8

Precision of Inference

When Joe Cawledge obtains a verbal test score of 30, we report his raw score, his standard score, or one or more percentiles. If someone wanted to know the mean score of a group of sophomores in a psychology experiment, we would have the same options there. Of course, neither raw score by itself would have any meaning, but what about the other two (standard score and percentile)? You learned in Chapter 5 to interpret them, didn't you?

Well, yes and no. We have briefly discussed the problem of generalizing from a sample: the question of whether it is appropriate, once we have measured a sample, to make statements about some population that we say the sample represents. Really, though, that is not an either–or question. A proper answer would have to indicate the level of confidence that we have in that answer, and the probable *error* of such an inference is often as important to know as the inference itself.

This chapter then, is all about errors.[1] Its title is not misleading, however, for it is by discovering the probable limits of error that we define the precision of our inferences. In many applications, that definition is actually more important than the score. There is no point in estimating parameters unless our estimates are *reliable*.

STANDARD ERRORS

The concepts "sampling distribution," "sampling error," and "standard error" are closely related. A *sampling distribution* is a distribution comprising estimates (from measurements of samples) of the magnitude of some property of a population; *sampling error* refers to the dispersion of those estimates; and the *standard error* is a measure of that dispersion—namely, the standard deviation of the sampling distribution.

An example will make this clear.

The Standard Error of the Mean ($s_{\bar{X}}$)

One of the most important properties of a sample mean is that in a normally distributed population it is the most stable measure of central tendency. By "most stable" I mean that it varies least among repeated random samples taken from the same population. Let us say we are interested in finding the mean Wechsler Adult Intelligence Scale IQ of American college students. For simplicity, let us assume that the population itself remains stable during the time that it takes us to sample it. Such a population would probably be distributed normally with respect to IQ scores; let's assume that it is. It is too large to be described, so we must draw a sample and infer from it the mean IQ of the population.

There are more ways than one of selecting a sample. Conceptually, the simplest method is *random* selection, in which the composition of the resulting sample is entirely a matter of chance. The ideal case would be the equivalent of picking numbers out of a lottery bowl: Each number would represent one student; the entire bowlful, all students. Then we would test all of the persons whose numbers had been drawn (our sample of the population with respect to IQ).

Now back to the stability of the mean. If we were to select, say, 1000 individuals in the manner described above, we could compute a mean of their IQs. If we were to return that 1000 to the population and then select another sample of the same number in the same manner, we could compute a second mean. Returning those subjects to the population, we could select a third sample, and a fourth, and so on ad infinitum. Each sample would have a mean, and not all of those means would be the same. In fact, they would form a distribution—a distribution having the same bell shape as that of the population and of each of the samples.[2,3] It would, however, be a much more compact distribution than any distribution of individuals. Figure 8-1 depicts distributions of the population, of a single sample, and of the means of many samples on a common scale of IQs.

In Figure 8-1, a common symbol for the mean of the means ($\mu_{\bar{X}}$, pronounced "mew sub ecks bar") has been placed at the mean of the population (μ_X, or "mew sub ecks"), and indeed if an infinite number of means were taken, their mean would be the same as the population mean. But, of course, since that is never done, we can never be sure precisely where the population mean really is. (Remember, we know the population only through the samples that we draw from it.)

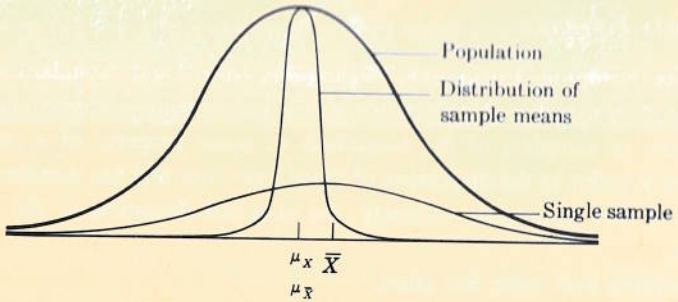


FIGURE 8-1 Superimposed distributions of the population, a sample, and an infinite number of sample means. μ_X = mean of the population (hypothetical); $\overline{X}$ = mean of one sample (obtained); and $\mu_{\overline{X}}$ = mean of an infinite number of sample means (hypothetical). The mean of the population is an example of a parameter; the mean of a sample is an example of a statistic. Note that the mean of the infinite number of sample means is the same as the mean of the population.

What we *can* do is discover the limits within which the population mean probably lies. The horizontally compact configuration in Figure 8-1 is the distribution of sample means; *when it is narrow*, we know that *the population mean is near our sample mean*, because when it is narrow, the population mean is near *every* sample mean.

So we have found that which we sought—pinned it down precisely! Unfortunately, however, it is only in fancy that we can take a very large number of samples. In real life, we get only one; so how does all this really help? It helps because although a knowledge of the sample mean will never tell us what the population mean is, we can *estimate* from the variability and the size of the sample what the variability of a distribution of sample means would be. Our estimate of the standard deviation of a hypothetical distribution of sample means is called a *standard error of the mean*, and its defining formula does include both the variability and the size of the sample:

$$s_{\overline{X}} = \frac{s}{\sqrt{n}} \tag{8-1}$$

where $s_{\overline{X}}$ is the estimated standard error of the mean, s is the estimated standard deviation of the population, and n is the size of the one sample that you have observed. From here on, however, I'll refer simply to "the standard error of the mean" without saying "estimated," because you will use it only when you lack access to the entire population of interest. If you *had* that access, you wouldn't *need* a standard error, because there would *be* no error.

