course Evaluations and Rude Motorists

One example that we will investigate when we discuss correlation and regression looks at the relationship between how students evaluate a course and the grade they expect to receive in that course. Many faculty feel strongly about this topic because even the best instructors turn to the semiannual course evaluation forms with some trepidation—perhaps the same amount of trepidation with which many students open their grade report form. Some faculty think that a course is good or bad independently of how well a student feels he or she will do in terms of a grade. Others feel that a student who seldom came to class and who will do poorly as a result will also unfairly rate the course as poor. Finally, there are those who argue that students who do well and experience success take something away from the course other than just a grade and that those students will generally rate the course highly. But the relationship between course ratings and student performance is an empirical question and, as such, can be answered by looking at relevant data. Suppose that in a random sample of 50 courses we find a general trend for students in a course in which they expect to do well to rate the course highly, and for students to rate courses in which they expect to do poorly as low in overall quality. How do we tell whether this trend in our small data set is representative of a trend among students in general or just an odd result that would disappear if we ran the study again? (For your own interest, make your prediction of what kind of results we will find. We will return to this issue later.)

A second example comes from a study by Doob and Gross (1968), who investigated the influence of perceived social status. They found that if an old, beat-up (low-status) car failed to start when a traffic light turned green, 84% of the time the driver of the second car in line honked the horn. However, when the stopped car was an expensive, high-status car, the following driver only honked 50% of the time. These results could be explained in one of two ways:

1. The difference between 84% in one sample and 50% in a second sample is attributable to sampling error (random variability among samples); therefore, we cannot conclude that perceived social status influences horn-honking behavior.
2. The difference between 84% and 50% is large and reliable. The difference is not attributable to sampling error; therefore, we conclude that people are less likely to honk at drivers of high-status cars.

Although the statistical calculations required to answer this question are different from those used to answer the one about course evaluations (because the first deals with relationships and the second deals with proportions), the underlying logic is fundamentally the same.

hypothesis testing

These examples of course evaluations and horn honking are two kinds of questions that fall under the heading of **hypothesis testing.** This chapter is intended to present the theory of hypothesis testing in as general a way as possible, without going into the specific techniques or properties of any particular test. I will focus largely on the situation involving differences instead of the situation involving relationships, but the logic is basically the same. (You will see additional material on examining relationships in Chapter 9.) I am very deliberately glossing over details of computation because my purpose is to explore the concepts of hypothesis testing without involving anything but the simplest technical details.

We need to be explicit about what the problem is here. The reason for having hypothesis testing in the first place is that data are ambiguous. Suppose that we want to decide whether larger classes receive lower student ratings. We all know that some large classes are terrific, and others are really dreadful. Similarly, there are both good and bad small classes. So if we collect data on large classes, for example, the mean of several large classes will depend to some extent on which large courses just happen to be included in our sample. If we reran our data collection with a new random sample of large classes, that mean would almost certainly be different. A similar situation applies for small classes. When we find a difference between the means of samples of large and small classes, we know that the difference would come out slightly differently if we collected new data. So a difference between the means is ambiguous. Is it greater than zero because large classes are worse than small ones, or because of the particular samples we happened to pick? Well, if the difference is quite large, it probably reflects differences between small and large classes. If it is quite small, it probably reflects just random noise. But how large is "large" and how small is "small"? That is the problem we are beginning to explore, and that is the subject of this chapter.

If we are going to look at either of the two examples laid out earlier, or at a third one to follow, we need to find some way of deciding whether we are looking at a small chance fluctuation between the horn-honking rates for low- and high-status cars or a difference that is sufficiently large for us to believe that people are much less likely to honk at those they consider higher in status. If the differences are small enough to attribute to chance variability, we may well not worry about them further. On the other hand, if we can rule out chance as the source of the difference, we probably need to look further. This decision about chance is what we mean by hypothesis testing.

4.2 Sampling Distributions

In addition to course evaluations and horn honking, we will add a third example, which is one to which we can all relate. It involves those annoying people who spend what seems to us an unreasonable amount of time vacating the parking space we are waiting for. Ruback and Juieng (1997) ran a simple study in which they divided drivers into two groups of 100 participants each—those who had someone waiting for their space and those who did not. Ruback and Juieng then recorded the amount of time that it took the driver to leave the parking space. For those drivers who had no one waiting, it took an average of 32.15 seconds to leave the space. For those who did have someone waiting, it took an average of 39.03 seconds. For each of these groups, the standard deviation of waiting times was 14.6 seconds. Notice that a driver took 6.88 seconds longer to leave a space when someone was waiting for it. (If you think about it, 6.88 seconds is a long time if you are the person doing the waiting.)

There are two possible explanations here. First, it is entirely possible that having someone waiting doesn't make any difference in how long it takes to leave a space, and that normally drivers who have no one waiting for them take, on average, the same length of time as do drivers who have someone waiting. In that case, the difference that we found is just a result of the particular samples we happened to obtain. What we are saying here is that if we had whole populations of drivers in each of the two conditions, the populations' means (μ_{nowait} and μ_{wait}) would be identical and any difference we find in our samples is sampling error. The alternative explanation is that the population means really are different and that people actually do take longer to leave a space when there is someone waiting for it. If the sample means had come out to be 32.15 and 32.18, you and I would probably side with the first explanation—or at least not be willing to reject it. If the means had come out to be 32.15 and 59.03, we would probably be likely to side with the second explanation—having someone waiting actually makes a difference. But the difference we found is actually somewhere in between, and we need to decide which explanation is more reasonable.

sampling distributions

We want to answer the question "Is the obtained difference too great to be attributable to chance?" To do this, we have to use what are called **sampling distributions,** which tell us specifically what degree of sample-to-sample variability we can expect by chance as a function of sampling error.

The most basic concept underlying all statistical tests is the sampling distribution of a statistic. It is fair to say that if we did not have sampling distributions, we would not have any statistical tests. Roughly speaking, sampling distributions tell us what values we might (or might not) expect to obtain for a particular statistic under a set of predefined conditions (e.g., what the sample differences between our two samples might be expected to be *if* the true means of the populations from which those samples came are equal?) In addition, the standard deviation of that distribution of differences between sample means (known as the "standard error" of the distribution) reflects the variability that we would expect to find in the values of that statistic (differences between means) over repeated trials. Sampling distributions provide the opportunity to evaluate the likelihood (given the value of a sample statistic) that such predefined conditions actually exist.

Basically, the sampling distribution of a statistic can be thought of as the distribution of values obtained for that statistic over repeated sampling (i.e., running the experiment, or drawing samples, an unlimited number of times). Sampling distributions are almost always derived mathematically, but it is easier to understand what they represent if we consider how they could, in theory, be derived empirically with a simple sampling experiment.

sampling distribution of the differences between means

We will take as an illustration the **sampling distribution of the differences between means** because it relates directly to our example of waiting times in parking lots. The sampling distribution of differences between means is the distribution of differences between means of an infinite number of random samples drawn under certain specified conditions (e.g., under the condition that the true means of our populations are equal). Suppose we have two populations with known means and standard deviations (Here we will suppose that the two population means are 35 and the standard deviation is 15, though what the values are is not critical to the logic of our argument.) Further suppose that we draw a very large number (theoretically an infinite number) of pairs of random samples from these populations, each sample consisting of 100 scores. For each sample we will calculate its sample mean and then the difference between the two means in that draw. When we finish drawing all the pairs of samples, we will plot the distribution of these differences. Such a distribution would be a sampling distribution of the difference between means and might look like the one presented in Figure 4.1. The center of this distribution is at 0.0, because we expect that, on average, differences between sample means will be 0.0. (The individual means themselves will be roughly 35.) We can see from this figure that differences between sample means of approximately −1.5 and 1.5, for example, are quite likely to occur when we sample from identical populations. We also can see that it is extremely unlikely that we would draw samples from these populations that differ by 4.5 or more. Knowing the kinds of values to expect for the difference of means of samples drawn from these populations allows us to turn the question around and ask whether an obtained sample mean difference can be taken as evidence in favor of the hypothesis that we actually are sampling from identical populations—or populations with the same mean.

Notice here that the most common event we would find in drawing pairs of samples is that the means don't differ. $(\mu_{\text{nowait}} - \mu_{\text{wait}}) = 0$. (That is the mode [and the mean] of that distribution.) It is also fairly common to find differences of 1.5 or 2, though it is rare to find differences of 4.5.

