3

CONCEPTUAL FOUNDATIONS OF STATISTICS

In this chapter, we examine the conceptual foundations of statistics. The goal is to give you an appreciation and conceptual understanding of some basic statistical tests used in educational research. As we suggested in the first chapter, statistics are tools that empirical researchers use for analysis of quantitative research. Statistical tools are useful means and not ends in themselves. We focus on conceptual understanding and not on the technical details of computing the statistics, which is most often done in statistical courses or by using statistical packages, such as SPSS (Statistical Package for the Social Sciences) or SAS (Statistical Analysis System). We begin with a review of some basic descriptive statistics and then move to the conceptual underpinnings of inferential statistics, which are used to test **research hypotheses**. Read the text with a pencil in hand; check the simple calculations.

MEASURES OF CENTRAL TENDENCIES

There are three common measures of central tendency: **mean**, **mode**, and **median**. The mean is the most widely known statistic; it is the average of a set of numbers or scores. Most students compute their average test scores in a course without difficulty, and they understand what it means or represents—it is their typical test score. *The arithmetic average of some set of numbers in statistics is called the mean*. Summing all the scores in the set and then dividing the sum by the number of scores is the calculation of the mean. Consider the set of numbers (1, 2, 2, 3, 4, 6). The mean is calculated as follows:

Mean = Sum of the scores (Σ (scores)) divided by N (number of scores)

or

$$\text{Mean} = \frac{\Sigma(\text{scores})}{N}$$

$$\text{Mean} = (1 + 2 + 2 + 3 + 4 + 6)/6 = 18/6 = 3.$$

The mean or average of the set is 3, which represents a typical score for this set of data points. If the scores are reasonably consistent—that is, they don't vary wildly—then the mean is a good indication of the central tendency. If there are a few extreme scores, however, the mean can be distorted. Consider the set of numbers (1, 1, 1, 7, 1, 7). In this case, the mean is still 3, but it is not really typical. A few large and extreme numbers can distort the mean, and therein lies the possible rub of using the mean to describe a set of scores as typical. For example, in the previous set of numbers (1, 1, 1, 7, 1, 7), 1 is clearly more typical than 3.

The mode is the most frequent number in a set of scores. In the above set (1, 1, 1, 7, 1, 7), the mode is 1, the most frequent number in the distribution, and in this case, it is a good standard to describe the typical score of this set of numbers. But again, just as with the mean, the mode can be misleading. For example, suppose you give a test to 30 students and most students score close to 88; in fact, when you compute the mean you get 88. Yet there were five people who got 100 and only three who actually scored 88. Which is the better measure of central tendency, the mean (88) or the mode (100)? Clearly, the mean is more typical of the distribution of scores.

The median is the middle score of a distribution of numbers. To compute the median, do the following:

1. *Rank the numbers* or scores from low to high.
2. *Find the middle* number or score:
    - If there is an odd number of scores, for example, 11 numbers in the set, simply add 1 to the total number and divide by 2; the resulting number represents how far to go to find the median. Consider the numbers in the set (1, 2, 2, 2, 3, 5, 6, 7, 7, 8, 12). Since the set has 11 numbers (an odd number), add 1 to 11 and divide by 2: 12/2 = 6. The sixth number in the set is 5, and it is the median or middle score.
    - But, if there is an even number of scores, simply average the two middle scores. For example, consider the set (1, 2, 2, 2, 4, 5, 6, 7, 7, 8), which has 10 scores in the distribution. You simply add the fifth and sixth scores and divide by 2; hence, in this example the median is (4 + 5)/2 = 4.5. The median is the middle score, which is 4.5 in this case.

The median is the middle score in the distribution of ranked numbers; it is the point at which half the numbers are larger and half are smaller. When there are a few very

high or very low scores, however, the median or mode may represent better the central tendency than does the mean.

In sum, the mean, mode, and median are the three most common measures of central tendency; they are indicators of how typical a given score is in a distribution of numbers, but none of these indicators gives you a sense of how the scores are distributed, that is, how much variability there is in the set of numbers.

MEASURES OF VARIABILITY

Let's now turn our attention to how much variability there is in a set of numbers. How are the scores distributed? How much do they vary? We consider three measures of variability: the **range**, the average deviation, and the **standard deviation** (*SD*).

The range is the difference between the highest and lowest scores in a set of numbers, but it is also given as the span of scores beginning with the lowest score and ending with the highest score, as in the range of 89 to 144 (or, alternately, the range is 55). The range is direct and simple, but a little crude because it only describes in broad strokes the limits of the scores; it does not tell us what is happening in between the extremes.

The average deviation from the mean is just what the phrase suggests: We find the mean, then find the deviation from the mean for each number (subtract the mean from the number), and then average all the deviations to get a typical departure of the scores from the mean. Conceptually, that makes sense, but unfortunately, we always get the same average deviation because half the scores will deviate above the mean and the other half below the mean; consequently, when you add the deviations you always get 0. Thus, the average deviation is always 0 and not useful. Take an example. Consider the set of numbers (1, 2, 3, 4, 5, 3, 3). The mean is 21/7 = 3. The deviations from the mean are –2, –1, 0, 1, 2, 0, and 0, respectively, and the sum is therefore 0. Zero divided by 7 is 0. Zero is always the average deviation from the mean because half the scores are above the mean and the other half are the same amount below the mean, and 0 divided by any number is 0. Try it yourself with a small set of numbers. Why bother with the average deviation from the mean? Only to help you understand the concept of a standard deviation from the mean.