BOX 8-1 Calculation of Standard Error of the Mean for the Collegiate IQs Example

$$s_{\overline{X}} = \frac{s}{\sqrt{n}}$$

$$s_{\overline{X}} = \frac{10.2}{\sqrt{10{,}000}} = \frac{10.2}{100} = 0.102$$

To calculate the *standard error of the mean*, all you need is the *size of the sample* and the *estimated standard deviation of the population*. In this case n = 10,000 and s = 10.2.

$$n = 10{,}000$$

$$s = 10.2$$

And of course you can't calculate that standard deviation unless you know the *mean of the sample*:

$$\overline{X} = 115.5$$

Formula (8-1) implies that the standard error is *small* when the sample is characterized by *small* variability in a *large* number of scores. A small standard error implies that samples of this size and variability have means that tend to be very close to each other. That in turn implies that the mean of our sample probably does not differ very much from the means of other samples that we might have taken from the population "collegiate IQs." It means that our obtained statistic is *reliable*.

Since you already know how to compute a standard deviation (s), the calculation of the standard error of a mean is straightforward, as you can see by referring to Box 8-1.

Other Types of Standard Error

Throughout the foregoing discussion of standard errors, actually only one kind—the standard error of the mean—has been presented. That limitation was imposed

because, although standard error may be a difficult concept to grasp initially, once it is understood in one situation it can be readily applied to others. The same logic applies to the standard error of a standard deviation, of an obtained score, of a difference between means (Chapter 9), of a correlation coefficient, and many others. In every case, a *small* standard error tells us that the sample statistic is a reliable estimate of the corresponding population parameter.

CONFIDENCE INTERVALS AND LEVELS OF CONFIDENCE

We now have a powerful tool to use in our attempt to locate the population mean, because now we can try some hypotheses and discover the probabilities of their being valid. Figure 8-2 indicates graphically how that might be done in the case of

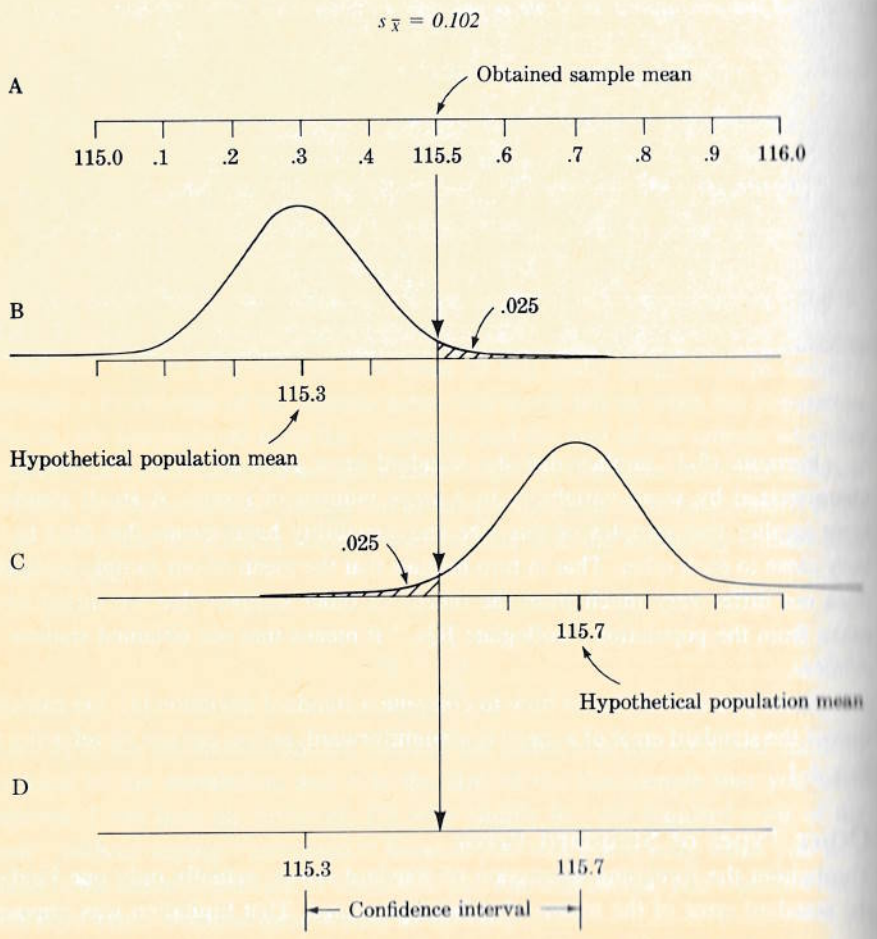


FIGURE 8-2 Establishing a confidence interval at the .95 level of confidence.

the mean IQ of college students if the standard error of the mean were $s_{\bar{X}} = 0.102$ (as it would be if calculated from the sample statistics listed in Box 8-1).

Essence of the Concept

The obtained mean IQ is 115.5. Figure 8-2 shows how that statistic plus the standard error of the mean can be used to answer the question, "Within what *limits* may we be reasonably sure that the population mean resides?" The interval enclosed by those limits is called a *confidence interval*, and the "reasonably sure" in the above question is expressed quantitatively as a *level of confidence*. Each of these important concepts depends on the other for its meaning. If I were to tell you that I am 100 percent sure that the true mean is merely *somewhere*, what would you learn about its location? And what information would you acquire if I were to say that the true mean just *may* be between 115.3 and 115.7? You have to know the interval if the level is to mean anything, and vice versa.

Actually, the confidence interval and the level of confidence are directly proportional to each other. Given a small standard error of the mean, we can say with *high* confidence that the true mean is within a *large* interval; but reducing the size of the interval reduces our confidence that the true mean is within that reduced interval. A moment's thought will convince you that this must be so in all cases.