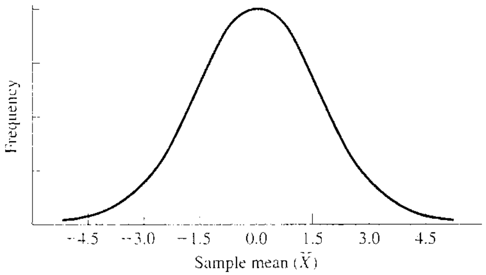

Figure 4.1 Distribution of difference between means, each based on 100 scores

4.3 Theory of Hypothesis Testing

Preamble

One of the major ongoing discussions in statistics in the behavioral sciences relates to hypothesis testing. The logic and theory of hypothesis testing has been debated for at least 75 years, but recently that debate has intensified considerably. The exchanges on this topic have not always been constructive (referring to your opponent's position as "bone-headedly misguided," "a perversion of the scientific method," or "ridiculous" usually does not win them to your cause), but some real and positive changes have come as a result. The changes are sufficiently important that much of this chapter, and major parts of the rest of the book, have been rewritten to accommodate them.

The arguments about the role of hypothesis testing concern several issues. First, and most fundamental, is hypothesis testing a sensible procedure? I think that it is, and whether it is or isn't, the logic involved is related to so much of what we do, and is so central to what you will see in the experimental literature, that you have to understand it whether you approve of it or not. Second, what logic will we use for hypothesis testing? The dominant logic has been an amalgam of positions put forth by R. A. Fisher and by Neyman and Pearson, dating from the 1920s and 1930s. (This amalgam is one to which both Fisher and Neyman and Pearson would express deep reservations, but it has grown to be employed by many, particularly in the behavioral sciences.) We will discuss that approach first, but follow it by more recent conceptualizations that lead to roughly the same point, but do so in what many feel is a more logical and rational process. Third, and perhaps most importantly, what do we need to consider *in addition to* traditional hypothesis testing? Running a statistical test and declaring a difference to be statistically significant at "$p < .05$" is no longer sufficient. A hypothesis test can only suggest whether a relationship is reliable or it is not, or that a difference between two groups is likely to be the result of chance, or that it probably is not. In addition to running a hypothesis test, we need to tell our readers something about the difference itself, about confidence limits on that difference, and about the power of our test. This will involve a change in emphasis from earlier editions, and will affect how I describe results in the rest of the book. I think the basic conclusion is that simple hypothesis testing, no matter how you do it, is important, but it is not enough. If the debate has done nothing else, getting us to that point has been very important. You can see that we have a lot to cover, but once you understand the positions and the proposals, you will have a better grasp of the issues than most people in your field.

The American Psychological Association recently put together a task force to look at the general issue of hypothesis tests, and its report is now available (Wilkinson, 1999; see also http://www.loyola.edu/library/ref/articles/Wilkinson.pdf). Further discussion of this issue was included in an excellent paper by Nickerson (2000). These two documents do a very effective job of summarizing current thinking in the field. These recommendations have influenced the coverage of material in this book, and you will see more frequent references to confidence limits and effect size measures than you would have seen in previous editions.

The Traditional Approach to Hypothesis Testing

For the next several pages, we will consider the traditional treatment of hypothesis testing. This is the treatment that you will find in almost any statistics text and is something that you need to fully understand. The concepts here are central to what we mean by hypothesis testing, no matter who is speaking about it.

We have just been discussing sampling distributions, which lie at the heart of the treatment of research data. We do not go around obtaining sampling distributions, either mathematically or empirically, simply because they are interesting to look at. We have important reasons for doing so. The usual reason is that we want to test some hypothesis. Let's go back to the sampling distribution of differences in mean times that it takes people to leave a parking space. We want to test the hypothesis that the obtained difference between sample means could reasonably have arisen had we drawn our samples from populations with the same mean. This is another way of saying that we want to know whether the mean departure time when someone is waiting is different from the mean departure time when there is no one waiting. One way we can test such a hypothesis is to have some idea of the probability of obtaining a difference in sample means as extreme as 6.88 seconds *if* we actually sampled observations from populations with the same mean. The answer to this question is precisely what a sampling distribution is designed to provide.

Suppose we obtained (constructed) the sampling distribution plotted in Figure 4.1. Suppose further, for the sake of argument, that our sample mean difference was 1.5 seconds and that we then determined from the sampling distribution that the probability of a sample mean difference as high as 1.5 seconds is .16. (How we determine this probability is not important here.) Our reasoning could then go as follows: "If we did in fact sample from populations with the same mean, the probability of obtaining a sample mean difference as high as 1.5 seconds is .16—that is not a terribly high probability, but it certainly isn't a low probability event. Because a sample mean difference at least as great as 1.5 is often obtained from populations with equal means, we have no reason to doubt that our two samples came from such populations."

Alternatively, suppose we obtained a sample mean difference of 10 seconds and calculated from the sampling distribution that the probability of a sample mean difference as large as 10, when the population means are equal, was only .0008. Our argument could then go like this: "*If* we did obtain our samples from populations with equal means, the probability of obtaining a sample mean difference as large as 10 is only .0008—an unlikely event. Because a sample mean difference that large is unlikely to be obtained from such populations, we can reasonably conclude that these samples probably came from populations with different means."

It is important to realize the steps in this example because the logic is typical of most tests of hypotheses. The actual test consisted of several stages:

research hypothesis

1. We wanted to test the hypothesis, often called the **research hypothesis**, that people backing out of a parking space take longer when someone is waiting.
2. We obtained random samples of behaviors under the two conditions.

null hypothesis

3. We set up the hypothesis (called the **null hypothesis**, H_0) that the samples were drawn from populations with the same means. This hypothesis states that leaving times do not depend on whether someone is waiting.
4. We then obtained the sampling distribution of the differences between means under the assumption that H_0 (the null hypothesis) is true (i.e., we obtained the sampling distribution of the differences between means when the population means are equal.)
5. Given the sampling distribution, we calculated the probability of a mean difference *at least as large* as the one we actually obtained between the means of our two samples.
6. On the basis of that probability, we made a decision: either to reject or fail to reject H_0. Because H_0 states the means of the populations are equal, rejection of H_0 represents a belief that they are unequal, although the actual value of the difference in population means remains unspecified.

The preceding discussion is oversimplified in the sense that in practice we also would need to consider (either directly or by estimation) the value of σ^2, the population variance(s), and N, the sample size(s). But again, those are specifics we can deal with when the time comes. The logic of the approach is representative of the logic of most, if not all, statistical tests.

1. Begin with a research hypothesis.
2. Set up the null hypothesis.
3. Construct the sampling distribution of the particular statistic on the assumption that H_0 is true.
4. Collect some data.
5. Compare the sample statistic to that distribution.
6. Reject or retain H_0, depending on the probability, under H_0, of a sample statistic as extreme as the one we have obtained.

The First Stumbling Block

I probably slipped something past you there, and you need to at least notice. This is one of the very important issues that motivates the fight over hypothesis testing, and it is something that you need to understand even if you can't do much about it. What I imagine that you would like to know is "what is the probability that the null hypothesis (drivers don't take longer when people are waiting) is true *given* the data we obtained?" But that is not what I gave you, and it is not what I am going to give you in the future. I gave you the answer to a different question, which is "what is the probability that I would have obtained these data *given* that the null hypothesis is true?" I don't know how to give you an answer to the question you would like to answer—not because I am a terrible statistician, but because the answer is much too difficult in most situations. However, the answer that I did give you is still useful—and is used all the time. When the police ticket a driver for drunken driving because he can't drive in a straight line and can't speak coherently, they are saying, "*if he were sober*, he would not behave this way. Because he behaves this way, we will conclude that he is not sober." This logic remains central to most approaches to hypothesis testing.