The standard deviation *from the mean is the extent to which scores vary from the mean—the typical deviation from the mean for a set of scores*. The standard deviation is conceptually similar to the average deviation, but it is more useful because it is not always 0, and it has some interesting mathematical and statistical properties, which we discuss later. Remember, the standard deviation is always *from the mean*; the mean is the point of

reference. How much are the scores deviating from the mean? What is the typical or standard deviation of scores from the mean? Let's consider the same set of numbers as before (1, 2, 3, 4, 5, 3, 3), and illustrate the computation of its standard deviation.

- Compute the *mean* as we did above; it better still be 3, but check it.
- Compute the *deviations from the mean*; subtract the mean from each score.
- Square each *deviation from the mean*; check these computations below:
- Sum the squared deviations:

$$\Sigma\,(\text{Score} - \text{Mean})^2 = (4 + 1 + 0 + 1 + 4 + 0 + 0) = 10.$$

- Divide the sum of squared deviations by the number of scores:

$$\Sigma\,(\text{Score} - \text{Mean})^2 / 7 = 10 / 7 = 1.43.$$

- Take the square root of the quotient to obtain the standard deviation: Square root of 1.43 = 1.196.

Deviation From the Mean (Score – Mean)	Deviation From the Mean Squared $(\text{Score} - \text{Mean})^2$
$(1 - 3) = -2$	$(1 - 3)^2 = -2^2 = 4$
$(2 - 3) = -1$	$(2 - 3)^2 = -1^2 = 1$
$(3 - 3) = 0$	$(3 - 3)^2 = 0^2 = 0$
$(4 - 3) = 1$	$(4 - 3)^2 = 1^2 = 1$
$(5 - 3) = 2$	$(5 - 3)^2 = 2^2 = 4$
$(3 - 3) = 0$	$(3 - 3)^2 = 0^2 = 0$
$(3 - 3) = 0$	$(3 - 3)^2 = 0^2 = 0$

Hence, the standard deviation of this set of numbers is 1.196, and the formula is

$$\text{Standard Deviation } (SD) = \sqrt{\frac{\Sigma\,(\text{Score} - \text{Mean})^2}{N}}$$

One small note: Statisticians use the shorthand expression *sum of squares* to refer to the sum of the deviations from the mean squared, which often confuses students. So remember that you *square all the deviations from the mean* and then calculate the sum to get the sum of squares; then you divide by the number of scores and take the square root of this quotient to get the standard deviation. Now you have the formula for computing the standard deviation, but it is just as important to know what *standard deviation* means—the extent to which your set of scores vary from the mean—the larger the standard deviation, the more widely the scores vary from the mean (see Figure 3.1); when the standard deviation is small, the variability is also small.

Knowing the mean and the standard deviation of a group of scores gives you a better understanding of an individual score. For example, suppose you received a score of 79 on a test. You would be pleased with the score if the mean of the test were 70 and the *SD* were 4 because your score would be a little more than 2 *SD*s above the mean, a score well above average.

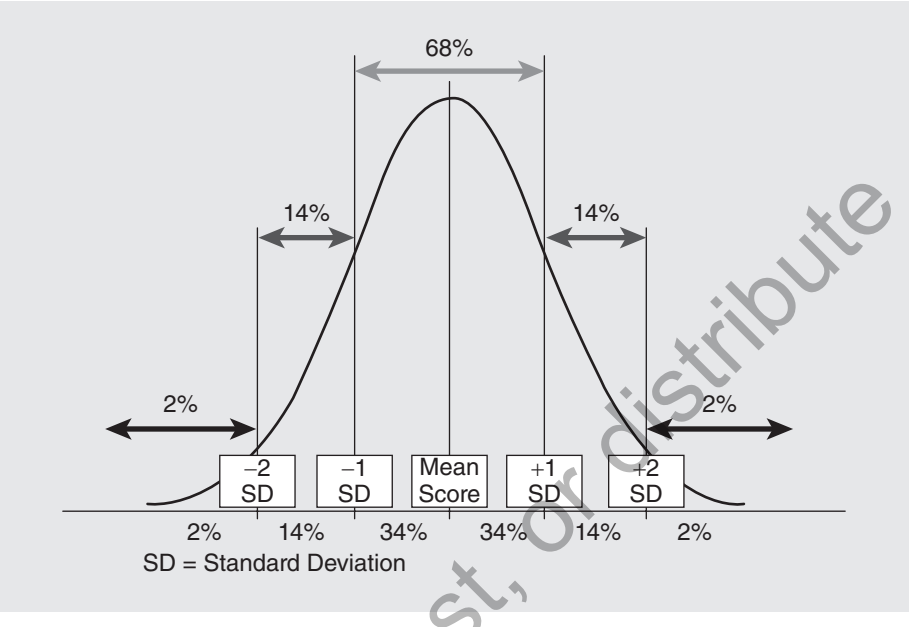
Consider the difference if the mean of the test had remained at 70, but the *SD* had been 16. In this case, your score of 79 would be less than 1 *SD* from the mean. You would be much closer to the middle of the group, with a score slightly above average, but not high. Knowing the standard deviation tells you much more than simply knowing the range of scores. No matter how the majority scored on the test, one or two students may do very well or very poorly and thus make the range very large.

