12 Testing a Model

How many times have you wished for one more hour to study for a midterm exam to increase your chances of getting an A? All you needed was that one extra hour of study. But does one more hour of study really make that much of a difference?

To address this question, a teacher could instruct students to come to a two-hour midterm examination with their notes and study materials. Then half the students would be given an opportunity to study and the other half would engage in some irrelevant activity such as watching soap operas. After an hour they would all take the midterm. The teacher would then see whether those who had the extra hour of study did better on the midterm than those who did not have the extra time. The teacher would know if one hour of study makes a difference, or more accurately, how much of a difference.

What was just described is a research study. Research can be used to help answer important questions such as the following.

1. Does divorce affect children's social development?
2. Does psychotherapy improve one's mental health?
3. Does television violence make children more aggressive?
4. Does bilingual education retard or accelerate the performance of children in schools?

Research is more than "men in tweed suits, cutting up frogs, paid for by huge government grants" (Woody Allen, in the movie *Sleeper*). Research helps us in understanding the world around us. Research in the behavioral and social sciences often involves testing statistical models.

What Is a Model?

A statistical model is a formal representation of a set of relationships between variables. Statistical models contain an outcome variable that is the focus of study. In studies of weight change, the outcome is weight change; in studies

of psychotherapy, it is adjustment; in studies of education, one often-studied outcome is reading skill. In research, the outcome of interest is called the *dependent variable*. A dependent variable is what is supposed to change in response to changing events. In statistical models, it is written on the left-hand side of the equal sign.

The variable that brings about changes in the dependent variable is called the *independent variable*. Examples of independent variables are type of psychotherapy, drug dosage, and age. The dependent variable is assumed to be some function of the independent variable. How the independent variable affects the dependent variable is represented on the right-hand side of the equation.

Sometimes the designation between independent and dependent variable depends on the variables under study and the researcher's theoretical orientation. For instance, researchers study the relationship between self-esteem and academic performance. Some designate self-esteem as the independent variable and academic performance as the dependent variable. Others reverse the designations.

Other variables that cause the dependent variable to vary besides the independent variable are represented by the *residual variable*. The residual variable represents the degree to which the researcher is ignorant about what causes the dependent variable. The residual variable is sometimes referred to as error or noise.

In simple equation form the model is

$$\begin{matrix}\text{dependent} \\ \text{variable}\end{matrix} = \begin{matrix}\text{effect of the} \\ \text{independent} \\ \text{variable}\end{matrix} + \begin{matrix}\text{residual} \\ \text{variable}\end{matrix}$$

By far the vast majority of models in the social and behavioral sciences take on this general form. The only major difference is that most models have more than one independent variable on the right-hand side, but the basic specification of the model remains the same.

In this model the independent variable and the residual variable are *added* together to cause the dependent variable. This is not the only way that the independent and the residual variable could combine. For instance, they could multiply. However, an additive formulation is by far the simplest and most common formulation. Most of the standard statistical models assume that the effect of the independent variable and the residual variable add together.

Instead of expressing the model as an equation, the model could be just as easily specified by a diagram; arrows could be drawn from cause to effect, as follows:

$$\begin{matrix}\text{independent} \\ \text{variable}\end{matrix} \longrightarrow \begin{matrix}\text{dependent} \\ \text{variable}\end{matrix} \longleftarrow \begin{matrix}\text{residual} \\ \text{variable}\end{matrix}$$

A representation of a model that uses arrows is called a *path diagram*.

To better understand a statistical model, consider the following example. A researcher, investigating the effect of owning a personal computer on grade-point average, made arrangements to give a personal computer to each of 30 students. Another 30 students served as a comparison group and they did not receive computers. One year later, the researcher measured the grade-point averages of the two groups. The independent variable is owning or not owning a personal computer, and the dependent variable is grade-point average. The residual variable represents any other causes of grade-point average besides owning a personal computer. The residual variable is the way of accounting for the fact that all students with computers (or without computers) do not have the same grade-point average.

Statistical models are a bit more complicated than the independent variable and the residual variable causing the dependent variable. In most models a constant is added to every person's score. In equation form,

$$\begin{matrix}\text{dependent}\\ \text{variable}\end{matrix} = \text{constant} + \begin{matrix}\text{effect of the}\\ \text{independent}\\ \text{variable}\end{matrix} + \begin{matrix}\text{residual}\\ \text{variable}\end{matrix}$$

In many models the constant term corresponds to the population mean of the dependent variable.