4.4 The Null Hypothesis

As we have seen, the concept of the null hypothesis plays a crucial role in the testing of hypotheses. People frequently are puzzled by the fact that we set up a hypothesis that is directly counter to what we hope to show. For example, if we hope to demonstrate the research hypothesis that college students do not come from a population with a mean self-confidence score of 100, we immediately set up the null hypothesis that they do. Or if we hope to demonstrate the validity of a research hypothesis that the means (μ_1 and μ_2) of the populations from which two samples are drawn are different, we state the null hypothesis that the population means are the same (or, equivalently, $\mu_1 - \mu_2 = 0$). (The term "null hypothesis" is most easily seen in this second example, in which it refers to the hypothesis that the difference between the two population means is zero, or *null*—some people call this the *nil null*, but that complicates the issue too much). We use the null hypothesis for several reasons. The philosophical argument, put forth by Fisher when he first introduced the concept, is that we can never prove something to be true, but we can prove something to be false. Observing 3,000 people with two arms does not prove the statement, "Everyone has two arms." However, finding one person with three arms does disprove the original statement beyond any shadow of a doubt. Although one might argue with Fisher's basic position—and many people have—the null hypothesis retains its dominant place in statistics.

alternative hypothesis

A second and more practical reason for employing the null hypothesis is that it provides us with the starting point for any statistical test. Consider the case in which you want to show that the mean self-confidence score of college students is greater than 100. Suppose further that you were granted the privilege of proving the truth of some hypothesis. What hypothesis are you going to test? Should you test the hypothesis that $\mu = 101$, or maybe the hypothesis that $\mu = 112$, or how about $\mu = 113$? The point is that in almost all research in the behavioral sciences we do not have a *specific* **alternative** (research) **hypothesis** in mind, but without one we cannot construct the sampling distribution we need. (This was one of the arguments raised against the original approach of Neyman and Pearson because they often spoke as if there were a specific alternative hypothesis to be tested, rather than just the diffuse negation of the null.) However, if we start off by assuming H_0: $\mu = 100$, we can immediately set about obtaining the sampling distribution for $\mu = 100$ and then, if our data are convincing, reject that hypothesis and conclude that the mean score of college students is greater than 100, which is what we wanted to show in the first place.

Statistical Conclusions

When the data differ markedly from what we would expect if the null hypothesis were true, we simply reject the null hypothesis and there is no particular disagreement about what our conclusions mean—we conclude that the null hypothesis is false. (This is not to suggest that we still don't need to tell our readers more about what we have found.) The interpretation is murkier and more problematic, however, when the data do not lead us to reject the null hypothesis. How are we to interpret a nonrejection? Shall we say that we have "proved" the null hypothesis to be true? Or shall we claim that we can "accept" the null, or that we shall "retain" it, or that we shall "withhold judgment"?

The problem of how to interpret a nonrejected null hypothesis has plagued students in statistics courses for more than 50 years, and it will probably continue to do so (but see Section 4.10). The idea that if something is not false then it must be true is too deeply ingrained in common sense to be dismissed lightly.

The one thing on which all statisticians agree is that we can never claim to have "proved" the null hypothesis. As was pointed out, the fact that the next 3,000 people we meet all have two arms certainly does not prove the null hypothesis that all people have two arms. In fact, we know that many perfectly normal people have fewer than two arms. Failure to reject the null hypothesis often means that we have not collected enough data.

The issue is easier to understand if we use a concrete example. Wagner, Compas, and Howell (1988) conducted a study to evaluate the effectiveness of a program for teaching high-school students to deal with stress. If this study found that students who participate in such a program had significantly fewer stress-related problems than did students in a control group who did not have the program, then we could, without much debate, conclude that the program was effective. However, if the groups did not differ at some predetermined level of statistical significance, what could we conclude?

We know we cannot conclude from a nonsignificant difference that we have proved that the mean of a population of scores of treatment subjects is the same as the mean of a population of scores of control subjects. The two treatments may lead to subtle differences that we were not able to identify conclusively with our relatively small sample of observations.

Fisher's position was that a nonsignificant result is an inconclusive result. For Fisher, the choice was between rejecting a null hypothesis and suspending judgment. He would have argued that a failure to find a significant difference between conditions could result from the fact that the students who participated in the program handled stress only *slightly* better than did control subjects, or that they handled it only slightly less well, or that there was no difference between the groups. For Fisher, a failure to reject H_0 merely means that our data are insufficient to allow us to choose among these three alternatives; therefore, we must suspend judgment. You will see this position return when we shortly discuss a proposal by Jones and Tukey (2000).

A slightly different approach was taken by Neyman and Pearson (1933), who took a much more pragmatic view of the results of an experiment. In our example, Neyman and Pearson would be concerned with the problem faced by the school board, who must decide whether to continue spending money on this stress-management program that we are providing for them. The school board would probably not be impressed if we told them that our study was inconclusive and then asked them to give us money to continue operating the program until we had sufficient data to state confidently whether or not the program was beneficial (or harmful). In the Neyman–Pearson position, one either rejects or *accepts* the null hypothesis. But when we say that we "accept" a null hypothesis, however, we do not mean that we take it to be proven as true. We simply mean that we will *act as if* it is true, at least until we have more adequate data. Whereas given a nonsignificant result, the ideal school board from Fisher's point of view would continue to support the program until we finally were able to make up our minds, the school board with a Neyman–Pearson perspective would conclude that the available evidence is not sufficient to defend continuing to fund the program, and they would cut off our funding.

This discussion of the Neyman–Pearson position has been much oversimplified, but it contains the central issue of their point of view. The debate between Fisher on the one hand and Neyman and Pearson on the other was a lively (and rarely civil) one, and present practice contains elements of both viewpoints. Most statisticians prefer to use phrases such as "retain the null hypothesis" and "fail to reject the null hypothesis" because these make clear the tentative nature of a nonrejection. These phrases have a certain Fisherian ring to them. On the other hand, the important emphasis on Type II errors (failing to reject a *false* null hypothesis), which we will discuss in Section 4.7, is clearly an essential feature of the Neyman–Pearson school. If you are going to choose between two alternatives (accept or reject), then you have to be concerned with the probability of falsely accepting as well as that of falsely rejecting the null hypothesis. Fisher would never accept a null hypothesis in the first place, so he did not need to worry much about the probability of accepting a false one.[1] We will return to this whole question in Section 4.10, where we will consider an alternative approach, after we have developed several other points. First, however, we need to consider some basic information about hypothesis testing so as to have a vocabulary and an example with which to go further into hypothesis testing. This information is central to any discussion of hypothesis testing under any of the models that have been proposed.

4.5 Test Statistics and Their Sampling Distributions

We have been discussing the sampling distribution of the mean, but the discussion would have been essentially the same had we dealt instead with the median, the variance, the range, the correlation coefficient (as in our course evaluation example), proportions (as in our horn-honking example), or any other statistic you care to consider. (Technically, the shapes of these distributions would be different, but I am deliberately ignoring such issues in this chapter.) The statistics just mentioned usually are referred to as **sample statistics** because they describe characteristics of samples. There is a whole different class of statistics called **test statistics,** which are associated with specific statistical procedures and which have their own sampling distributions. Test statistics are statistics such as t, F, and χ^2, which you may have run across in the past. (If you are not familiar with them, don't worry—we will consider them separately in later chapters.) This is not the place to go into a detailed explanation of any test statistics. I put this chapter where it is because I don't want readers to think that they are supposed to worry about technical issues. This chapter is the place, however, to point out that the sampling distributions for test statistics are obtained and used in essentially the same way as the sampling distribution of the mean.

sample statistics

test statistics

As an illustration, consider the sampling distribution of the statistic t, which will be discussed in Chapter 7. For those who have never heard of the t test, it is sufficient to say that the t test is often used, among other things, to determine whether two samples were drawn from populations with the same means. Let μ_1 and μ_2 represent the means of the populations from which the two samples were drawn. The null hypothesis is the hypothesis that the two population means are equal, in other words, H_0: $\mu_1 = \mu_2$ (or $\mu_1 - \mu_2 = 0$). If we were extremely patient, we could empirically obtain the sampling distribution of t when H_0 is true by drawing an infinite number of pairs of samples, all from two identical populations, calculating t for each pair of samples (by methods to be discussed later), and plotting the resulting values of t. In that case, H_0 must be true because we forced it to be true by drawing the samples from identical populations. The resulting distribution is the sampling distribution of t when H_0 is true. If we later had two samples that produced a particular value of t, we would test the null hypothesis by comparing our sample t to the sampling distribution of t. We would reject the null hypothesis if our obtained t did not look like the kinds of t values that the sampling distribution told us to expect when the null hypothesis is true.

I could rewrite the preceding paragraph, substituting χ^2, or F, or any other test statistic in place of t, with only minor changes dealing with how the statistic is calculated. Thus, you can see that all sampling distributions can be obtained in basically the same way (calculate and plot an infinite number of statistics by sampling from identical populations).

4.6 Using the Normal Distribution to Test Hypotheses

Much of the discussion so far has dealt with statistical procedures that you do not yet know how to use. I did this deliberately to emphasize the point that the logic and the calculations behind a test are two separate issues. However, we now can use what you already know about the normal distribution to test some simple hypotheses. In the process, we can deal with several fundamental issues that are more easily seen by use of a concrete example.