NORMAL DISTRIBUTION

Standard deviations are especially useful if the distribution of scores is normal. You have heard of a **normal distribution** before; it is the bell-shaped curve that describes many naturally occurring physical and social phenomena, such as height and intelligence. Most scores in a normal distribution fall toward the middle, with fewer and fewer scores toward the ends, or the *tails*, of the distribution. The mean of a normal distribution is also its midpoint. Half the scores are above the mean, and half are below it. Furthermore, the mean, median, and mode are identical in a normal distribution.

As you can see in Figure 3.1, when the distribution of scores is normal, the percentage of scores falling within each area of the curve is known. Scores have a tendency toward the middle or mean. In fact, 68% of all scores are located in the area from 1 *SD* below to 1 *SD* above the mean. About 16% of the scores are beyond 1 *SD* above the mean. Of this higher group, only 2% are greater than 2 *SD*s above the mean. Similarly, only about 16% of the scores are beyond 1 *SD* below the mean, and of that group only about 2% are beyond 2 *SD*s below the mean.

Figure 3.1 The Normal Distribution



Standard scores are based on the standard deviation. A ***z* score** *is a standard score that tells how many standard deviations above or below the mean a score falls*. In the example described earlier, where you were fortunate enough to get a 79 on a test where the mean was 70 and the *SD* was 4, your *z* score would have been greater than 2 *SD*s above the mean (actually 2.25 *SD*s above the mean), which means that your score is higher than 98% of those who took the test. To determine your place in a normal distribution, you need to convert your raw score to a standard score, which is a simple process—simply subtract the mean from the raw score and divide the difference by the standard deviation. The formula is

$$Z = \frac{\text{Raw score} - \text{Mean}}{\text{Standard deviation}}$$

POPULATIONS AND SAMPLES

So far, all our statistics have described properties of **populations**. The population, or universe, *contains all the elements of the set*. If you have all the elements of the set you are studying, for example, all the scores for all students in your class, then you have the results for that universe. You can compute the *exact* or actual mean, mode, median, range, and standard deviation for the population; there is no need to estimate.

For the most part, however, researchers are interested in samples of a population. A **sample** *is a subgroup of the population*. If we want to generalize about the third-grade students in the country, the population is all-third grade students in America. It is impractical, if not impossible, to get information on all such students, so researchers limit the population to third grade students in a state. Even this population may be too large for practical purposes, so we take a subgroup of these students as a sample. We would like to get a representative sample so that our conclusions are general to the population.

We need to add a few more refinements to our definitions because we are usually directly concerned with samples rather than populations. *Statistics are the characteristics of samples. Parameters are the characteristics of populations.* That is, measures of central tendencies (mean, mode, median) and indicators of variability (range and standard deviation) are parameters, which are estimated from the sample. One formula, the standard deviation, needs to be altered slightly to get a better estimate of the actual standard deviation of the population. In other words, when using a sample to estimate the standard deviation of a population, divide by $n - 1$ (number in the sample minus 1). This revised formula yields a better estimate of the standard deviation for the population; this slightly altered way of calculating the variance is called the mean square and has other mathematical and statistical properties that make it useful. Thus, the standard deviation for a sample is best defined as

$$SD = \sqrt{\frac{\sum(\text{score} - \text{Mean})^2}{n-1}}$$

Thus far in our analysis, we have used the standard deviation as a measure of the variability. A related concept that is more useful in statistics is the variance of a set of scores. The variance of a sample is its standard deviation squared.

$$\text{Variance}\,(V) = \frac{\sum(\text{score} - \text{Mean})^2}{n-1}$$

The variance and the mean are the two key concepts used in most statistical analyses. Both are summaries of a set of scores; the mean is a measure of central tendency and the variance a measure of variability. We started our discussion of variability with the standard deviation because we assumed that it was more familiar, but now the related concept of variance becomes our chief index of variability.

Much of statistical analysis is explaining the variance in the dependent variable. *Does the independent variable cause the dependent variable to vary or lean in a certain direction?* That is the key problem of inferential statistics. We ask the question "Were the results of my study a consequence of the independent variable, or were they a result of chance?"

In other words, we measure our actual findings against the chance model. We attempt to eliminate chance as an explanation of our results in order to buttress the argument that our independent variable, not chance, made the difference. To reiterate the central thesis of inferential statistics, "Did the results occur by chance, or are they a function of our independent variable?" Statistics and probability help us answer this basic question. This is not a book on statistics; however, a basic conceptual understanding of statistical tests is essential if we are to grasp the nature and meaning of quantitative research.