The residual term is a necessary part of a statistical model. It is also called the disturbance, error, or noise. The mean of the residual variable is set to zero. This is not a mathematical necessity but is merely a convention. Also, it is very often assumed that the residual variable has a normal distribution with a given variance. It should be noted that it is the residual and not the dependent variable that is assumed to have a normal distribution. Also, it is commonly assumed that the variance of the residual does not vary as a function of the independent variable. Many of the assumptions of the model refer to the residual variable. In sum, the residual is a normally distributed variable with a zero mean.

Model Comparison

In this chapter the logic of model testing is presented. It is first illustrated for one type of model and then the general procedure is discussed. A very simple model is one in which the dependent variable equals a constant plus the residual variable.

$$\begin{matrix}\text{dependent}\\ \text{variable}\end{matrix} = \text{constant} + \begin{matrix}\text{residual}\\ \text{variable}\end{matrix}$$

It is this model that will be considered in this chapter. There is no independent variable effect in the model. This model will be called the *complete model* because later an even simpler version of the model is considered. The model

has two parameters: the constant and the standard deviation of the residual variable. The constant in this model is the population mean of the dependent variable, and the standard deviation of the residual variable is the standard deviation of the dependent variable.

In models, a parameter can be fixed or free. If the parameter is free, it must be estimated from the data. If it is fixed, then the researcher sets the parameter to some a priori value. In this chapter, the constant is set to some a priori value, as in the following examples.

1. Eighty-seven persons are asked to learn pairs of words like "cat-package." They are then presented the word "cat" and are asked to recall whether the other word was "package" or "glass." Because there were two alternatives for each word pair, the probability of being correct is .5. There are ten such trials, and if subjects were only guessing, they would be expected to be correct on five of the ten trials. The dependent variable is the number correct out of ten and the a priori constant is 5.0.
2. Twenty persons aged 50 were asked at what age they would ideally prefer to retire. The researcher sought to compare the preferred age of retirement of persons to the standard retirement age of 65. The dependent variable is preferred retirement age and the a priori constant is 65.0.
3. Robinson and Hastie (1985) had 40 undergraduates read a mystery story "The Poisoned Philanthropist," in which there are five suspects. The subject had to estimate the probability that any given suspect was guilty. For each subject the five probabilities are summed. The dependent variable is total probability and the a priori constant is 1.00. (The mean probability of the subjects was over 2.00.)
4. In an extrasensory perception study, twelve proclaimed psychics were asked to guess whether a head or a tail results when a coin is flipped. The coin was flipped 30 times. By chance, each psychic would be correct 15 times. The dependent variable is the number of correct judgments and the a priori constant is 15.0.

When the constant is fixed and not free, the researcher is specifying a simple and restricted version of the model:

$$\begin{matrix}\text{dependent}\\ \text{variable}\end{matrix} = \text{constant} + \begin{matrix}\text{residual}\\ \text{variable}\end{matrix}$$

This model is restricted in that the constant is not free to take on any value but instead is fixed or set to some a priori value. A model in which a parameter of the complete model is fixed is called the *restricted model*. In the restricted model under consideration, the constant parameter is fixed or restricted to some a priori value. The hypothesis of interest is whether the parameter equals the value to which it is restricted. This hypothesis is referred to as the null hypothesis. The *null hypothesis* is the constraint on the complete model that is present in the restricted model. (It is common to symbolize the null

hypothesis as H_0.) For instance, for the model that the psychics are guessing, the null hypothesis is that the constant equals 15.0. Although in testing the restricted model the interest is primarily the null hypothesis, it is not uniquely tested. Rather, the plausibility of a model, of which the null hypothesis is a part, is evaluated. The *alternative hypothesis* is the hypothesis that is true if the null hypothesis is false. (It is common to symbolize the alternative hypothesis as H_A.) It states that the constant is free to take on any value. So, for the psychic example, the alternative hypothesis is that the constant does not equal 15.0.

Model testing is always model comparison. The restricted model is compared to the complete model. The restricted model is a simpler model which is identical to the complete model except that one of the parameters in the restricted model is fixed to some value. If the restricted model is not contradicted by the data, then the restricted model is retained for reasons of simplicity, and the more complicated complete model is not considered. However, if the restricted model is contradicted by the data, the restricted model is rejected, and the more complicated complete model must be adopted.

To illustrate the difference between the complete and restricted models, consider the three presented in Table 12.1. Model I is the simplest of the three. In it, the dependent variable is not caused by any independent variable. In Model II the dependent variable is caused by variable A, and in Model III it is caused by both variables A and B. Considering III as the complete model, II would be a restricted model for III. The restriction present in Model II is that

TABLE 12.1 **Illustration of Complete and Restricted Models**

Model I
dependent variable = constant + residual
Model II
dependent variable = constant + effect of independent variable A + residual
Model III
dependent variable = constant + effect of independent variable A + effect of independent variable B + residual

independent variable *B* does not cause the dependent variable. Considering II as the complete model and I as the restricted model, the restriction present in Model I is that variable *A* has no effect. So, Model II can be considered either as a complete or restricted model: If II is compared to III, it is a restricted model; if compared to I, it is a complete model.