You will recall that there is a variance in the means of random samples from the same population. That being the case, if the population mean were, say, 0.2 of a point *below* the obtained sample mean (Figure 8-2B), how many of a thousand sample means would be as *high* as the one we got? In the drawing, you can see that .025, or 25 out of every thousand, sample means would be that high. We can now state that the population mean is not lower than 115.3, and we can state it with a level of confidence of $1.000 - .025 = .975$.

The lower limit is only half of what we need to define a confidence interval, but with that behind us, the rest is easy. All we do is to repeat the above process, except this time we test the hypothesis that the population mean is *above* the obtained mean. Now we ask, "If the population mean were 0.2 of a point *above* the obtained sample mean (Figure 8-2C), what is the probability that a sample mean as *small* as ours would occur?" Again the probability is .025.

Testing our first hypothesis informed us that there is a probability of only .025 that the obtained mean is lower than 115.3; the second test yielded a probability of .025 that it is higher than 115.7. Now let us put the two hypotheses together: "What is the probability that the population mean lies outside the interval 115.3–115.7?" The answer is, of course, $.025 + .025$, or .05. So if we say that the population mean is somewhere within that interval, we can do so at a *confidence level* of $1.0 - .05 = .95$, and we can speak of it as the *95 percent confidence interval*.

Some Exercises to Consolidate the Concept

The amount of deviation of an obtained mean from a hypothetical population mean is somewhat arbitrary; it depends on where you imagine the true mean to be. But the

shape of the distribution of sample means in Figure 8-2 is *not* arbitrary; it is estimated from sample size and variability by computing the standard error of the mean.

The deviation of an obtained mean from a population mean is a deviation score; the standard error of the mean ($s_{\bar{X}}$) is essentially a standard deviation. A deviation score divided by a standard deviation is a standard score. So when we divide our imagined deviation by our hypothetical standard deviation, we obtain a standard score. In Figure 8-2B, the deviation score is 0.2 and the standard deviation is 0.102; the resulting standard score is 0.2/0.102 = 1.96. If we were to consult a table for a normal curve with a standard score of 1.96, we would find that the area in the smaller portion of the curve is indeed .025. If you were learning to *do* inferential statistics, there would be such a table at the back of this book, and you would get much practice in using it. However, since your objective is to *understand* inferential statistics, it is better that you deal with such problems graphically. The following exercises will give you several opportunities to do that. Don't skip over them. On the other hand, don't be too concerned about precision; just do what *looks* right. That will be close enough for the purpose at hand. However, if you think you need help in estimating areas under a normal curve, turn back to Figure 5-4, page 52.

Trace the curve in Figure 8-2 on a piece of paper, take a pair of scissors, and cut around it so that you have a distribution that can be moved about in order to test various hypotheses. Try the limits of 115.4–115.6, 115.2–115.8, 115.1–115.9, and 115.0–116.0; estimate in each case the probability that the obtained mean is *outside* the interval and then estimate the probability that it is inside. You can do that for any interval you wish, but in a real-life situation, you would decide in advance how much risk of error you were willing to accept and conversely what confidence level you would require. Then you would find the score limits that corresponded to that level of confidence.

In other words, you would do just what I did in Figure 8-2. I initially decided that I wanted to illustrate a confidence level of .95, which leaves a probability of .05 that the obtained mean lies *outside* the confidence interval. Half of that .05 (i.e., .025) is found above and half below that interval, so I proceeded as follows:

1. First, I slid the hypothetical distribution of means downward from 115.5 until the vertical line from the obtained mean cut off what appeared to me to be about .025 of the distribution. Then I designated the mean of that distribution as the lower limit of my confidence interval; if the population mean were any lower than that, fewer than 0.25 of repeated samples would be expected to have means as high as the one in my sample.
2. Having established the lower limit, I repeated the procedure in order to find the upper. I asked how high the distribution of sample means had to be to put only .025 of them below the obtained mean of my sample.
3. Having found that, I had both limits of the confidence interval that corresponded to a confidence level of .95.

EFFECT OF *n* ON STANDARD ERROR

We have already noted that the variability of a hypothetical distribution of means ($s_{\bar{X}}$) is estimated from the variability of an actual distribution of individual cases (S). But another factor is extremely important in making that estimate: the *number* of individual cases (n) in the sample.

The relation of n to $s_{\bar{X}}$ may be less comprehensible initially than that of s to $s_{\bar{X}}$, since the latter relation is between two forms of the same concept; however, the relation to n can be quickly illuminated by means of a few simple diagrams. Figure 8-3A represents the actual distribution of an entire population; each case in the distribution is an individual member of that population. The three drawings below it are hypothetical distributions; each case within each of them is the mean of a sample from the above population of individuals. The differences among the three distributions are striking—and those differences are a function of differences in n.

You may have been surprised to see that the first of the hypothetical distributions of means (Figure 8-3B) was exactly the same as the original distribution of