STATISTICAL TESTS

One more time, here is the *basic statistical question*: Is what I found in my research significantly different from what I would expect to find by chance? What you as a researcher need to do is to compare your actual results with the chance model. Do the results vary enough from chance to conclude that something else is causing the variance or variability in the dependent variable? Statistics provide critical ratios, such as the *t* ratio, the *F* ratio, or chi square, which enable us to answer the chance question with confidence (see "Elements of a Proposal," Appendix A).

t Test

All critical ratios work the same way, and we illustrate a few so that you understand what is happening and why. The ***t* test** is a good place to begin because it is a clear, straightforward statistical application. If we are doing a study in which the independent variable has only two categories and the dependent variable is continuous, then the appropriate statistic is a *t* test.

For example, suppose we want to know if college men and women are significantly different with respect to liberal attitudes toward premarital sex. Note that the population is all students at College A. Assume that we select a representative sample of men and women from College A, and we have all students in the sample respond to a reliable and valid scale measuring their attitudes. Assume further that the higher the score on the scale, the more liberal their attitudes. How can we test the results of our little research problem?

First, we divide the sample into two groups, male and female; the independent variable has only two categories. Then we compute the mean scores for men and for women on the dependent variable—liberal attitudes toward premarital sex. Finally, we ask whether the means for the men and women were significantly different. *The t test is an appropriate statistical procedure when the independent variable has two and only two categories and the dependent variable is continuous.*

Here is how the *t* test works. To assess whether there is a significant difference (one not explained by the chance model), we compare what we found—the *actual difference* in scores between men and women—with the *difference expected by chance*. The ratio between the actual difference and the difference due to chance is a *t* ratio. A *t* test is defined as

$$t = \frac{\text{Actual difference in the means}}{\text{Difference expected by chance}}$$

The larger the ratio, the greater the probability that the difference is not a function of chance. If the ***t* value** is 1, what does that mean? The actual difference between the means is exactly what to expect if nothing but chance is working; chance is explaining this relationship. But, if the *t* value is 2, it is more likely that something other than chance is operating.

Let's continue a little further without getting bogged down in statistical calculations. The general formula for a *t* test is as follows:

$$t = \frac{\text{Mean from Group 1} - \text{Mean from Group 2}}{\text{Standard error of the difference between the means of the two groups}}$$

There are several important aspects of this general formula:

1. We are examining the actual difference between the means of the two groups.
2. We are comparing the actual difference with what is expected by chance.
3. Statisticians can determine what is expected by chance by computing the standard error of the difference between the two means.
4. A *t* ratio is computed that indicates the extent to which the results depart from the chance model: The greater the *t* value, the greater the likelihood that chance is not explaining the relationship.

Fortunately for us, using any one of a number of statistical packages, the computer calculates the standard error of the difference between the means as well as the *t* ratio and its **level of significance** (*p* value).

A ***p* value** is a probability level that indicates the level of significance, that is, *the probability that the results are a function of chance*. When you read research publications,

you find statements like ($t = 2.62, p < .01$). This means that a *t* test produced a *t* ratio of 2.62, which was significant beyond the .01 level of significance ($p < .01$); hence, we can be quite confident that the chance model does not explain the relationship. By convention, most researchers accept a relation as statistically significant if the *p* value is equal to or less than .05. What that means is that the relation could have occurred by chance only 5 times or less out of 100.

Let's return to the question of whether men and women at College A have different attitudes toward premarital sex.

- First, we add up all the scores for the men and divide by the number of men (mean score for men) and do the same for the women (mean score for women).
- Next, we subtract the scores (mean score of men minus mean score of women).
- Then, we compute the standard error of the difference between the means of men and women (the difference we would expect to get by chance).
- Finally, we compare the two by computing a *t* ratio (actual difference divided by the standard error of the difference).

Fortunately, our laptop and SPSS computer program does all this as quickly as we can hit the Execute button. The results include the *t* value and give us its level of significance.

What would it mean in our research project if we obtained the following: ($t = 1.02$, $p > .95$)? The answer is that a *t* value of 1.02 is not statistically significant. We can tell this just by looking at the *t* value because 1 would indicate perfect chance to explain the result. The $p > .95$ indicates that more than 95 times out of 100, chance would explain our results. Hence, in College A, we can conclude with great confidence that there is no significant difference between men and women in their attitudes toward premarital sex.

F Test

The independent variable is not always a dichotomous variable, one with only two categories. Sometimes the independent variable has more than two categories. If so, we cannot use the *t* test. We need a more general test that does essentially the same thing, that is, produces a critical ratio to check the departure from the chance model. In a case when there is more than two categories in the independent variable and a continuous dependent variable, the more general *F* ratio provides our answer. An *F* test

is done using a statistical procedure called **analysis of variance (ANOVA)**. There are a variety of ANOVAs, and we now focus on the least complex; however, conceptually all ANOVAs are similar in that one or more ***F* values** are computed to answer the question of the deviation from the chance model question.

Let's illustrate a simple one-way ANOVA with an example. Suppose I want to test the effectiveness of three teaching approaches with graduate students in education—teacher directed, student directed, and shared. I am teaching a large group of 90 students in an introductory course in education, about a third of all the beginning graduate education students. *Does my teaching approach make any difference in the mastery of key concepts in education?* Assume that I can divide the group into three similar subgroups; probably the best way to do this is to assign the students to the groups at random. Assume further that the 90 students are representative of all beginning graduate education students at my university.