Consider the hypothetical data in Table 12.2 from the experiment with twelve psychics. A coin was flipped 30 times and each time each "psychic" guessed whether it came up heads or tails. Because there are two sides to a coin, pure guessing would lead to accuracy on 15 trials (or 1/2 times 30). So, if there were a large number of supposed psychics who were only guessing, they would on the average be correct on 15 out of 30 trials. But there is not a large number—only twelve. The question is whether the numbers in Table 12.2 are compatible with the view that the psychics are fakers who are just guessing. The mean of the twelve numbers is 16.0 and the standard deviation is 2.04.

Although the psychics did not do a stunning job at the task, their results seem to be better than chance. Only one had an exactly chance performance of 15. Of the remaining eleven, there were eight who did better than chance and only three who did worse than chance. The mean of the twelve is 16.0, a full one "guess" better than chance. The conclusion might be drawn that the psychics did better than chance.

However, if they were merely guessing, then about half the time they would appear to do better than chance and about half the time they would appear to do worse than chance. Even if it is believed that the twelve were just guessing, it is totally unrealistic to expect each psychic to be correct exactly 15 out of 30 trials or even for the sample mean of the twelve psychics to be exactly 15. Just because the sample mean is greater than the chance value of 15.0 does not necessarily refute the view that the supposed psychics were just guessing. Sampling error is to be expected, and so it would be expected that they would score better than chance about half the time. At issue is whether the value of 16.0 obtained by the psychics is within the limits of reasonable sampling error.

In the restricted model, the constant is set at 15.0. The restricted model presumes that the psychics are guessing. In the complete model the constant may be any value, and so it is compatible with the view that the psychics are not guessing.

If the restricted model were true (that psychics are guessing), then the

TABLE 12.2 **Guesses of Twelve Psychics (Hypothetical Data)**

15	17	19	13
16	16	18	16
19	14	13	16

sample mean of the number of correct guesses should be near 15.0. It happens that this particular sample mean is 16.0, one unit greater than the a priori value of 15.0. At issue is whether 16.0 is near enough to 15.0 to be explained by sampling error. The standard deviation of the guesses can be used to gauge how near 15.0 the sample mean should be. Assuming that the psychics were guessing, the smaller the standard deviation, the nearer the sample mean should be to 15.0. Also as the sample size gets larger, the sample mean should be nearer to 15.0. So, as the standard deviation gets smaller and the sample size larger, the sample mean should approach its a priori value.

Both the standard deviation and the sample size are in the formula for the standard error of the sample mean minus an a priori constant. As is presented in the previous chapter, the standard error of the mean minus a constant equals the standard deviation of the observations divided by the square root of the sample size. The difference between the sample mean and the a priori mean can be divided by its standard error to obtain

$$\frac{\bar{X} - M}{s/\sqrt{n}}$$

where $\bar{X}$ is the sample mean, M the a priori mean, n the sample size, and s the standard deviation of the observations. This value normalizes the difference between the sample mean and the presumed population mean to take into account sample size and variability.

For the psychic example, $\bar{X}$ is 16.0, M is 15.0, n is 12, and s is 2.04. The sample mean minus its a priori value divided by its standard error is as follows:

$$\frac{16.0 - 15.0}{2.04/\sqrt{12}} = 1.698$$

Thus, the sample mean is 1.698 standard errors above the mean. The question now is just how unlikely is this type of outcome. If $\bar{X}$ was ten standard errors above or below the a priori constant, it would be known almost for certain that the psychics were not guessing because it is virtually impossible to obtain a value ten standard errors above the mean. Alternatively, if it were only one standard error or less above the mean, it is still plausible to believe that they are guessing. But the value of 1.698 standard error above the mean for the psychic example is ambiguous. The "psychics" did better than chance, but it is not clear whether their success might have been due to sampling error.

The Test Statistic and Its Sampling Distribution

The quantity $(\bar{X} - M)/(s/\sqrt{n})$ is called the *test statistic*. Of prime concern is how unusual is a test statistic of 1.698. To determine exactly how unlikely a

value like 1.698 is, the distribution of the quantity $(\bar{X} - \mathrm{M})/(s/\sqrt{n})$ must be known. The quantity $(\bar{X} - \mathrm{M})/(s/\sqrt{n})$ is computed from sample data and so it is a statistic. As described in Chapter 9, the distribution of a statistic is called a *sampling distribution*. Given the restricted model, if the residual variable is normally and independently distributed, then $(\bar{X} - M)/(s/\sqrt{n})$ has a t distribution with $n - 1$ degrees of freedom. Figure 12.1 shows the theoretical t distribution for eleven degrees of freedom.