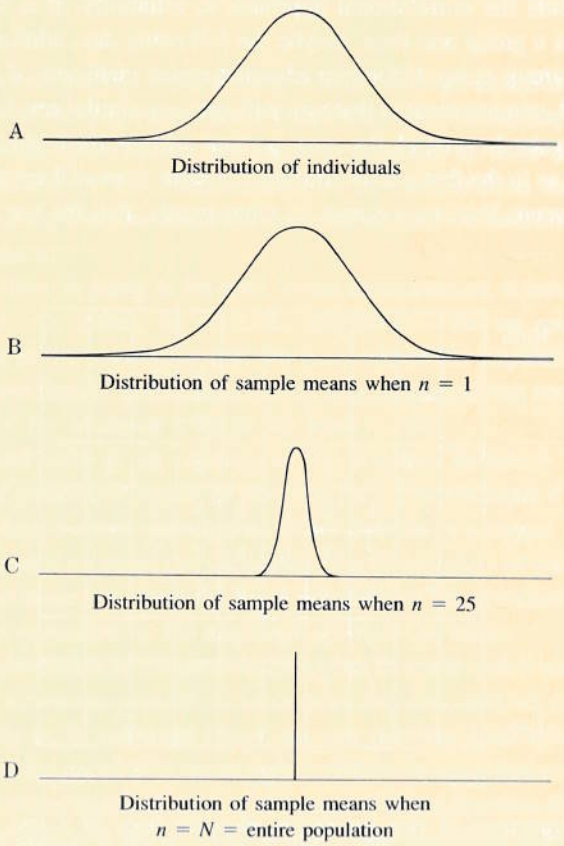


FIGURE 8-3 Distributions of individuals and of the means of samples of three different sizes.

individual subjects. But a moment's reflection will convince you that it could be no other way, because when $n = 1$, every sample *is* an individual subject. Similarly, you can see that if every sample included the entire population (Figure 8-3D), the variability among sample means would have to be precisely zero (every mean would be exactly the same as every other), and the statistic $\bar{X}$ (which is here equal to μ) would be *completely reliable*. Figure 8-3C represents an intermediate situation in which every sample has an n of 25; there you can see some variability, but not nearly as much as where the distribution is made up of samples of 1 (Figure 8-3B). In every case, the selection of samples is random, and each sample is returned to the population before the next one is taken.

TWO KINDS OF RELIABILITY

You encountered one index of reliability in Chapter 6: the coefficient r_{xx} that emerges when you correlate test X with itself. Although there are various ways of obtaining a reliability coefficient that are different from the method cited there, that one well represents the correlational approach to reliability. It is the test–retest method: You test a group and then, maybe the following day, administer the same test to the same group again. If the two administrations yield sets of scores that are highly correlated, you are assured that you will get very similar results if you do the same thing yet again: In general, the same persons will get the highest scores in the second and third as in the first administration, the same ones will get the lowest, and similarly in between. You are assured, in other words, that the test is reliable.

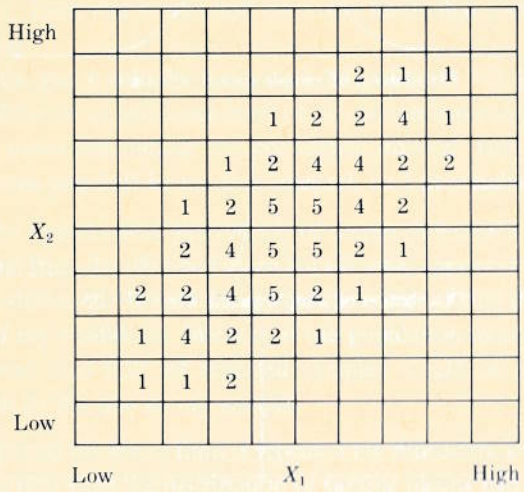


FIGURE 8-4 Scatterplot of two successive administrations of a test, using digits to represent scores that may be stacked atop one another so that only the top one can be seen. The digit in each cell represents the number of persons whose scores are in that cell. The correlation is $r_{xx} = .75$, which is rather low for a reliability coefficient.

Precision of inferences depends on the reliability of measures. In Figure 8-2, the confidence interval is small (the measurement is precise) because at a given level of confidence (in this case .95) the statistic of interest (in this case, the mean) does not vary much from sample to sample. If we had been able to include the entire population in our calculation, that statistic would not have varied at all (Figure 8-3D) and we should have had perfect reliability and perfect precision.

So you are now acquainted with two indices of reliability. One, the correlation coefficient, quantifies the tendency of many individuals to be related to each other in the same way across successive measurements; the other, the standard error, quantifies the sampling error of a single statistic. The two indices are therefore different. On the other hand, they share the essence of reliability: a similarity of results over successive trials. That similarity can be seen graphically as (1) the compactness of a scatterplot of individual scores on two successive administrations of the same test (see Figure 8-4) and (2) the compactness of a distribution of sample statistics (see Figures 8-2 and 8-3).

SUMMARY

Every measure is derived from a sample, and although samples are by definition representative of the population from which they are drawn, they differ from each other even when drawn from the same population. If there are no *systematic* (biased) deviations of the characteristics of the sample from those of the parent population, we say that the deviations are *random*, and we refer to them collectively as *error variance*.

Now, if there is going to be error in our measurement, it behooves us to know its magnitude—or rather, its *probable* magnitude—for we might otherwise be overconfident in the statements that we make about the population from which a sample has been drawn. The most common way of quantifying error is to compute a *standard error* for whatever statistic we wish to cite.

A standard error is an estimate of the standard deviation of a hypothetical distribution of values that would be obtained for a given statistic if repeated samples were drawn from a single population. If the statistic in question were a mean, for example, we could estimate the variability of a distribution of means of successive samples of the same size from the same population. The standard deviation of that distribution is estimated by the *standard error of the mean*.

The standard error of the mean is important because with it we can mark off the *limits* within which the population mean probably lies, and we can ascertain what that probability is. In technical terms, we can state the *level of confidence* of our hypothesis that the population mean resides within the specified *confidence interval*.

The size of the standard error of the mean ($s_{\bar{X}}$) is a function of the variability of the sample, as indicated above. But it is also a function of the *number of individuals* (n) in the sample; n varies all the way from *one* individual at one extreme to *all the individuals in the population* at the other. If each sample includes only one individ-

ual, the standard error is as large as the standard deviation of the population would be if we could get it. If the sample size is the same as that of the population, the standard error is zero, because repeated samples would all have the same mean. Other sample sizes have intermediate effects on the standard error.