What is my independent variable in the research problem? How many independent variables do I have? Three? No, actually I have only one independent variable (teaching approach) with three variations or categories (teacher-directed, student-directed, and shared approaches). The independent variable is a manipulated categorical variable. I, the researcher, will manipulate the variable by teaching each group in one of three ways. What is the dependent variable? I am interested in mastery of basic education concepts, and I have a final exam that I developed over the years that is reliable and valid; that is, it taps the content that I am interested in having students master in a consistent manner. The dependent variable is measured by my test and is continuous: The higher the test score, the greater the level of mastery of basic concepts. *The F test is an appropriate statistical procedure when the independent variable has two or more categories and the dependent variable is continuous.*

Here is how an *F* test is computed using ANOVA. At the end of the term, I compute the mean score on mastery for each of the three groups. Almost certainly, there will be differences in the means, but the question is essentially the same here as it was for the *t* test: Is there a *significant* difference *among* the three mean scores? I proceed by doing the following:

- First, compute the mean for each of the three groups on the mastery exam.
- Next, calculate the total variance for the entire sample. That is, combine all three groups into one, and compute the overall mean for the entire 90 students. To compute the variance for the entire group, which is called the total variance (V_t), use the following formula described earlier:

$$\text{Variance (V)} = \frac{\sum(\text{score} - \text{Mean})^2}{n-1}$$

- Now compute the variance between the groups. To do this, we treat each of the three means for the groups as data points and use our variance formula. The **between-group variance** is the variance caused by the independent variable; it is also called **systematic or experimental variance**.
- The variance due to error is commonly called the **within-group variance**; it is also called **error variance**. This computation is a little more difficult to explain, but conceptually it is the variance "left over" from the total variance after the between-groups or experimental variance is removed from the total variance. The within-group variance is a measure of chance variation.
- Finally, calculate the *F* ratio, which is the variance produced by the independent variable divided by the variance due to chance.

$$F = \frac{\text{Variance due to the independent variable}}{\text{Variance due to change}} = \frac{\text{Between-groups variance}}{\text{Within-groups variance}}$$

We have come a long way to show that the *F* ratio using ANOVA is essentially the same as a *t* ratio in that both compare actual findings in relation to chance and yield an index and a probability level to enable us to make confident judgments about the nature of our relationships. A significant *F* ratio in this kind of problem simply means that there is a significant difference *among the three groups*. To find which pairs of means are different, we must do some further post hoc analyses, which can be found in any good statistics book. But, the idea is the same: *Compare your actual results with what you would expect by chance*.

Chi-Square Test

Sometimes, both the independent and the dependent variables are categorical. If so, we need another statistical tool called the **chi-square (χ^2)** test to compute the critical ratio for such situations. Suppose you want to examine the relationship between gender and graduation. *Is the gender of the students related to whether or not one graduates?* What are the independent and dependent variables of this research problem? What kind of variable is each in terms of measurement? Gender is the independent variable; it is the presumed cause and has two variations or categories: male and female. Graduation is

the dependent variable, and it also has two categories: graduate and no graduate. One might also consider graduation as a continuous variable, that is, graduation rate, but to illustrate the chi-square, we cast graduation as a dichotomous variable.

We decide to go back to the freshman class of four years ago and see how many men and women graduated at the end of four years. We select a random sample of 100: 50 men and 50 women. We summarize the results of our research in a 2 × 2 cross break or contingency table (see Table 3.1).

Table 3.1 Summary of the Results of our Analysis

	Men	Women	
Graduate	15	35	50
No graduate	35	15	50
	50	50	100

As we examine the results in the 2 × 2 table, we see that in our sample, women may be more likely to graduate than men, but what is the likelihood that the results can be explained by chance? In other words, we need to *compare what we found in this analysis with what we would expect to find by chance*. Do the results here represent a major departure from the chance model? We need a critical ratio. *The chi-square test is the appropriate statistic when both variables are categorical*. The chi-square is a test of frequency counts. What do the numbers in the cells of our 2 × 2 table represent? Yes, frequencies—the number of students in each cell. The chi-square is an index of the actual results compared with those expected by chance. Examine the formula for chi-square:

$$\text{Chi-square}(\chi^2) = \sum\left[\frac{(f_0 - f_e)^2}{f_e}\right]$$

Now, we use the formula and the previous results obtained and summarized in our 2 × 2 cross break. The chance model would predict 25 students in each cell; that is, the expected frequency for each cell (f_e) is 25. Now compare the expected with the actual for each cell by subtracting the expected frequency (f_e) from the observed frequency (f_o), squaring the difference, and then dividing the difference by the expected frequency (f_e). Let's do the computations for each cell and sum them as the formula instructs.

$$\chi^2 = [(15 - 25)^2/25] + [(35 - 25)^2/25] + [(15 - 25)^2/25] + [(35 - 25)^2/25]$$

$$\chi^2 = [100/25] + [100/25] + [100/25] + [100/25]$$

$$\chi^2 = 4 + 4 + 4 + 4$$

$$\chi^2 = 16.$$

The χ^2 index is 16. What would the chi-square have been if only chance were working? Each cell would have had the number 25, and χ^2 would have been 0. Run the numbers, and make sure you see why the answer is 0. Thus, our index of departure from chance in this example is 16. Using a computer program, we would have found ($\chi^2 = 16$, $p < .01$). The results show that a χ^2 of 16 is statistically significant beyond the .01 level of significance; that is, these results would occur by chance less than 1 time out of 100. Our conclusion would be that women are more likely to graduate from the college than are men. Note that, as always, our conclusion is probabilistic, not certain. The point of this exercise is to demonstrate the meaning of yet another critical ratio, one that works when both variables are categorical.