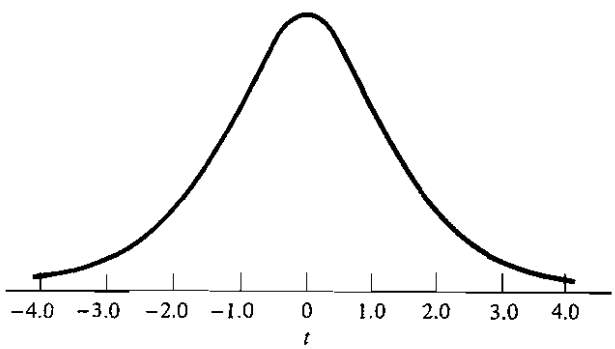
As can be seen in Figure 12.1, the t distribution is a symmetric unimodal distribution whose mean is zero. Its variance depends on its degrees of freedom and is always greater than one for finite degrees of freedom. As the degrees of freedom increase, the variance of t approaches one. Consequently, the tails of the t distribution are a bit fatter than the standard normal or Z distribution. The number of degrees of freedom for t in this case is $n - 1$. A t value may be denoted by $t(df)$ where df stands for degrees of freedom. So for the psychic example, the df are twelve minus one, or eleven.

Because t is a continuous distribution, the probability that $t(11)$ exactly equals any particular value, such as 1.698, is zero. What is needed is not the probability that t is 1.698 but rather the probability of obtaining a value of 1.698 or greater. At issue is the probability of obtaining a value at least as large as the test statistic.

There are two ways the restricted model could be wrong. The population mean could be larger than the a priori value or it could be smaller. For the psychic example, they could do better than chance (better than 15), which they did, or they could have performed worse than chance (worse than 15). So if the null hypothesis is wrong, there are two directions or sides that it could be wrong.

If the null hypothesis is false, either the psychics could do better than chance or worse than chance. Only one of the two may be plausible. For the particular example, it does not seem very reasonable that the psychics could

FIGURE 12.1 **The t distribution with 11 degrees of freedom.**



be operating at a level worse than chance. However, if such a result did occur, it should be considered as unusual. So even though there is little reason to expect the psychics to do worse than chance, it remains a possibility, and both alternative hypotheses need to be considered: that $(\bar{X} - M)/(s/\sqrt{n})$ is very positive or very negative. For the psychic example the probability of obtaining a value greater than 1.698 or less than –1.698 must be determined.

If the restricted model were clearly false, the value of $(\bar{X} - M)/(s/\sqrt{n})$ would tend to be either very positive or very negative. From Figure 12.1 it can be seen that if the restricted model were false, the value of t would fall in either *tail* of the t distribution. It is for this reason the test is called *two-tailed.* Although it is not recommended, a researcher might wish to consider only one direction or tail. For instance, only the probability that $(\bar{X} - M)/(s/\sqrt{n})$ is greater than 1.698. Such a test is called a *one-tailed test.* A one-tailed test is not recommended because if the value of the test statistic is quite unusual but in the wrong direction, most researchers would still consider it significant. Also, almost all computer programs output two-tailed p values.

As has been stated, under the restricted model, the test statistic $(\bar{X} - \mathrm{M})/(s/\sqrt{n})$ has a t distribution with $n - 1$ or eleven degrees of freedom. Using the t distribution it can be determined how likely a value greater than 1.698 or less than –1.698 actually is. Such a value would occur by chance about 12% of the time or about one out of eight times. It must now be decided whether 12% of the time is sufficiently unusual to reject the restricted model.

Significance Level and p *Value*

Researchers who test statistical models have established a fairly standard, though arbitrary, criterion for judging how unusual a result must be to reject the restricted model. They have, by informal convention, required that the result and even more extreme results must occur no more than 5% of the time before the restricted model is rejected. The question now becomes: Given the restricted model, how often will the absolute value of $(\bar{X} - M)(s/\sqrt{n})$ be 1.698 or more? If it would occur 50% of the time, the result would not be considered unusual. But if it only occurs once in 20 times, the result would be unusual.

This 5% criterion is said to be the *significance level.* It is the standard of proof that is required for the restricted model to be deemed implausible. Other standards are also used. A common alternative standard is the .01 (or 1%) significance level. A result is judged to be improbable if it would occur by chance only once in 100 times. More stringent levels of once in 1000 are sometimes used, and less stringent rules of once in ten and even once in five are infrequently used. The choice of the significance level depends on the type of error that the researcher is more willing to accept. (See later section on errors in model comparison.)