An important use of the normal distribution is to test hypotheses, either about individual observations or about sample statistics such as the mean. In this chapter, we will deal with individual observations, leaving the question of testing sample statistics until later chapters. Note, however, that in the usual case we test hypotheses about sample statistics such as the mean rather than about individual observations. I am starting with an example of an individual observation because the explanation is somewhat clearer. Because we are dealing with only single observations, the sampling distribution invoked here will be the distribution of individual scores (rather than the distribution of means or differences between means). The basic logic is the same, and we are using an example of individual scores only because it simplifies the explanation and is something with which you have had experience.

Psychologists who study neurological functioning have a battery of tests at their disposal. A common test is simple finger tapping speed, which is useful for diagnosing hidden

[1] Excellent discussions of the differences between the theories of Fisher on the one hand, and Neyman and Pearson on the other can be found in Chapter 4 of Gigerenzer, Swijtink, Porter, Daston, Beatty, & Krüger (1989), Lehman (1993), and Oakes (1990). The central issues involve the concept of probability, the idea of an infinite population or infinite resampling, and the choice of a critical value, among other things. The controversy is far from a simple one.

brain damage. (For example, people with brain damage to the dorsal lateral frontal lobes are especially slow in the speed of finger tapping, but are often unaware of their loss of behavioral competency.) For a simple example, assume we know that the mean rate of finger tapping of normal healthy adults is 100 taps in 20 seconds, with a standard deviation of 20, and that tapping speeds are normally distributed in the population. We already know that the tapping rate is slower among people with dorsal lateral frontal lobe damage. Suppose that an individual has just been sent to us who taps at a rate of 70 taps in 20 seconds. Is his score sufficiently below the mean for us to assume that he did not come from a population of neurologically healthy people? This situation is diagrammed in Figure 4.2, in which the arrow indicates the location of our piece of data (the person's score).

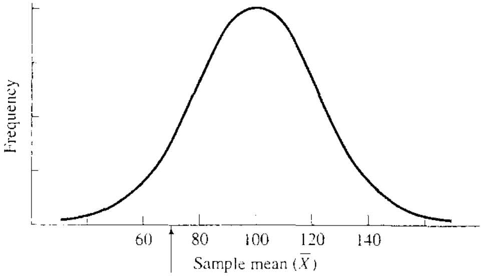

Figure 4.2 Location of a person's tapping score on a distribution of scores of neurologically healthy people

The logic of the solution to this problem is the same as the logic of hypothesis testing in general. We begin by assuming that the individual's score *does* come from the population of healthy scores. This is the **null** hypothesis (H_0). If H_0 is true, we automatically know the mean and the standard deviation of the population from which he was supposedly drawn (100 and 20, respectively). With this information, we are in a position to calculate the probability that a score *as low as* his would be obtained from this population. If the probability is very low, we can reject H_0 and conclude that he did not come from the healthy population. Conversely, if the probability is not particularly low, then the data represent a reasonable result under H_0, and we would have no reason to doubt its validity and thus no reason to doubt that the person is healthy. Keep in mind that we are not interested in the probability of a score *equal* to 70 (which, because the distribution is continuous, would be infinitely small) but, rather, in the probability that the score would be at least as low as (i.e., less than or equal to) 70.

The individual had a score of 70. We want to know the probability of obtaining a score *at least as low as* 70 if H_0 is true. We already know how to find this—it is the area below 70 in Figure 4.2. All we have to do is convert 70 to a z score and then refer to Appendix z (page 694).

$$z = \frac{X - \mu}{\sigma} = \frac{70 - 100}{20} = \frac{-30}{20} = -1.5$$

From Appendix z, we can see that the probability of a z score of -1.5 or below is .0668. (Locate $z = 1.50$ in the table and then read across to the column headed "Smaller Portion.")

decision-making

At this point, we have to become involved in the **decision-making** aspects of hypothesis testing. We must decide whether an event with a probability of .0668 is sufficiently unlikely to cause us to reject H_0. Here we will fall back on arbitrary conventions that have been established over the years. The rationale for these conventions will become clearer as we go along, but for the time being keep in mind that they are merely conventions. One convention calls for rejecting H_0 if the probability under H_0 is less than or equal to .05 ($p \leq .05$), and another convention—one that is more conservative with respect to the probability of rejecting H_0—calls for rejecting H_0 whenever the probability under H_0 is less than or equal to .01. These values of .05 and .01 are often referred to as the **rejection level**, or the **significance level**, of the test. (When we say that a difference is statistically significant at the .05 level, we mean that a difference that large would occur less than 5% of the time if the null were true.) Whenever the probability obtained under H_0 is less than or equal to our predetermined significance level, we will reject H_0. Another way of stating this is to say that any outcome whose probability under H_0 is less than or equal to the significance level falls in the **rejection region** because such an outcome leads us to reject H_0.

rejection level (significance level)

rejection region

For the purpose of setting a standard level of rejection for this book, we will use the .05 level of statistical significance, keeping in mind that some people would consider this level to be too lenient.[2] For our particular example, we have obtained a probability value of $p = .0668$, which obviously is greater than .05. Because we have specified that we will not reject H_0 unless the probability of the data under H_0 is less than .05, we must conclude that we have no reason to decide that the person did not come from a population of healthy people.

More specifically, we conclude that a finger-tapping rate of 70 reasonably could have come from a population of scores with a mean equal to 100 and a standard deviation equal to 20. It is important to note that we have not shown that this person is healthy, but only that we have insufficient reason to believe that he is not. It may be that he is just acquiring the disease and therefore is not quite as different from normal as is usual for his condition. Or maybe he has the disease at an advanced stage but just happens to be an unusually fast tapper. This is an example of the fact that we can never say that we have proved the null hypothesis. We can conclude only that this person does not tap sufficiently slowly for an illness, if any, to be statistically detectable.

4.7 Type I and Type II Errors

Whenever we reach a decision with a statistical test, there is always a chance that our decision is the wrong one. Although this is true of almost all decisions, statistical or otherwise, the statistician has one point in her favor that other decision makers normally lack. She not only makes a decision by some rational process, but she can also specify the conditional probabilities of a decision's being in error. In everyday life, we make decisions with only subjective feelings about what is probably the right choice. The statistician, however, can state quite precisely the probability that she would make an erroneously rejection of H_0 if it were true. This ability to specify the probability of erroneously rejecting a true H_0 follows directly from the logic of hypothesis testing.

[2] The particular view of hypothesis testing described here is the classical one that a null hypothesis is rejected if the probability of obtained the data when the null hypothesis is true is less than the predefined significance level, and not rejected if that probability is greater than the significance level. Currently, a substantial body of opinion holds that such cut-and-dried rules are inappropriate and that more attention should be paid to the probability value itself. In other words, the classical approach (using a .05 rejection level) would declare $p = .051$ and $p = .150$ to be (equally) "statistically nonsignificant" and $p = .048$ and $p = .0003$ to be (equally) "statistically significant." The alternative view would think of $p = .051$ as "nearly significant" and $p = .0003$ as "very significant." Although this view has much to recommend it, especially given current trends to move away from only reporting statistical significance of results, it will not be wholeheartedly adopted here. Most computer programs do print out exact probability levels, and those values, when interpreted judiciously, can be useful. The difficulty comes in defining what is meant by "interpreted judiciously."

critical value

Consider the finger-tapping example, this time ignoring the score of the individual sent to us. The situation is diagrammed in Figure 4.3, in which the distribution is the distribution of scores from healthy subjects, and the shaded portion represents the lowest 5% of the distribution. The actual score that cuts off the lowest 5% is called the **critical value.** Critical values are those values of X (the variable) that describe the boundary or boundaries of the rejection region(s). For this particular example, the critical value is 67.

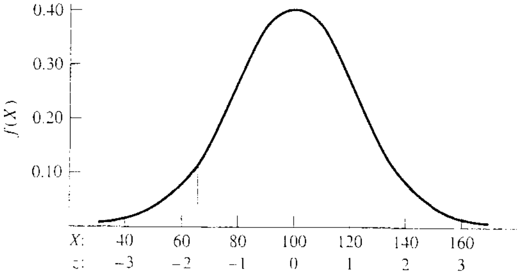

Figure 4.3 Lowest 5% of scores from clinically healthy people

Type I error

α (alpha)

If we have a decision rule that says to reject H_0 whenever an outcome falls in the lowest 5% of the distribution, we will reject H_0 whenever an individual's score falls in the shaded area; that is, whenever a score as low as his has a probability of .05 or less of coming from the population of healthy scores. Yet by the very nature of our procedure, 5% of the scores from perfectly healthy people will themselves fall in the shaded portion. Thus, if we actually have sampled a person who is healthy, we stand a 5% chance of his score being in the shaded tail of the distribution, causing us erroneously to reject the null hypothesis. This kind of error (rejecting H_0 when it is actually true) is called a **Type I error,** and its conditional probability (the probability of rejecting the null hypothesis given that it is true) is designated as **α (alpha)**, the size of the rejection region. In the future, whenever we represent a probability by α, we will be referring to the probability of a Type I error.