Effect Size

The three tests that we examined thus far—the *t* test, the *F* test, and chi-square—are statistics that help us answer the basic statistical question: *Is what I found in my analysis significantly different than I would expect to find by chance*? None of these statistics, however, tells us anything about the magnitude of the relation. Increasingly, researchers want to know the strength of the relation, that is, its **effect size**. *The magnitude of the independent variable's effect on the dependent variable is the effect size*. Suffice it to say that when using *t* tests, analysis of variance, or chi-square analysis, we must do additional computations to determine effect size. For example, a contingency coefficient and an omega-squared (Hays, 1994; Kerlinger & Lee, 2000) are relatively straightforward computations that will tell us the magnitude of the effect size. The point here is that the *F* and *t* values and chi-square tell us if there is a statistically significant relationship, but they do not indicate the magnitude of the relation; other indices are needed.

We turn next to coefficients of correlation, which not only answer the question of statistical significance, but also indicate the magnitude of the relationship between the independent and dependent variable—the proportion of variance in the dependent variable explained by the independent variable. Correlation coefficients, unlike the statistics explored thus far, answer *both* the statistical significance and the effect size questions.

Linear Regression and Coefficient of Correlation

What if both the independent and the dependent variables are continuous? We need another statistic: A **coefficient of correlation (*r*)** will give us the answer to whether the relation is likely a chance one or not. But, another useful feature of the correlation is that we can use it to test not only the departure from the chance model, but also the strength of the relation. A coefficient of correlation *is a number that indicates the magnitude of the relation between two continuous variables such that the higher the absolute value of the correlation, the stronger the relation.* Correlations range in value from –1 to +1. If the two variables vary together, they have a positive correlation, which means that as the value of one increases, so does the other. If the correlation is negative, then as the independent variable changes, the dependent variable changes in the opposite direction. Which is stronger, a correlation of +1 or one of -1? Neither. Both are perfect correlations; both are as high as they can get, but in opposite directions. The sign of the correlation represents the direction of the relation and has nothing to do with its strength. So $r = -.85$ is a stronger correlation than $r = +.41$ because the sign merely indicates whether the variables are varying in the same or opposite direction.

The calculations of coefficients of correlations are a little more tedious and not as self-evident as the other statistics that we have discussed, so we will not spend much time with the formulas and computations. Instead, we illustrate the correlations with a table. Correlations describe linear relations, which are straight lines when graphed. The relation between two variables, x and y, is a set of ordered pairs. That means that for every value of x there is one corresponding value of y. We can express the pairs of values in set notation, or we can simply express them in a table or graph or both. Consider the relations between *three* sets of ordered pairs (relations) as expressed in Table 3.2.

The first set of ordered pairs (1) has a correlation coefficient of +1; the numbers vary together. For each change in the independent variable x, there is a corresponding change in the dependent variable y of the same magnitude and direction. In the second set of ordered pairs (2), sometimes a change in x produces a positive change in y and sometimes a negative change; there is no systematic pattern in the relation; there is no relation ($r = 0$). Finally, in the third set (3), for each change in x there is a corresponding change in y of the same magnitude except in the opposite direction; we have a perfect **negative correlation** ($r = –1$); x and y vary together in opposite directions. In brief, the correlation coefficient provides an index of the extent to which the two variables vary together and the direction of the variation.

A computer program will provide you with correlation coefficients and levels of significance. Consider the statement ($r = –.52$, $p < .01$). The correlation is negative:

Table 3.2 Correlations for Three Sets of Numbers

(1) $r = 1.00$		(2) $r = 0$		(3) $r = -1.00$	
x	*y*	*x*	*y*	*x*	*y*
1	1	1	2	1	5
2	2	2	5	2	4
3	3	3	3	3	3
4	4	4	1	4	2
5	5	5	4	5	1

As *x* increases, *y* decreases. The relation is statistically significant; that is, chance is unlikely to explain the relationship; in fact, in less than 1 time in 100, the two variables would not be related. The correlation coefficient also suggests how strong the relation is between the two variables. Square the coefficient of correlation and multiply it by 100, and you have an estimate of the percentage of the variance in the dependent variable (*y*) caused by the independent variable (its effect size). For example, if $r = .50$, then the independent variable *x* explains 25% of the variance in *y*. If $r = 0$, then none of the variance in *y* is explained by *x*. If $r = -.83$, then about 69% of the variance in *y* is explained by *x*. An important point: *What scientists try to do with their research and statistics is to identify independent variables that explain the variance in the dependent variable*. Explaining variance in a dependent variable is an important goal of scientific research.