The standard error of the mean has been used here as an illustration of the concept of *sampling error*. Everything that has been said about that here can also be said of the standard errors of other statistics.

The standard error is one kind of *reliability*. Another is the correlation of a test with itself that you first encountered in Chapter 6.

Sample Applications

Education

You are superintendent of a large urban school district. You want to know the average achievement levels of elementary school children in each grade. Funds are limited, so you test a random sample of 10 percent of the students at each grade level, using a battery of nationally standardized achievement tests. From that information you calculate the mean achievement level of the sample at each grade level. What else might you want to know besides the mean?

Political Science

You have developed an Index of Political Participation and, as a part of the standardization procedure, have applied it to each of several middle-class neighborhoods. You report the mean of the scores (50 points) to potential users of the index, but since no measure is perfectly reliable, you also want to report the limits within which the *true* mean probably lies. How do you proceed?

Psychology

You have administered an inkblot test to a 10-year-old boy. The results indicate that the boy is mildly disturbed and in need of psychotherapy. Since you know, however, that the inkblot measure is not perfectly reliable (i.e., the results may vary from one administration to the next), you wonder whether psychotherapy should be recommended. (The next administration of the inkblot test might indicate that the child is well within the normal limits.) The test manual might contain something that would help you to decide how much confidence to place in the test score. What would you look for in the manual?

Social Work

A family service agency utilizes a Family Life Involvement Profile (FLIP) to measure a family's psychological functioning. The scale yields a score that indicates whether a family's functioning is inadequate, marginal, or adequate. In your assessment of a family, you obtain a FLIP score that indicates a marginally functioning family. But you wonder how accurate the score is—how much confidence you can have in the rating. How can you quantify that uncertainty?

Sociology

A member of your research class notes that you have only a *sample* of professors in the study of authoritarianism cited in the Chapter 5 sociology application (page 55). He asserts that since you have only a sample, you do not really know where the true population mean is and that consequently your conclusions are meaningless. How do you respond?

9

Significance of a Difference between Two Means

Often, in both basic and applied research, it is important to know whether two populations are different from each other. From the point of view of the researcher, the question is better stated in the negative: *Are the two samples that I have measured merely two random samples from the same population*? The true answer to that question is either yes or no, but one can never know with absolute certainty which it is. One must therefore state one's answer in probabilistic terms. The probability that the true answer is yes (that the two samples have been drawn from a single population) depends on two factors: (1) the size and direction of the obtained difference between the two means and (2) the variability of a hypothetical distribution of differences between means when pairs of samples are taken at random from a single population. Note the effects of these two factors as you read through the chapter.

One way to illustrate the effect of variability is to analyze carefully a single example. In the following section, we shall examine an experimental study in the field of education, from the design of the experiment to the announcement of the results, concentrating throughout on basic ideas rather than computations.

AN EXAMPLE

Imagine that we have invented a new—and we hope better—method of teaching French grammar to American high school students who have no knowledge of that language. Is it really better than the prevailing method? To find out, we take two samples from the population "naive American high school students," make sure that the two are initially random samples from a single population, and teach one group (hereafter to be known as the *control group*) by the traditional method and the other (the *experimental group*) by the new method. Then, after 150 hours of teaching time, we test both groups and compare their mean scores. The objective of our study is to ascertain whether the two are *still* random samples of the same population insofar as their knowledge of French grammar is concerned.

Say the difference between the means of the two groups is 10 points. Is the obtained difference significant? Is it large enough that we may reject the hypothesis that the control and experimental groups remain, after the teaching as they had been before, random samples from a single population with respect to knowledge of French grammar? That single population might now be labeled "American high school students who have had 150 hours of instruction in French grammar." Our hope is that, instead, there are *two* populations—one superior to the other with respect to knowledge of French grammar—and that the experimental group represents the superior population.

This example is admittedly very abstract, because you have to imagine a population that does not exist—that is, a population made up of an infinite number of American high school students who have been taught French grammar by the new method. We suspect that such a hypothetical population would be superior to the one with which we are more familiar, but before we can be sure of that, we must disprove the hypothesis that the two groups are merely two samples of a single population. That is the hypothesis of no difference, or the *null hypothesis* (abbreviated H_0), and our test of significance is really a test of the null hypothesis: We attempt to *dis*prove the hypothesis that there is no real difference between the groups—that the observed difference is merely a chance difference resulting from ordinary sampling error.

That may sound easy, but there is a complication. Like virtually every conclusion that emanates from statistical reasoning, this one must be stated not in absolute terms, but in terms of probability. We will not get a yes or a no out of our statistical test. Rather, it will give us the probability that we are *rejecting a true null hypothesis*, and although that probability could conceivably approach zero, we'll never get a flat-out no.*

We therefore adopt arbitrarily some level of probability that is an *acceptable approximation* of zero, and when a test reveals a probability below that level, we

* If we do not reject the null hypothesis, we must regard it as tenable (our data offer insufficient evidence against it) but not necessarily true.

reject the null hypothesis. (After that, we may entertain other hypotheses.) Whether a particular difference—like the 10 points we obtained between our control and experimental groups after the language instruction—meets that preset criterion depends upon the *variability of a hypothetical distribution* in much the same way that the size of the confidence interval did when we were discussing the stability of the mean. It might be worth your while to review that section before proceeding with this one; it begins on page 91.

In fact, a knowledge of the standard errors of the means of both groups is essential to our present purpose. This time, however, we are dealing primarily with a hypothetical distribution not of means but of *differences between* means. In this case, each difference is between the means of two groups of American high school students taught by different methods.