Keep in mind the "conditional" nature of the probability of a Type I error. I know that sounds like jargon, but what it means is that you should be sure you understand that when we speak of a Type I error we mean the probability of rejecting H_0 *given that it is true*. We are not saying that we will reject H_0 on 5% of the hypotheses we test. We would hope to run experiments on important and meaningful variables and, therefore, to reject H_0 often. But when we speak of a Type I error, we are speaking only about rejecting H_0 in those situations in which the null hypothesis happens to be true.

Type II error

β (beta)

You might feel that a 5% chance of making an error is too great a risk to take and suggest that we make our criterion much more stringent, by rejecting, for example, only the lowest 1% of the distribution. This procedure is perfectly legitimate, but realize that the more stringent you make your criterion, the more likely you are to make another kind of error—failing to reject H_0 when it is false and H_1 is true. This type of error is called a **Type II error,** and its probability is symbolized by **β (beta)**.

The major difficulty of Type II errors stems from the fact that if H_0 is false, we almost never know what the true distribution (the distribution under H_1) would look like for the population from which our data came. We know only the distribution of scores under H_0. Put in the present context, we know the distribution of scores from healthy people but not from nonhealthy people. It may be that people suffering from some neurological disease tap, on average, considerably more slowly than healthy people, or it may be that they tap, on average, only a little more slowly. This situation is illustrated in Figure 4.4, in which the distribution labeled H_0 represents the distribution of scores from healthy people (the set of observations expected under the null hypothesis), and the distribution labeled H_1 represents our hypothetical distribution of nonhealthy scores (the distribution under H_1). Remember that the curve H_1 is only hypothetical. We really do not know the location of the nonhealthy distribution, other than that it is lower (slower speeds) than the distribution of H_0. (I have arbitrarily drawn that distribution with a mean of 80 and a standard deviation of 20.)

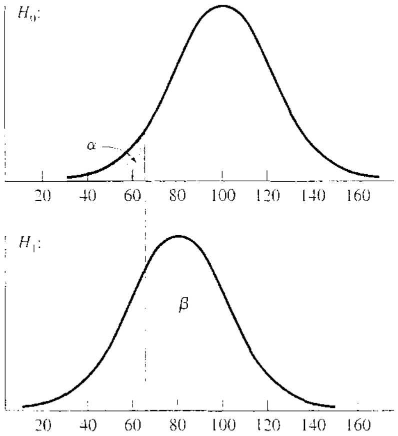

Figure 4.4 Areas corresponding to α and β for tapping speed example

The darkly shaded portion in the top half of Figure 4.4 represents the rejection region. Any observation falling in that area (i.e., to the left of about 67) would lead to rejection of the null hypothesis. If the null hypothesis is true, we know that our observation will fall in this area 5% of the time. Thus, we will make a Type I error 5% of the time.

The lightly shaded portion in the bottom half of Figure 4.4 represents the probability (β) of a Type II error. This is the situation of a person who was actually drawn from the nonhealthy population but whose score was not sufficiently low to cause us to reject H_0.

In the particular situation illustrated in Figure 4.4, we can actually calculate β by using the normal distribution to calculate the probability of obtaining a score *greater than* 67 (the critical value) if $\mu = 80$ and $\sigma = 20$. The actual calculation is not important for your understanding of β; because this chapter was designed specifically to avoid calculation, I will simply state that this probability (i.e., the area labeled β) is .74. Thus, for this example, 74% of the time when we have a person who is actually nonhealthy (i.e., H_1 is actually true), we will make a Type II error by failing to reject H_0 when it is false (as medical diagnosticians, we leave a lot to be desired).

From Figure 4.4, you can see that if we were to reduce the level of α (the probability of a Type I error) from .05 to .01 by moving the rejection region to the left, it would reduce the probability of Type I errors but would increase the probability of Type II errors. Setting α at .01 would mean that $\beta = .908$. You can see that there is room for debate about what level of significance to use. The decision rests primarily on your opinion concerning the relative importance of Type I and Type II errors for the kind of study you are conducting. If it were important to avoid Type I errors (such as telling someone that he has a disease when he

Table 4.1 Possible outcomes of the decision-making process

	True State of the World	
Decision	H_0 True	H_0 False
Reject H_0	Type I error $p = \alpha$	Correct decision $p = 1 - \beta =$ Power
Don't reject H_0	Correct decision $p = 1 - \alpha$	Type II error $p = \beta$

does not), then you would set a stringent (i.e., small) level of α. If, on the other hand, you want to avoid Type II errors (telling someone to go home and take an aspirin when in fact he needs immediate treatment), you might set a fairly high level of α. (Setting $\alpha = .20$ in this example would reduce β to .44.) Unfortunately, in practice most people choose an arbitrary level of α, such as .05 or .01, and simply ignore β. In many cases, this may be all you can do. (You will probably use the alpha level that your instructor recommends.) In other cases, however, there is much more you can do, as you will see in Chapter 8.

I should stress again that Figure 4.4 is purely hypothetical. I was able to draw the figure only because I arbitrarily decided that speeds of nonhealthy people were normally distributed with a mean of 80 and a standard deviation of 20. The calculated answers would be different if I had chosen to draw it with a mean of 70 or a standard deviation of 10. In most everyday situations, we do not know the mean and the variance of that distribution and can make only educated guesses, thus providing only crude estimates of β. In practice, we can select a value of μ under H_1 that represents the *minimum* difference we would like to be able to detect because larger differences will have even smaller βs.

power

From this discussion of Type I and Type II errors, we can summarize the decision-making process with a simple table. Table 4.1 presents the four possible outcomes of an experiment. The items in this table should be self-explanatory, but there is one concept—power—that we have not yet discussed. The **power** of a test is the probability of rejecting H_0 when it is actually false. Because the probability of *failing* to reject a false H_0 is β, then power must equal $1 - \beta$. Those who want to know more about power and its calculation will find power covered in Chapter 8.

4.8 One- and Two-Tailed Tests

The preceding discussion brings us to a consideration of one- and two-tailed tests. In our tapping example, we knew that nonhealthy subjects tapped more slowly than healthy subjects; therefore, we decided to reject H_0 only if a subject tapped too slowly. However, suppose our subject had tapped 180 times in 20 seconds. Although this is an exceedingly unlikely event to observe from a healthy subject, it did not fall in the rejection region, which consisted *solely* of low rates. As a result, we find ourselves in the position of not rejecting H_0 in the face of a piece of data that is very unlikely, but not in the direction expected.

The question then arises as to how we can protect ourselves against this type of situation (if protection is thought necessary). The answer is to specify before we run the experiment that we are going to reject a given percentage (say 5%) of the *extreme* outcomes, both those that are extremely high and those that are extremely low. But if we reject the lowest 5% and the highest 5%, then we would reject H_0 a total of 10% of the time when it is actually true, that is, $\alpha = .10$. We are rarely willing to work with α as high as .10 and prefer to see it set no higher than .05. The way to accomplish this is to reject the lowest 2.5% and the highest 2.5%, making a total of 5%.

one-tailed (directional) test

two-tailed (nondirectional) test

The situation in which we reject H_0 for only the lowest (or only the highest) tapping speeds is referred to as a **one-tailed**, or **directional**, **test**. We make a prediction of the direction in which the individual will differ from the mean, and our rejection region is located in only one tail of the distribution. (That makes sense when we know that brain damage is only associated with slow tapping speeds.) When we reject extremes in both tails, we have what is called a **two-tailed**, or **nondirectional**, **test**. It is important to keep in mind that although we gain something with a two-tailed test (the ability to reject the null hypothesis for extreme scores in either direction), we also lose something. A score that would fall in the 5% rejection region of a one-tailed test may not fall in the rejection region of the corresponding two-tailed test because now we reject only 2.5% in each tail.