A final observation about a correlation coefficient—it is mathematically the coefficient of *x* in the formula for a straight line, as expressed by the following equation:

$$y = mx + b.$$

Think of the set of ordered pairs that represents the relation between the independent variable, *x*, and the dependent variable, *y*, as a graph of a line that passes through those points such that the line represents the best fit for all the points; mathematically, that means the sum of all the distances from the points to the line (sometimes called a regression line) would be as small as possible. If we standardize *x* and *y*, then the coefficient of *x* is the correlation coefficient for the regression line for the relation of *x* and *y*. In sum, a correlation coefficient for a relation in which both variables have been standardized is the slope of its regression line. The regression line for two variables will take the form of $y = mx + b$, where *m* is the slope of the line and the correlation coefficient for the standardized data and *b* is the *y*-intercept. Perhaps we are getting a little too technical, so let's move on.

Multiple Regression and Multiple Coefficient of Correlation

Thus far, all the tests that we describe are bivariate; that is, they examine the relation between *one independent and one dependent variable*. In the actual world, relationships are more complicated. Typically, dependent variables are influenced by more than one variable at a time; thus, we need multivariate statistics. You should be beginning to realize that there are statistics for just about any relation you can imagine, but most are designed to answer basic questions: Can I reject the chance model as a good explanation? How strong is this relationship?

Just as a simple correlation (r) tells us whether the chance model can be rejected and how strong the relation is between x and y, a **multiple correlation (R)** tells us the same thing. But, in the case of the multiple R we have a little more information because *R represents how much variance in the continuous dependent variable y is explained by a set of continuous independent variables ($x_1, x_2, x_3 \ldots x_n$)*. Moreover, each x variable has a coefficient, which is sometimes called a regression coefficient or **beta weight**. So, a multiple regression analysis produces a multiple R, which represents the combined influence of all the independent variables on the dependent variable y, as well as a regression coefficient or beta weight for each independent variable (x). The coefficients represent the strength of the relation between that x and the dependent y, controlling for the other xs—that is, taking out the influence of the other independent variables. Consider the following formula for a multiple regression line:

$$y = ax_1 + bx_2 + cx_3 + i \text{ (the intercept).}$$

Note that this equation is simply an extension of simple regression; multiple regression is an extension of simple regression where there are multiple independent variables predicting a single dependent variable. In the regression equation above, we have three independent variables instead of only one. For example, we might be trying to predict student achievement (y) based on the IQ (x_1), motivation (x_2), and sense of optimism (x_3) of students. If we had data from some sample of students on these variables, we could use a standard statistical package to run a multiple regression analysis on this set of variables. The analysis would first compute an R, which would tell us how strong the relation is between this set of variables and student achievement. For example, what is the combined impact of IQ, motivation, and sense of optimism on student achievement? If we square the R, then R^2 is a good estimate of how much of the variance in student achievement is explained by the combination of IQ, motivation, and sense of optimism. The program also computes a t value or F ratio to gauge the likelihood that the relation is a matter of chance. Furthermore, the analysis yields a standardized beta coefficient for each independent variable, which tells us how much influence each independent variable has relative to the other independent variables,

and, of course, for each coefficient there will be a corresponding test to determine its departure from chance.

Remember that the two variables in simple correlation are both continuous; this is also the case in multiple regression—all the variables are typically continuous.

Hierarchical Linear Modeling (HLM)

HLM is simply multiple regression performed with hierarchical data, data that are clustered or nested within different units of analysis. For example, in school effects research, students are nested in schools. There are two different levels of analysis: the individual (students) and the school.

Why use HLM if it is just multiple regression? Let's go back to the purpose of multiple regression—to test the relationship among a set of independent variables and a single dependent variable. But, what if the question is how important is the school versus the individual in influencing some dependent variable? HLM enables us to answer this question efficiently. A standard multiple regression analysis of student achievement across schools does not simultaneously account for both student and school effects; HLM does.

To estimate a school effect in standard multiple regression, the researcher needs to aggregate individual student data to the school level by calculating a mean student score for each school. In the process of aggregating, however, we lose the differential effect of individual students on the dependent variable. So, how does HLM differ? Let's say we are interested in the relationship between collective faculty trust in students and student achievement. This question involves hierarchical data because student achievement is measured at the individual student level, but collective faculty trust is measured as a school property. Students are nested in schools; therefore, it is necessary to separate the variance in student achievement due to differences in schools and differences in individual students. In our example, we might find that 20% of the variance in student achievement is attributed to school differences while the other 80% is due to individual differences or to chance.

Remember the purpose of research is to explain the variance of a dependent variable, in this case achievement. There are several student factors that may explain differences in student achievement at the individual student level. Likewise, there are many school factors that may explain achievement differences across schools. We need to include these multiple variables in the regression model. In our example, we might use a measure of SES (socioeconomic status) as well as IQ at the individual level, and a measure of school poverty along with collective faculty trust at the school level. With this model, we now are accounting for achievement differences due to

student SES, IQ, school poverty, and collective trust of the faculty. In this case, the four independent variables are student SES, IQ, school poverty, and collective faculty trust. Note that the independent variables are at different levels—student SES and IQ are individual variables while school poverty and collective faculty trust are school variables. The reason we use HLM is because the multiple independent variables are at two different levels.