TEST OF SIGNIFICANCE: THE *z* RATIO

Is a difference of 10 points sufficiently large to be significant, or is it small enough that it might easily have occurred by chance—by ordinary sampling error? The answer is a function not only of the size of the difference itself but, very importantly, of the sampling errors of the two groups being tested—as indicated by the standard errors of the means. The next six paragraphs, together with Figures 9-1 through 9-4, concern a situation in which the standard error of the mean is rather small and is the same for both groups. (Just accept my figures for the standard errors throughout this chapter. Concentrate now on what you can *do* with a standard error once you have it.) You may find this discussion difficult to follow the first time through; it would be easier just to learn the formula and its practical implications and be done with it. But the diagrams hold your best hope of really understanding how the standard error of the mean relates to the standard error of a difference between means and how the latter functions in the testing of hypotheses.

Look at Figure 9-1. It shows two hypothetical distributions of sample means on a scale of French grammar scores. The two graphs occupy the same space because the drawing was made on the hypothesis that the two groups are in reality *not* of two kinds with respect to achievement in French grammar, but are merely two sets of random samples from a single population. That is, of course, the null hypothesis (H_0), and we shall try our best to disprove it (or rather, to make it untenable).

In order to test the null hypothesis with respect to differences, it is necessary to think in terms of a scale not of scores but of *differences between* scores—specifically, of differences between control group means and experimental group means. Look again at Figure 9-1 and imagine taking pairs of means (one control and one experimental group in each pair) at random from the two distributions. (Remember that according to the null hypothesis, the two are actually one.) Most of the intrapair differences would be close to zero, but a few—just by chance, remember—would be substantially larger. In fact, if you were to compare the largest experimental group mean with the smallest mean of the controls in this example, the

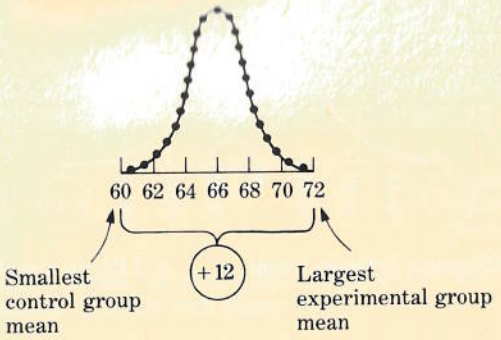


FIGURE 9-1 Sample means, $s_{\bar{X}} = 2$, when control (——) and experimental (• • •) populations are identical.

difference would be 12 raw score points, or 6 standard errors in the distribution of means (Figure 9-1).

That difference, $\bar{X}_e - \bar{X}_c$, is in the direction we have predicted, so we'll call it positive. But if all the samples are drawn from the same population, as the null hypothesis would have it, there should be as many negative differences as there are positive, and the largest of them should be just as impressive as the positive difference we just found. Figure 9-2 represents the same two distributions as Figure 9-1; if you will compare the *lowest* experimental with the *highest* control group mean, you will indeed discover that the difference is again 12, but this time in the negative direction.

Figures 9-1 and 9-2, then, represent the same pair of distributions analyzed in two different ways: one as positive, the other as negative differences. Figure 9-3 combines positive and negative differences into a distribution of differences. Positives are on the right, negatives are on the left, and the mean is zero—the null

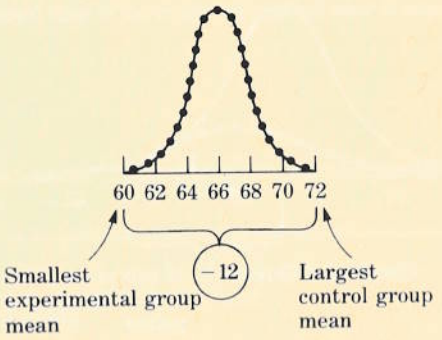


FIGURE 9-2 Same as Figure 9-1 except that different extreme means are identified (control, ——; experimental, • • •).

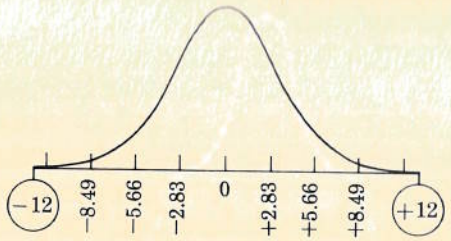


FIGURE 9-3 Differences between means, $s_{\bar{X}_e - \bar{X}_c} = 2.83$, on a scale of raw scores.

hypothesis again. If we were to compute an infinite number of differences between pairs of means, selecting each pair at random from a single population, that is about what the distribution would look like.

With that much behind us, the testing of the null hypothesis should be relatively easy. All we have to do is place our obtained difference within the hypothetical distribution of differences, as in Figure 9-4, and we can see immediately that only a very few positive differences of that magnitude (10 or larger) occur simply by chance among samples taken from a single population. (Asking only about *positive* differences constitutes a *one-tail test*. You will find a discussion of one- versus two-tail tests begins on page 112.)

By inspection it appears that very few such differences would be as large as the one we obtained in our experiment if the samples were from the same population; see Figure 9-4. We are justified, therefore, in rejecting the null hypothesis (H_0) as untenable. Formula (9-1) offers a more precise estimate of the variability of differences between means[1]

$$s_{\bar{X}_e - \bar{X}_c} = \sqrt{s^2_{\bar{X}_c} + s^2_{\bar{X}_e}} \qquad (9\text{-}1)$$

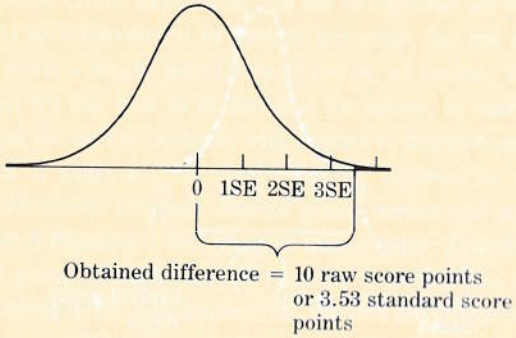


FIGURE 9-4 Same as Figure 9-3 but on a scale of standard error units ($s_{\bar{X}_e - \bar{X}_c}$). The shaded area is so small that it is hard to see.