In the finger-tapping example, the decision between a one- and a two-tailed test might seem reasonably clear-cut. We know that people with a given disease tap more slowly; therefore, we care only about rejecting H_0 for low scores—high scores have no diagnostic importance. In other situations, however, we do not know which tail of the distribution is important (or both are), and we need to guard against extremes in either tail. The situation might arise when we are considering a campaign to persuade children not to start smoking. We might find that the campaign leads to a decrease in the incidence of smoking. Or, we might find that campaigns run by adults to persuade children not to smoke simply make smoking more attractive and exciting, leading to an increase is the number of children smoking. In either case, we would want to reject H_0.

In general, two-tailed tests are far more common than one-tailed tests for several reasons. The investigator may have no idea what the data will look like and therefore has to be prepared for any eventuality. Although this situation is rare, it does occur in some exploratory work.

Another common reason for preferring two-tailed tests is that the investigators are reasonably sure the data will come out one way but want to cover themselves in the event that they are wrong. This type of situation arises more often than you might think. (Carefully formed hypotheses have an annoying habit of being phrased in the wrong direction, for reasons that seem so obvious after the event.) The smoking example is a case in point, where there is some evidence that poorly contrived antismoking campaigns actually do more harm than good. A frequent question that arises when the data may come out the other way around is, "Why not plan to run a one-tailed test and then, if the data come out the other way, just change the test to a two-tailed test?" This kind of question comes from people who have no intention of being devious but who just do not fully understand the logic of hypothesis testing. If you start an experiment with the extreme 5% of the left-hand tail as your rejection region and then turn around and reject any outcome that happens to fall in the extreme 2.5% of the right-hand tail, you are working at the 7.5% level. In that situation, you will reject 5% of the outcomes in one direction (assuming that the data fall in the desired tail), and you are willing also to reject 2.5% of the outcomes in the other direction (when the data are in the unexpected direction). There is no denying that 5% + 2.5% = 7.5%. To put it another way, would you be willing to flip a coin for an ice cream cone if I have chosen "heads" but also reserve the right to switch to "tails" after I see how the coin lands? Or would you think it fair of me to shout, "Two out of three!" when the coin toss comes up in your favor? You would object to both of these strategies, and you should. For the same reason, the choice between a one-tailed test and a two-tailed one is made *before* the data are collected. It is also one of the reasons that two-tailed tests are usually chosen.

Although the preceding discussion argues in favor of two-tailed tests, and although in this book we generally confine ourselves to such procedures, there are no hard-and-fast rules. The final decision depends on what you already know about the relative severity of different kinds of errors. It is important to remember that with respect to a given tail of a distribution, the difference between a one-tailed test and a two-tailed test is that the latter just uses a different cutoff. A two-tailed test at $\alpha = .05$ is more liberal than a one-tailed test at $\alpha = .01$.[3]

If you have a sound grasp of the logic of testing hypotheses by use of sampling distributions, the remainder of this course will be relatively simple. For any new statistic you encounter, you will need to ask only two basic questions:

1. How and with which assumptions is the statistic calculated?
2. What does the statistic's sampling distribution look like under H_0?

If you know the answers to these two questions, you can accomplish your test by calculating the test statistic for the data at hand and comparing the statistic to the sampling distribution. Because the relevant sampling distributions are tabled in the appendices, all you really need to know is which test is appropriate for a particular situation and how to calculate its test statistic. (Of course, there is more to statistics than just hypothesis testing, so perhaps I'm doing a bit of overselling here. There is a great deal to understanding the field of statistics beyond how to calculate, and evaluate, a specific statistical test. Calculation is the easy part, especially with modern computer software.)

4.9 What Does It Mean to Reject the Null Hypothesis?

conditional probabilities

One of the common problems that even well-trained researchers have with the null hypothesis is the confusion over what rejection really means. I mentioned this earlier when I discussed the fact that we calculate the probability of the data given that the null is true, rather than the probability of the null being true given the data. Suppose that we test a null hypothesis about the difference between two population means and reject it at $p = .045$. There is a temptation to say that such a result means that the probability of the null being true is .045. But that is *not* what this probability means. What we have shown is that *if the null hypothesis were true*, the probability of obtaining a difference between means as great as the difference we found is only .045. That is quite different from saying that the probability that the null is true is .045. What we are doing here is confusing the probability of the hypothesis given the data, and the probability of the data given the hypothesis. These are called **conditional probabilities**, and will be discussed in Chapter 5. The probability of .045 that we have here is the probability of the data given that H_0 is true, written $p(D \mid H_0)$—the vertical line is read "given." It is not the probability that H_0 is true given the data, written $p(H_0 \mid D)$. The best discussion of this issue that I have read is in an excellent paper by Nickerson (2000). Let me illustrate my major point with an example.

Suppose that I create a computer-generated example where I know for a fact that the data for one sample came from a population with a mean of 54.28, and the data for a second sample came from a population with a mean of 54.25. (It is very easy to use a program like SPSS to generate such samples.) Here I *know for a fact* that the null hypothesis is false. In other words, the probability that the null hypothesis is true is .00—that is, $p(H_0) = .00$. However, if I have two small samples I might happen to get a result such as 54.26 and 54.36, and that result would have a very high probability of occurring even in the situation where the null hypothesis is true and both means were, say, 54.28. Thus, the probability of the data given a true null hypothesis might be .75, for example, and yet we know that the probability that the null is really true is exactly .00. (Using probability terminology, we can write $p(H_0) = .00$ and $p(D \mid H_0) = .75$). Alternatively, assume that I created a situation where I know that the null is true. For example, I set up populations where both means are 54.00. It is easy to imagine getting samples with means of 53 and 54.5. If the null is really true, the probability of getting means this difference may be .33, for example. Thus, the probability that the null is true is fixed, by me, at 1.00, yet the probability of the data when the null is true is .33. (Using probability terminology again, we can write $p(H_0) = 1.00$ and $p(D \mid H_0) = .33$.) Notice that in both of these cases there is a serious discrepancy between the probability of the null being true and the probability of the data given the null. You will see several instances like this throughout the book whenever I sample data from known populations. Never confuse the probability value associated with a test of significant with the probability that the null hypothesis is true. They are very different things.

4.10 An Alternative View of Hypothesis Testing

What I have presented so far about hypothesis testing is the traditional approach. It is found in virtually every statistics text, and you need to be very familiar with it. However, there has recently been an interest in different ways of looking at hypothesis testing, and a new approach proposed by Lyle Jones and John Tukey (2000) avoids some of the problems of the traditional approach.

We will begin with an example comparing two population means that is developed further in Chapter 7. Adams, Wright, and Lohr (1996) showed a group of homophobic heterosexual males and a group of nonhomophobic heterosexual males a videotape of sexually explicit erotic homosexual images, and recorded the resulting level of sexual arousal in the participants. The researchers were interested in seeing whether there was a difference in sexual arousal between the two categories of viewers. (Notice that I didn't say which group they expected to come out with the higher mean, just that there would be a difference.)

The traditional hypothesis testing approach would set up the null hypothesis that $\mu_h = \mu_n$, where μ_h is the population mean for homophobic males, and μ_n is the population mean for nonhomophobic males. The traditional alternative (two-tailed) hypothesis is that $\mu_\eta \neq \mu_\nu$. Many people have pointed out that the null hypothesis in such a situation is never going to be true. It is not reasonable to believe that if we had a population of all homophobic males their mean would be exactly equal to the mean of the population all nonhomophobic males to an unlimited number of decimal places. Whatever the means are, they will certainly differ by *at least* some trivial amount. So we know before we begin that the null hypothesis is false, and we might ask ourselves why we are testing the null in the first place. (Many people have asked that question.)