Thus far, our analyses focused on examining the relations between independent variables and one dependent variable. When statistical analyses simultaneously test relations among *both* multiple independent and multiple dependent variables, the procedures are called multivariate. We could continue building our set of statistical procedures. For example, what if we are interested in multiple independent variables and multiple dependent variables? There are, of course, statistical tests for such circumstances, multivariate analysis of variance (MANOVA), canonical correlation, and structural equation modeling (SEM). But, we have gone far enough to give you a flavor of statistics, what they do, and when and how they are employed.

SUMMARY

If you have carefully read and studied this chapter, you now have a good working repertoire of statistical procedures and tests; however, the chapter is not a substitute for a set of statistics courses, but it should provide the conceptual understanding that you need to begin to analyze and frame quantitative research. Let's review the key points:

- The mean, mode, and median are measures of central tendencies.
- The range and standard deviation are measures of variability.
- The basic purpose of inferential statistics is to answer the question "Were the results of my study a consequence of the independent variable, or were they a result of chance?"
- Your inventory of statistical tests includes
  - the *t* test for the difference between two means,
  - ANOVA and the *F* test,
  - the chi-square test,
  - coefficients of correlation (*r*),
  - multiple regression and multiple correlations (*R*), and
  - hierarchical linear modeling (HLM).
- Which test is appropriate depends on the nature of the independent and dependent variables, that is, whether they are continuous or categorical and nested or not (see Table 3.3 for a summary).

Table 3.3 Types of Variables and Appropriate Statistical Tests

Independent Variable	Dependent Variable	Statistical Test
Dichotomous	Continuous	*t* test
Categorical	Continuous	*F* test (ANOVA)
Categorical	Categorical	Chi-square (χ^2)
Continuous	Continuous	Correlation (*r*)
Multiple and continuous	Continuous	Multiple correlation (*R*)
Multiple, continuous, and nested	Continuous	Hierarchical Linear Modeling (HLM)

CHECK YOUR UNDERSTANDING

1. An educational researcher conducted an experiment with two groups: an experimental group (A) and a control group (B). A was a taught using "dynamic inquiry," and B was taught in a traditional way. At the end of the unit, a performance test was given to both groups, and their scores were as follows:

A	B
3	6
5	5
1	7
4	8
2	4

Using the formulas in this chapter, compute the mean, standard deviation, and variance for Group A and Group B. Based on the results, develop a hypothesis relating dynamic inquiry and effectiveness.

2. The following scores are the result of a test of reading comprehension in a fourth grade class:

0, 2, 4, 1, 3, 5, 2, 4, 6, 6, 4, 2, 5, 3, 1, 4, 2, 0

What are the mean, mode, and median for this set of scores? What are the range, average deviation, and standard deviation? In your own *words*, not in statistical terms, describe the variance and central tendency of this distribution.

3. A student score of 600 on the SAT (Scholastic Aptitude Test) is the same as a standard score of 1. How does this student compare with all those who have taken the test? What if the SAT score is 300 or a standard score of -2? What is a standard score? (*Hint:* For the SAT test, the mean score is 500, and *SD* = 100.)

4. Compute a *t* value for Exercise 1 above, assuming that the standard error of the difference between the two means is 1. Interpret what that *t* value means. Is the difference in the means of the two groups statistically significant?

5. You computed a correlation between the socioeconomic status of your students and their math achievement scores ($r = .70$). Interpret what this correlation means. If this is a true correlation, what can you as a teacher do to improve performance? How much does SES help or hinder your task?

6. You just read an interesting article where the researcher shows that the multiple regression of home background (HB), intelligence (IQ), and motivation (M) on achievement produces an R^2 of .87 and the standardized beta weights are .31, .41, and .34, respectively. How strong is the relation? Which variable is the most important in explaining achievement? What is the relative influence of each of the independent variables? What conclusions can you draw?

7. A school district wants to examine the influence of academic optimism on student achievement. Data are collected for academic optimism of the schools, the socioeconomic status of students, the attendance record of students, and the mathematics achievement of students. What are the independent variables? Dependent variable? School-level variables? Individual-level variables? What kind of analysis is required? Why?

KEY TERMS

ANOVA (analysis of variance) (p. 55)
Beta weight (p. 61)
Between-group variance (p. 56)
Chi-square (χ^2) (p. 56)
Coefficient of correlation (*r*) (p. 59)
Effect size (p. 58)
Error variance (p. 56)
Experimental variance (p. 56)
***F* value (p. 55)**
Hierarchical linear modeling (HLM) (p. 62)
Level of significance (p. 53)
Mean (p. 45)
Median (p. 45)
Mode (p. 45)
Multiple correlation (*R*) (p. 61)
Multiple regression (p. 61)
Negative correlation (p. 59)
Normal distribution (p. 49)
Population (p. 50)
***p* Value (p. 53)**
Range (p. 47)
Research hypothesis (p. 000)
Sample (p. 51)
Standard deviation (p. 47)
Standard score (p. 50)
Systematic variance (p. 56)
***t* Test (p. 52)**
***t* Value (p. 53)**
Within-group variance (p. 56)
***z* Score (p. 50)**