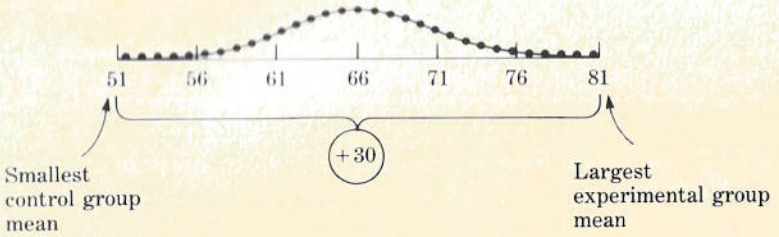


FIGURE 9-5 Sample means, $s_{\bar{X}} = 5$, when control (——) and experimental (• • •) populations are identical. Compare with Figure 9-1.

where $s_{\bar{X}_e - \bar{X}_c}$ is the standard error of the difference between means*, $s_{\bar{X}_c}$ is the standard error of the mean for the control group, and $s_{\bar{X}_e}$ is the standard error of the mean for the experimental group. Since our interest is in the underlying logic of the significance test rather than in the precision afforded by the formula, we need only note that, like Figures 9-1 through 9-4, the formula shows that the variability of random differences between means taken in pairs ($s_{\bar{X}_e - \bar{X}_c}$) is proportional to variability among means taken singly ($s_{\bar{X}_c}$ and $s_{\bar{X}_e}$).

To illustrate that point further and to emphasize the importance of $s_{\bar{X}}$ in determining the fate of the null hypothesis, turn to page 108 and see how our 10-point obtained difference would have fared had the sample means been less stable, as they are in Figures 9-5 through 9-8. In Figures 9-5 and 9-6, the standard error of the mean is 5 (instead of the $s_{\bar{X}} = 2$ of Figures 9-1 and 9-2), and in Figures 9-7 and 9-8, the standard error of the difference is 7.07 (instead of the 2.83 of Figures 9-3 and 9-4). Whereas in Figure 9-4 our difference was 3.53 standard errors ($z = 3.53$) above the mean and had a probability of .0002, in Figure 9-8 a difference of 10 is only 1.41 standard errors away from the mean and has a probability of .0793.

* Again, as with the standard error of the mean, this standard error is necessarily "estimated."

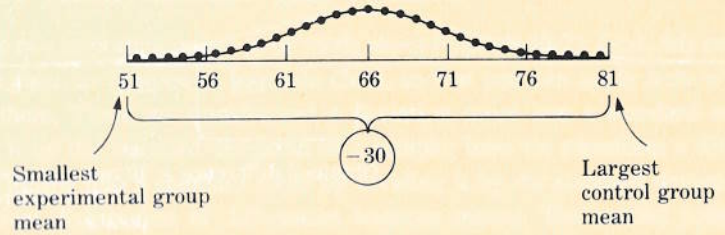


FIGURE 9-6 Same as Figure 9-5 except that different extreme means are identified (control, ——; experimental, • • •). Compare with Figure 9-2.

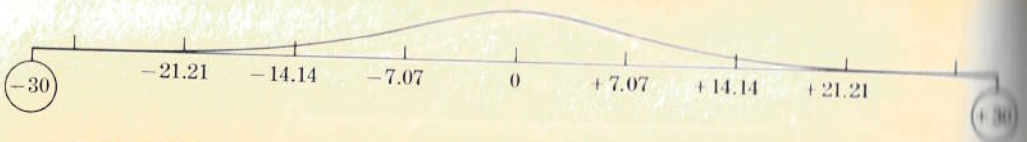


FIGURE 9-7 Differences between means, $s_{\overline{X}_e - \overline{X}_c} = 7.07$, on a scale of raw scores. Compare with Figure 9-3.

In the latter case, if we were to take 1000 pairs of samples at random from a single population, 79 of the intrapair differences would be at least as large as ours and in the same direction. Had we obtained our difference in *those* circumstances, we would have had to accept the null hypothesis as tenable. Incidentally, do not be disturbed if you cannot estimate the above values with this kind of precision simply by looking at the graphs; I did it by formula.[2]

In summary, Figures 9-1 through 9-4 show tests of the significance of differences when $s_{\overline{X}_c}$ and $s_{\overline{X}_e}$ are small, and Figures 9-5 through 9-8 when they are large. Distributions are all hypothetical. All of the examples in this section have been graphed in order to expose the structure of their operations. Normally, however, those operations are done algebraically for greater accuracy. In the language-teaching study, the formula would be

$$z = \frac{(\overline{X}_e - \overline{X}_c) - 0}{s_{\overline{X}_e - \overline{X}_c}} \qquad (9\text{-}2)$$

where z is the ratio of an *obtained difference* $(\overline{X}_e - \overline{X}_c)$ to the *standard error of the difference* between means $s_{\overline{X}_e - \overline{X}_c}$ in a distribution of such differences when the mean of the distribution is zero. In Figure 9-4, that ratio is 10/2.83, which makes it a much higher ratio than the 10/7.07 of Figure 9-8. The zero in the formula has no effect on the outcome; it is there only to remind you of precisely what the formula is supposed to do: namely, to ascertain how far our obtained difference is from the mean of a distribution of differences, all of which are the results of sampling errors. The mean of such a distribution would indeed be zero.[3]

Box 9-1 shows how a z could be calculated from the defining formula (9-2).