Jones and Tukey (2000) and Harris (2005) have argued that we really have three possible hypotheses or conclusions we could draw—Jones and Tukey speak primarily of "conclusions." One is that $\mu_h < \mu_n$, another is that $\mu_h > \mu_n$, and the third is that $\mu_h = \mu_n$. This third hypothesis is the traditional null hypothesis, and we have just said that it is never going

[3] One of the reviewers of an earlier edition of this book made the case for two-tailed tests even more strongly: "It is my (minority) belief that what an investigator *expects to be true* has absolutely no bearing *whatsoever* on the issue of one- versus two-tailed tests. Nature couldn't care less what psychologists' theories predict, and will often show patterns/trends in the opposite direction. Since our goal is to know the truth (not to prove we are astute at predicting), our tests must always allow for testing *both* directions. I say *always* do two-tailed tests, and if you are worried about β, jack the sample size up a bit to offset the loss in power" (D. Bradley, personal communication, 1983). I am personally inclined toward this point of view. Nature is notoriously fickle, or else we are notoriously inept at prediction. On the other hand, a second reviewer takes exception to this position. While acknowledging that Bradley's point is well considered, Rodgers, engaging in a bit of hyperbole, argues, "To generate a theory about how the world works that implies an expected direction of an effect, but then to hedge one's bet by putting some (up to ⅓) of the rejection region in the tail other than that predicted by the theory, strikes me as both scientifically dumb and slightly unethical. . . . Theory generation and theory testing are much closer to the proper goal of science than truth searching, and running one-tailed tests is quite consistent with those goals" (J. Rodgers, personal communication, 1986). Neither Bradley nor I would accept the judgment of being "scientifically dumb and slightly unethical," but I presented the two positions in juxtaposition because doing so gives you a flavor of the debate. Obviously there is room for disagreement on this issue.

to be true when means are carried to enough decimal places. These three hypotheses lead to three courses of action. If we test the first ($\mu_h < \mu_n$) and reject it, we conclude that homophobic males are more aroused than nonhomophobic males. If we test the second ($\mu_h > \mu_n$) and reject it, we conclude that homophobic males are less aroused than nonhomophobic males. If we cannot reject either of those hypotheses, we conclude that we have insufficient evidence to make a choice—the population means are almost certainly different, but we don't know which is the larger.

The difference between this approach and the traditional one may seem minor, but it is important. In the first place, when Jones and Tukey tell us something, we should definitely listen. These are not two guys who just got out of graduate school—they are two very highly respected statisticians. (If there were a Nobel Prize in statistics, John Tukey would have won it.) In the second place, this approach acknowledges that the null is never strictly true, but that sometimes the data do not allow us to draw conclusions about which mean is larger. So, instead of relying on fuzzy phrases like "fail to reject the null hypothesis" or "retain the null hypothesis," we simply do away with the whole idea of a null hypothesis and just conclude, "we can't decide whether μ_h is greater than μ_n, or is less than μ_n." In the third place, this looks as if we are running two one-tailed tests, but with an important difference. In a traditional one-tailed test, we must specify *in advance* which tail we are testing. If the result falls in the extreme of that tail, we reject the null and declare that $\mu_h < \mu_n$, for example. If the result does not fall in that tail, we must not reject the null, no matter how extreme it is in the other tail. But that is not what Jones and Tukey are suggesting. They do not require you to specify the direction of the difference before you begin.

Jones and Tukey are suggesting that we do not specify a tail in advance, but that we collect our data and determine whether the result is extreme in either tail. If it is extreme in the lower tail, we conclude that $\mu_h < \mu_n$. If it is extreme in the upper tail, we conclude that $\mu_h > \mu_n$. And if neither of those conditions applies, we declare that the data are insufficient to make a choice. (Notice that I didn't once use the word "reject" in the last few sentences. I said "conclude." The difference is subtle, but I think that it is important.)

But Jones and Tukey go a bit further and alter the significance level. First, we know that the probability that the null is true is .00. (In other words, $p(\mu_h = \mu_n) = 0$.) The difference may be trivially small, but there is a difference nonetheless. We cannot make an error by not rejecting the null because saying that we don't have enough evidence is not the same as incorrectly rejecting a hypothesis. As Jones and Tukey wrote,

> With this formulation, a conclusion is in error only when it is "a reversal," when it asserts one direction while the (unknown) truth is in the other direction. Asserting that the direction is not yet established may constitute a wasted opportunity, but it is not an error. We want to control the rate of error, the reversal rate, while minimizing wasted opportunity, that is, while minimizing indefinite results. (p. 412)

So one of two things is true—either $\mu_h > \mu_n$ or $\mu_h < \mu_n$. If $\mu_h > \mu_n$ is actually true, meaning that homophobic males are more aroused by homosexual videos, then the only error we can make is to erroneously conclude the reverse—that $\mu_h < \mu_n$. And the probability of that error is, at most, .025 if we were to use the traditional two-tailed test with 2.5% of the area in each tail. If, on the other hand, $\mu_h < \mu_n$, the only error we can make is to conclude that $\mu_h > \mu_n$, the probability of which is also at most .025. Thus, if we use the traditional cutoffs of a two-tailed test, the probability of a Type I error is at most .025. Jones and Tukey go on to suggest that we could use the cutoffs corresponding to 5% in each tail (the traditional two-tailed test at $\alpha = .10$) and still have only a 5% chance of making a Type I error. Although this is true, I think that you will find that many traditionally trained colleagues, including journal reviewers, will start getting a bit "squirrelly" at this point, and you might not want to push your luck.

I wouldn't be surprised if at this point students are throwing up their hands with one of two objections. First would be the claim that we are just "splitting hairs." My answer to that is "no, we're not." These issues have been hotly debated in the literature, with some people arguing that we abandon hypothesis testing altogether (Hunter, 1997). The Jones–Tukey formulations make sense of hypothesis testing and increase statistical power if you follow all their suggestions. (I believe that they would prefer the phrase "drawing conclusions" to "hypothesis testing.") Second, students could very well be asking why I spent many pages laying out the traditional approach and then another page or two saying why it is all wrong. I tried to answer that at the beginning—the traditional approach is so ingrained in what we do that you cannot possibly get by without understanding it. It will lie behind most of the studies you read, and your colleagues will expect that you understand it. That there is an alternative, and better, approach does not release you from the need to understand the traditional approach. And unless you change α levels, as Jones and Tukey recommend, you will be doing almost the same things but coming to more sensible conclusions.

4.11 Effect Size

Earlier in the chapter I mentioned that there was a movement afoot to go beyond simple significance testing to report some measure of the size of an effect. In fact, some professional journals are already insisting on it. I will expand on this topic in some detail later, but it is worth noting here that I have already sneaked a measure of effect size past you, and I'll bet that nobody noticed. When writing about waiting for parking spaces to open up, I pointed out that Ruback and Juieng (1997) found a difference of 6.88 seconds, which is not trivial when you are the one doing the waiting. I could have gone a step further and pointed out that, because the standard deviation of waiting times was 14.6 seconds, we are seeing a difference of nearly half a standard deviation. Expressing the difference between waiting times in terms of the actual number of seconds or as being "more than half a standard deviation" provides a measure of how large the effect was—and a very reputable measure. There is much more to be said about effect sizes, but at least this gives you some idea of what we are talking about.

I should say one more thing on this topic. One of the difficulties in understanding the debates about hypothesis testing is that for years statisticians have been very sloppy in selecting their terminology. Thus, for example, in rejecting the null hypothesis, it is very common for researchers to report that they have found a "significant difference." Most readers could be excused for taking this to mean that the study has found an "important difference," but that is not at all what is meant. When statisticians and researchers say "significant," that is shorthand for "statistically significant." It merely means that the difference, even if trivial, is not due to chance. The recent emphasis on effect sizes is intended to go beyond statements about chance, and tell the reader something, though perhaps not much, about "importance." I will try in this book to insert the word "statistical" before "significant," when that is what I mean, but I can't promise to always remember.

4.12 A Final Worked Example

A number of years ago the mean on the verbal section of the Graduate Record Exam (GRE) was 489 with a standard deviation of 126. The statistics were based on all students taking the exam in that year, most of whom were native speakers of English. Suppose we have an application from an individual with a Chinese name who scored particularly low (e.g., 220). If this individual were a native speaker of English, that score would be sufficiently low for us to question his suitability for graduate school unless the rest of the documentation is considerably better. If, however, this student were not a native speaker of English, we would probably disregard the low score entirely, on the grounds that it is a poor reflection of his abilities.

I will stick with the traditional approach to hypothesis testing in what follows, though you should be able to see the difference between this and the Jones and Tukey approach. We have two possible choices here, namely that the individual is or is not a native speaker of English. If he is a native speaker, we know the mean and the standard deviation of the population from which his score was sampled: 489 and 126, respectively. If he is not a native speaker, we have no idea what the mean and the standard deviation are for the population from which his score was sampled. To help us to draw a reasonable conclusion about this person's status, we will set up the null hypothesis that this individual is a native speaker, or, more precisely, he was drawn from a population with a mean of 489; $H_0 : \mu = 489$. We will identify H_1 with the hypothesis that the individual is not a native speaker ($\mu \neq 489$). (Note that Jones and Tukey would [simultaneously] test H_1: $\mu < 489$ and H_2: $\mu > 489$, and would associate the null hypothesis with the conclusion that we don't have sufficient data to make a decision.)