The significance level is symbolized by the Greek letter alpha (α). Con-

ventionally, alpha is fixed to .05 or once in 20. *The usual significance level used in research is .05.* There is nothing magical about .05, just as there is nothing magical about setting the legal definition of being drunk at 0.1% blood alcohol. Some cutoff must be set and for various reasons the .05 is the value taken for alpha. The .05 significance level often means that the test statistic must be more than twice as large as its standard error to be significant.

A *p value* is the probability of obtaining a value equal to or more extreme than the test statistic. The *p* value for the psychic experiment is .12. If the *p* value is less than or equal to the significance level, then the null hypothesis is rejected and the test statistic is said to be *statistically significant.* If the *p* value is greater than the significance level, the null hypothesis is retained and the test statistic is said to be *statistically insignificant.* For the psychic example, because .12 is greater than .05, the null hypothesis is retained.

The significance level is usually set at .05. How is the *p* value determined? Computer packages routinely calculate the *p* value of the test statistic. Without a computer, one can use Appendix D to determine the approximate *p* value of a test statistic distributed as *t*.

To use Appendix D, one first determines the degrees of freedom of the *t* statistic. For this model the degrees of freedom equal the sample size less one or $n - 1$. One locates in the first column in Appendix D the degrees of freedom, $n - 1$. If the exact value for degrees of freedom is not in the first column, one uses the closest value that is smaller than the actual degrees of freedom. One "rounds down" to the nearest value. For instance, 105 is not in the table and so 100 would be used. One then reads across the row and finds the value that is the closest to the test statistic without being larger than the test statistic. One then reads up the column to determine the approximate *p* value. The exact *p* value is always less than or equal to the approximate *p* value. Thus, this method results in a conservative estimate of the *p* value. The numbers in the table are called *critical values* because they are the values that the test statistic must exceed to be statistically significant.

The 1.698 value for the psychic example does not exceed the .05 level of significance. The null hypothesis is retained and the 1.698 value is judged to be not significant. So on the basis of this data set, there is no reason to believe that the "psychics" have any special powers beyond mere guessing.

If one wishes to consider only one tail of the *t* distribution (a one-tailed test), the sign of the test statistic must match the prediction of the researcher before it is tested. If it does match, then the *p* value should be divided in half. (Note, the *p* value and not the significance level is divided by two.) For instance, if it is considered only that the psychics would do better than chance and not worse, then the test statistic must be positive. Because it is, the *p* value would be .06. If $\bar{X}$ had been less than 15.0 (even a lot less), the restricted model would be retained. As was stated earlier, one-tailed tests are not recommended.

For the quantity $(\bar{X} - M)/(s/\sqrt{n})$ to be distributed as t with $n - 1$ degrees of freedom, it must be assumed that the residual variable has a normal distribution and that observations are randomly and independently sampled.

Even if the normality assumption is violated moderately, there is only a slight effect on the p values. Thus, unless the numbers are highly skewed or bimodal, one need not worry about the normality assumption. The reason for this is the central limit theorem described in Chapter 10. As sample size increases, the distribution of $\bar{X}$ approaches a normal distribution regardless of the shape of the distribution of the scores used to compute $\bar{X}$.

The random and independent sampling assumptions are more important for testing hypotheses concerning the constant. Random sampling ensures that persons are representative of the population. The independence assumption requires that persons do not interact with one another, be observed only once, and that the person be the sampling unit.

The Summary of the Logic of Model Comparison

The logic of model testing involves the following steps. First, a model is specified from theory that contains the parameter of interest. This is the *complete model*. A restricted version of the same model is constructed with some reasonable constraint on the parameter of interest. The constraint is called the *null hypothesis*, and the model with the constraint is called *the restricted model*.

The model describes the behavior in the population. Sample data are gathered from the population. The researcher computes a statistic from the sample data. The statistic, called the *test statistic*, has a distribution such as Z, t, χ^2, or F if the restricted model is true. Given the distribution, the probability of obtaining a value as or more extreme than the test statistic can be determined. This probability is called the *p value*. The p value can be exactly computed by using a computer program or it can be approximated by using tables. If the p value is less than the *significance level*, which is usually set at .05, the null hypothesis is rejected and the test statistic is said to be *statistically significant*. If the p value is greater than the significance level, the restricted model and null hypothesis are retained and the test statistic is said to be *not significant*. This information as applied to the test for psychic powers is summarized in Table 12.3.

Model testing can be viewed as a series of "let's assume" and, "given all this" statements. First, *let's assume* that the null hypothesis is true. Second, *let's assume* a restricted model which contains the null hypothesis is true. Third, from the data a number called the test statistic is computed. Fourth, *given all this*, the test statistic has a sampling distribution. Fifth, *given all*

TABLE 12.3 **Steps in Model Testing Illustrated for the Test of Psychic Powers in Table 12.2**

Complete Model	
$\text{number correct} = \text{constant} + \text{residual variable}$	
Restricted Model	
$\text{number correct} = 15.0 + \text{residual variable}$	
Null Hypothesis	Alternative Hypothesis
constant = 15.0	constant ≠ 15.0
Test Statistic	
$t(n-1) = \dfrac{\bar{X} - M}{s/\sqrt{n}}$	
$t(11) = \dfrac{16.0 - 15.0}{2.04/\sqrt{12}} = 1.698$	
p Value	
exact .12	
approximate .20	

this, if the test statistic's p value is less than or equal to the significance level (usually set at .05), the null hypothesis is rejected. The complete model is never directly tested, rather it is the restricted model that is tested. If the restricted model is judged to be implausible, it is rejected and the complete model is adopted.

A second example is used to apply these ideas. A researcher in a school district wants to determine whether the children in the district score above national norms on a test. The norm on the test is 200. The scores of nine children randomly and independently sampled are 240, 230, 220, 190, 220, 200, 250, 230, and 190. The complete model is

$$\text{test score} = \text{constant} + \text{residual}$$

and the restricted model is

$$\text{test score} = 200 + \text{residual}$$

The mean of the nine scores is 218.89 and the standard deviation is 21.47. The test statistic is

$$t(8) = \frac{218.89 - 200}{21.47/\sqrt{9}}$$

which equals 2.639. Using Appendix D, the test statistic 2.639 yields a p value of .05. (The exact p value is .030.) Because the p value is less than or equal to the significance level of .05, the test statistic of 2.639 is statistically significant at the .05 level. The null hypothesis that the constant equals 200 is rejected. Because 218.89 is above 200, it is thus concluded that the children in the district score above the national norm of 200.

For this example, the truth of the assumption of random sampling is essential in the test of the restricted model. If the students were not randomly sampled but only the district's brightest students were studied, the conclusion that students in the district score above the norm would be unjustified. Also, if the students shared answers on the test, then the sample data would not be independent, and thus the conclusion would be unjustified.

Errors in Model Comparison

There are two major types of errors that can be made in the comparison of statistical models. To understand these errors, four hypothetical yet possible results of testing a restricted model must be considered. The restricted model can be actually true or it can be false. For instance, the psychics could be just guessing or they could be true psychics. Of course, one never knows with perfect certainty whether any model is valid or not and so some idealized knowledge is being considered. The results of the statistical analysis can lead to rejection of the restricted model or its retention. For instance, the psychics could be operating significantly above chance or they could be operating at chance levels. Table 12.4 lists these four outcomes.

TABLE 12.4 **Four Possible Results of Model Testing**

	Statistical Analysis	
Reality	Retain Restricted Model	Reject Restricted Model
Restricted Model True	Retain a True Model	Type I Error (alpha)
Restricted Model False	Type II Error (beta)	Reject a False Model

For two of the outcomes in Table 12.4, the correct conclusion is drawn. In the top left-hand cell, the restricted model is correctly retained. For instance, the psychics are fakers and it is also concluded that their mean is not significantly above chance. In the bottom right-hand cell, a false restricted model is correctly rejected. For instance, the psychics are true psychics and it is also concluded that their performance is above chance. In both cases the sample data and the statistical test mirror reality. In the best of all worlds, one would hope to make the correct decision every time. However, statistical tests of models do not allow for inferences about the nature of reality with total certainty. Statistical logic never brings with it certainty; rather, statistical logic results in only a probability.

There are two types of errors. The first is the error of falsely rejecting a restricted model that is actually true. This is called a Type I error. For instance, if the psychics were fakers, it might be falsely concluded that their performance is above chance. The second error is to retain the restricted model that is actually false. This is a Type II error. For instance, if the psychics were true psychics, it might be mistakenly concluded that they are not performing significantly above chance levels.

The probability of making a Type I error is called alpha and is identical to the significance level. Alpha is usually set at .05 or one out of 20. The probability of making a Type II error is symbolized by *beta*, β. Its value is not set by the researcher like alpha, but rather it is largely determined by the number of persons in the study. Beta is then smaller if more persons are studied. The probability of correctly concluding that the restricted model is false is called *power*. Power then equals one minus the probability of making a Type II error. In Chapters 13 and 16, methods for determining power are presented.