For the traditional approach we now need to choose between a one-tailed and a two-tailed test. In this particular case, we will choose a one-tailed test on the grounds that the GRE is given in English, and it is difficult to imagine that a population of nonnative speakers would have a mean higher than the mean of native speakers of English on a test that is given in English. (*Note:* This does not mean that non-English speakers may not, singly or as a population, outscore English speakers on a fairly administered test. It just means that they are unlikely to do so, especially as a group, when both groups take the test in English.) Because we have chosen a one-tailed test, we have set up the alternative hypothesis as H_1: $\mu < 489$.

Before we can apply our statistical procedures to the data at hand, we must make one additional decision. We have to decide on a level of significance for our test. In this case, I have chosen to run the test at the 5% level, instead of at the 1% level, because I am using $\alpha = .05$ as a standard for this book and also because I am more worried about a Type II error than I am about a Type I error. If I make a Type I error and erroneously conclude that the student is not a native speaker when in fact he is, it is very likely that the rest of his credentials will exclude him from further consideration anyway. If I make a Type II error and do not identify him as a nonnative speaker, I am doing him a real injustice.

Next, we need to calculate the probability of a student receiving a score *at least as low as* 220 when $H_0 : \mu = 489$ is true. We first calculate the z score corresponding to a raw score of 220:

$$z = \frac{X - \mu}{\sigma} = \frac{(220 - 489)}{126} = \frac{-269}{126} = -2.13$$

We then go to tables of z to calculate the probability that we would obtain a z value less than or equal to -2.13. From Appendix z, we find that this probability is .017. Because this probability is less than the 5% significance level we chose to work with, we will reject the null hypothesis on the grounds that it is too unlikely that we would obtain a score as low as 220 if we had sampled an observation from a population of native speakers of English who had taken the GRE. Instead, we will conclude that we have an observation from an individual who is not a native speaker of English.

It is important to note that in rejecting the null hypothesis, we could have made a Type I error. We know that if we do sample speakers of English, 1.7% of them will score this low. It is possible that our applicant was a native speaker who just did poorly. All we are saying is that such an event is sufficiently unlikely that we will place our bets with the alternative hypothesis.

4.13 Back to Course Evaluations and Rude Motorists

We started this chapter with a discussion of the relationship between how students evaluate a course and the grade they expect to receive in that course. Our second example looked at the probability of motorists honking their horns at low- and high-status cars that did not move when a traffic light changed to green. As you will see in Chapter 9, the first example uses a correlation coefficient to represent the degree of relationship. The second example simply compares two proportions. Both examples can be dealt with using the techniques discussed in this chapter. In the first case, if there were no relationship between the grades and ratings, we would expect that the true correlation in the population of students is .00. We simply set up the null hypothesis that the population correlation is .00 and then ask about the probability that a sample of observations would produce a correlation as large as the one we obtained. In the second case, we set up the null hypothesis that there is no difference between the proportion of motorists *in the population* who honk at low- and high-status cars. Then we calculate the probability of obtaining a difference in sample proportions as large as the one we obtained (in our case, .34) if the null hypothesis is true. I do not expect you to be able to run these tests now, but you should have a general sense of the way we will set up the problem when we do learn to run them.

Key Terms

Sampling error (Introduction)
Hypothesis testing (4.1)
Sampling distributions (4.2)
Sampling distribution of the differences between means (4.2)
Research hypothesis (4.3)
Null hypothesis (H_0) (4.3)
Alternative hypothesis (H_1) (4.4)

Sample statistics (4.5)
Test statistics (4.5)
Decision making (4.6)
Rejection level (significance level) (4.6)
Rejection region (4.6)
Critical value (4.7)
Type I error (4.7)

α (alpha) (4.7)
Type II error (4.7)
β (beta) (4.7)
Power (4.7)
One-tailed test (directional test) (4.8)
Two-tailed test (nondirectional test) (4.8)
Conditional probabilities (4.9)

Exercises

4.1 Suppose I told you that last night's NHL hockey game resulted in a score of 26–13. You would probably decide that I had misread the paper and was discussing something other than a hockey score. In effect, you have just tested and rejected a null hypothesis.

a. What was the null hypothesis?

b. Outline the hypothesis-testing procedure that you have just applied.

4.2 For the past year, I have spent about $4.00 a day for lunch, give or take a quarter or so.

a. Draw a rough sketch of this distribution of daily expenditures.

b. If, without looking at the bill, I paid for my lunch with a $5 bill and received $.75 in change, should I worry that I was overcharged?

c. Explain the logic involved in your answer to part (b).

4.3 What would be a Type I error in Exercise 4.2?

4.4 What would be a Type II error in Exercise 4.2?

4.5 Using the example in Exercise 4.2, describe what we mean by the rejection region and the critical value.

4.6 Why might I want to adopt a one-tailed test in Exercise 4.2, and which tail should I choose? What would happen if I chose the wrong tail?

4.7 A recently admitted class of graduate students at a large state university has a mean Graduate Record Exam (GRE) verbal score of 650 with a standard deviation of 50. (The scores are reasonably normally distributed.) One student, whose mother just happens to be on the board of trustees, was admitted with a GRE score of 490. Should the local newspaper editor, who loves scandals, write a scathing editorial about favoritism?

4.8 Why is such a small standard deviation reasonable in Exercise 4.7?

4.9 Why might (or might not) the GRE scores be normally distributed for the restricted sample (admitted students) in Exercise 4.7?

4.10 Imagine that you have just invented a statistical test called the Mode Test to test whether the mode of a population is some value (e.g., 100). The statistic (M) is calculated as

$$M = \frac{\text{Sample mode}}{\text{Sample range}}$$

Describe how you could obtain the sampling distribution of M. (*Note:* This is a purely fictitious statistic as far as I am aware.)

4.11 In Exercise 4.10, what would we call M in the terminology of this chapter?

4.12 Describe a situation in daily life in which we routinely test hypotheses without realizing it.

4.13 In Exercise 4.7, what would be the alternative hypothesis (H_1)?

4.14 Define "sampling error."

4.15 What is the difference between a "distribution" and a "sampling distribution"?

4.16 How would decreasing α affect the probabilities given in Table 4.1?

4.17 Give two examples of research hypotheses, and state the corresponding null hypotheses.

4.18 For the distribution in Figure 4.4, I said that the probability of a Type II error (β) is .74. Show how this probability was obtained.

4.19 Rerun the calculations in Exercise 4.18 for $\alpha = .01$.

4.20 In the example in Section 4.12, how would the test have differed if we had chosen to run a two-tailed test?

4.21 Describe the steps you would go through to develop the example given in this chapter about the course evaluations. In other words, how might you go about determining whether there truly is a relationship between grades and course evaluations?

4.22 Describe the steps you would go through to test the hypothesis that motorists are ruder to fellow drivers who drive low-status cars than to those who drive high-status cars.

Discussion Questions

4.23 In Chapter 1, we discussed a study of allowances for fourth-grade children. We considered that study again in the exercises for Chapter 2, where you generated data that might have been found in such a study.

a. Consider how you would go about testing the research hypothesis that boys receive more allowance than girls do. What would be the null hypothesis?

b. Would you use a one- or a two-tailed test?

c. What results might lead you to reject the null hypothesis and what might lead you to retain it?

d. What single thing might you do to make this study more convincing?

4.24 Simon and Bruce (1991), in demonstrating a different approach to statistics called "Resampling statistics,"[4] tested the null hypothesis that the mean price of liquor (in 1961) for the 16 "monopoly" states, where the state owned the liquor stores, was different from the mean price in the 26 "private" states, where liquor stores were privately owned. (The means were $4.35 and $4.84, respectively, giving you some hint at the effects of inflation.) For technical reasons, several states don't conform to this scheme and could not be analyzed.

a. What is the null hypothesis that we are really testing?

b. What label would you apply to $4.35 and $4.84?

c. If these are the only states that qualify for our consideration, why are we testing a null hypothesis in the first place?

d. Can you think of a situation where it does make sense to test a null hypothesis here?

4.25 Discuss the different ways that the traditional approach to hypothesis testing and the Jones and Tukey approach would address the question(s) inherent in the example of waiting times for a parking space.

4.26 What effect might the suggestion that experimenters report effect sizes have on the conclusions we draw from future research studies in psychology?

[4] The home page containing information on this approach is available at http://www.resample.com/. I will discuss resampling statistics at some length in Chapter 18.