There are two other important errors in model testing that need to be considered. One is to draw the incorrect conclusion when the restricted model is rejected. If the p value of the test statistic is less than or equal to the significance level, then the null hypothesis is rejected. But what is rejected is the restricted model and not necessarily the null hypothesis. It can be that some other aspect of the restricted model is false. For instance, it might be that the assumption that the residual variable has a normal distribution is false. Just because the restricted model is false does not imply that the null hypothesis is false. This error will be referred to as an *assumption violation*.

Second, when the restricted model is rejected, the null hypothesis is rejected. Just because the null hypothesis is rejected, it does not mean that the result necessarily supports the researcher's theoretical position. For instance, it may be that the psychics do not perform at chance levels, not because they do better than chance but because they do worse than chance. If it is concluded that the psychics did better than chance, an error would be made. This is an error about the direction in which the null hypothesis is false. This error will be referred to as *choosing the wrong direction*, and can be avoided by careful examination of the data.

Remainder of the Book

So far only the simplest model in which the dependent variable equals the constant plus the residual variable has been considered. The restricted version of this model constrains the constant to equal some fixed value determined a priori. The remainder of the book considers more complex models. These more complex models are outlined in Table 12.5.

In Chapter 13 a model is presented in which the dependent variable is caused by an independent variable, and the independent variable is a nominal variable with only two levels. In Chapter 14 the independent variable is still a nominal variable, but it may have more than two levels. In Chapter 15 there are two nominal independent variables that both cause the dependent variable. In Chapter 16 both the independent variable and the dependent variable are measured at the interval level. In Chapter 17, both the independent and dependent variables are not at the interval level of measurement, but rather are at the nominal level of measurement. Finally, in Chapter 18, the dependent variable is at the ordinal level of measurement.

When the independent variable is at the nominal level of measurement, it is possible to have the same persons in all conditions or have different persons.

TABLE 12.5 **Taxonomy of Models**

Dependent Variable at the Interval Level of Measurement

No independent variable
Test of the constant, Chapter 12

Nominally measured independent variable
One independent variable
Dichotomous: two-sample *t* test, Chapter 13
Multilevel: one-way analysis of variance, Chapter 14
Two independent variables: two-way analysis of variance, Chapter 15

Intervally measured independent variable
Regression, Chapter 16

Dependent Variable at the Nominal Level of Measurement

No independent variable: chi-squared goodness-of-fit test, Chapter 17

Nominally measured independent variable: chi-squared test of independence, Chapter 17

Dependent Variable at the Ordinal Level of Measurement

Nominally measured independent variable
Dichotomous: Mann-Whitney *U* test, Chapter 18
Multilevel: Kruskal-Wallis analysis of variance, Chapter 18

Ordinally measured independent variable: rank-order coefficient, Chapter 18

When different persons are in each group, the design is said to have *independent groups.* When the same persons are in each group, the design is said to have *nonindependent groups;* nonindependent groups with two groups are commonly referred to as *paired groups;* and multiple-group designs in which groups are nonindependent are called *repeated measures* designs.

A nonindependent group design can come about even when different persons are at each level of the independent variable. Whenever there is some factor that links together observations across the different conditions, the design can be considered a nonindependent groups design. So, if persons are from the same family, litter, or class, the design can be considered a nonindependent groups design.

The statistical procedures presented in Table 12.5 presume that the groups are independent. In Table 12.6 are the statistical procedures for nonindependent groups. For each procedure, the independent variable is a nominal variable. Different statistical tests are used for nominal, ordinal, and interval-dependent variables.

Summary

A *model* is a formal set of relationships between variables. The *dependent variable* in the model is the outcome and the *independent variable* is presumed to bring about the change in the dependent variable. Most models have a *constant* that is added to every score and a *residual variable* that is added to the constant. The residual variable represents all other causes of the dependent variable besides the independent variable.

The model under consideration is called the *complete model.* In model testing a *restricted* version of the model is proposed that is identical to the

TABLE 12.6 **Statistical Procedures for Nonindependent Groups**

Intervally Measured Dependent Variable
Dichotomous independent variable: paired t-test, Chapter 13
Multilevel independent variable: repeated-measures analysis of variance, Chapter 15
Nominally Measured Dependent Variable
McNemar test, Chapter 17
Ordinally Measured Dependent Variable
Dichotomous independent variable: sign test, Chapter 18
Multilevel independent variable: Friedman two-way analysis of variance, Chapter 18

complete model, except that there is one constraint on one parameter of the complete model. This constraint is referred to as the *null hypothesis*.

In this chapter the complete model assumes that the dependent variable is equal to a constant plus the residual variable. In the restricted model the constant is fixed or set to some a priori value. If the restricted model were true, then the sample mean should differ from the a priori value within the limits of sampling error. Given the restricted model and the assumptions of random sampling, independence, and normality, the following quantity is distributed as t with $n - 1$ degrees of freedom.

$$\frac{\bar{X} - M}{s/\sqrt{n}}$$

where $\bar{X}$ is the sample mean, M the a priori constant, s the sample standard deviation, and n the sample size. The quantity $(\bar{X} - M)/(s/\sqrt{n})$ is called the *test statistic*.

The probability of obtaining a value as or more extreme than the test statistic is called the *p value*. If the p value is less than or equal to the significance level, then it is concluded that the null hypothesis is false and the test statistic is said to be *statistically significant*. If the p value is greater than the significance level, the null hypothesis is retained and the test statistic is said to be *not statistically significant*. The standard significance level is .05.

There are two major errors in model testing. A *Type I error* is rejecting the restricted model when, in fact, it is true. A *Type II error* is a failure to reject the restricted model when it is not true. The probability of making a Type I error is denoted as *alpha* and is set by the significance level. The probability of making a Type II error is denoted as *beta* and is determined by the sample size and other factors. Two other errors are (a) rejecting the restricted model not because the null hypothesis is false but because the assumptions are false, and (b) interpreting that the null hypothesis is false in the wrong direction.

Problems

1. For the following degrees of freedom, find the critical value for the following significance levels for the t distribution.

	df	*Alpha*
a.	12	.01
b.	23	.05
c.	76	.001
d.	209	.02
e.	17	.10
f.	48	.05

2. For the following t values and degrees of freedom determine the p value.

   a. $t(24) = -1.583$ b. $t(78) = 1.990$ c. $t(24) = 3.145$
   d. $t(19) = -3.117$ e. $t(28) = 2.963$ f. $t(77) = 1.942$

3. A prison official wishes to determine whether the inmates in a prison score above a national norm on a personality test. The scores of nine randomly chosen inmates are

   15, 18, 23, 41, 19, 25, 31, 43, 51

   The norm is 25. Are prisoners above the norm?

4. In a memory experiment, guessing would lead to a score of 10. The scores of six subjects are

   9, 15, 12, 17, 13, 10

   Is it reasonable to assume that subjects are guessing at this task?

5. What value would $(\bar{X} - M)/(s/\sqrt{n})$ have to equal or exceed (ignoring sign) to be significant at the .05 significance level for the following degrees of freedom?

   a. 15 b. 59 c. 25 d. 190

6. Explain the difference between a Type I and a Type II error.

7. Test the null hypothesis that the population means equals 50.

   63, 51, 43, 55, 60, 36, 40, 57, 54

8. Eight married couples were asked what proportion of the housework each did. The proportions were summed for both members. Test a restricted model that the constant is 100.

   109, 121, 98, 95, 105, 112, 123, 134

9. For the following two models, which is the restricted and which is the complete model?

$$\begin{matrix}\text{dependent} \\ \text{variable}\end{matrix} = \text{constant} + \begin{matrix}\text{effect of the} \\ \text{independent} \\ \text{variable}\end{matrix} + \begin{matrix}\text{residual} \\ \text{variable}\end{matrix}$$

$$\begin{matrix}\text{dependent} \\ \text{variable}\end{matrix} = \text{constant} + \begin{matrix}\text{residual} \\ \text{variable}\end{matrix}$$

   What is the restriction in the restricted model?

10. Imagine a psychologist who is interested in subliminal perception. Stimuli, either an A or B, are flashed on a tachistoscope. The subject responds

by saying whether an A or B was flashed. Each subject is presented with 15 trials. The number correct for ten subjects are

10, 15, 12, 6, 11, 9, 8, 12, 11, 8

a. Determine the constant if subjects were guessing.
b. Test the restricted model that subjects are only guessing in this task.

11. Fifteen different groups of subjects were asked to estimate the population of Phoenix, Arizona, in units of 100 thousands. The estimates are as follows:

9	8	12	10	8
6	9	6	9	6
11	11	7	7	5

The correct answer is 8 (hundred thousand). Do groups on this task tend to significantly over- or underestimate the population of Phoenix?

12. A company advertises that its cars get 30 miles per gallon gas mileage. An inquiring car dealer measures the miles per gallon of 20 cars. She obtains the following:

30, 25, 28, 30, 27, 34, 41, 25, 28, 30,
28, 35, 31, 34, 32, 31, 26, 31, 24, 32

What should she conclude about the manufacturer's claim?