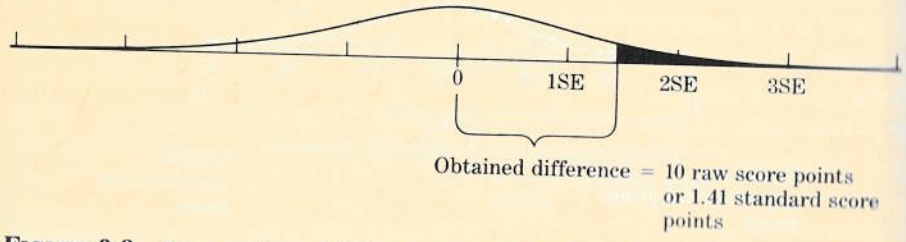


FIGURE 9-8 Same as Figure 9-7 but on a scale of standard error units ($s_{\overline{X}_e - \overline{X}_c}$). Compare with Figure 9-4; here the shaded area is much larger.

BOX 9-1 Calculation of a Test of Significance of a Difference between Two Means [See Equation (9-2) and Figures 9-1, 9-2, 9-3, and 9-4]

(1) $\overline{X}_e - \overline{X}_c = 10$

(2) $s_{\overline{X}_c} = 2 \qquad s_{\overline{X}_e} = 2$

(3) $s_{\overline{X}_e - \overline{X}_c} = \sqrt{s^2_{\overline{X}_c} + s^2_{\overline{X}_e}} = \sqrt{8} = 2.83$

$$\boxed{z = \frac{(\overline{X}_e - \overline{X}_c) - 0}{s_{\overline{X}_e - \overline{X}_c}}} = \frac{10}{2.83} = 3.53$$

If there are 50 students in the control group and 50 in the experimental, $p \leq .01$.

Before you can calculate the ratio z of a difference between means to the standard error of a difference between means, you must:

measure all of the individuals in both samples,

calculate their means and the difference between their means (row 1),

calculate the standard error of the mean for each sample (row 2), and

calculate the standard of error of the difference between means (row 3).

The ratio of row 1 to row 3 is:

$$\frac{\text{row 1}}{\text{row 3}} = \frac{\text{difference between means}}{\text{standard error of difference between means}} = z$$

TEST OF SIGNIFICANCE: THE *t* RATIO

There are two important points to be made about Formula (9-2): First, this ratio is of the same form as the one in Formula (5-1), so that you can apply here what you learned there. There we transformed a deviation score into a standard score by dividing it by the standard deviation of a sample; here we transform a difference between means into a standard score by dividing it by the standard deviation of a hypothetical distribution of differences between means. The second point to be made about Formula (9-2) is that although the general form of this new ratio is the same as that of Formula (5-1), its content is somewhat different: Specifically, the

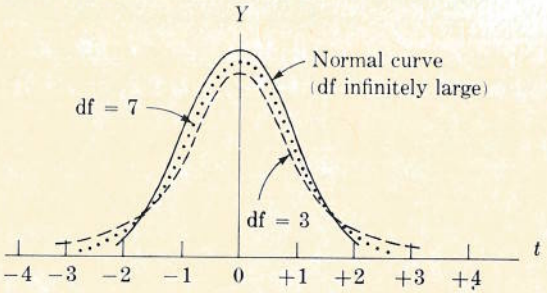


FIGURE 9-9 Distribution of t ratios at three different sample sizes. [Adapted from Henry L. Alder and Edward B. Roessler, *Introduction to Probability and Statistics*, 6th ed., W. H. Freeman and Company. Copyright © 1977.]

distribution here is hypothetical, not actual, and the z ratio can be used only when the distribution is actual—that is, when we can measure every member of the population. Because measuring every person in an entire population is not always feasible, we need a ratio that will work when we can measure only a *sample* of the population.

We may regard Equation (9-2) as the basic formula for the test significance because it clearly reveals the relation between the difference and the standard error of the difference. Unless you have measured the entire population, however, it is not the z ratio that we will test for significance, but a ratio called t. The t ratio is essentially the same as z, but it is referred for significance testing to a distribution that changes its shape (because the probability of obtaining the various scores in it changes) as n gets smaller. When n is very large, the difference between z and t distributions is negligible; when n is as large as the population, the two are identical.

Because t was developed specifically for use with small samples, it is important to use degrees of freedom (df) in place of n when discussing it (see pages 87–88). In Figure 9-9, it is clear that when the t distribution is used, the probabilities to be inferred from various placements on the baseline are in many instances quite different if degrees of freedom is small rather than large. Most notably, when degrees of freedom is small, extremely large t ratios (either positive or negative) make up a larger-than-normal part of the distribution.

Because it is appropriate for use with either large or small samples, the t test is used almost universally in place of z whenever inferences must be made from accessible samples to inaccessible populations. But because z is a construct that you already understand, it is best to think of t as a modified z.

SIGNIFICANCE LEVELS

We have seen in Figures 9-5 through 9-8 that in certain circumstances a difference of 10 points between two sample means can occur frequently by chance. However,

through 9-8, but in those of Figures 9-1 through 9-4; there it is clear that the probability is extremely small that, with respect to their knowledge of French grammar, our control and experimental groups are random samples of a single population. Now we may announce to an eagerly waiting public that our new method of teaching French grammar is almost certainly better than the traditional one—at least under the circumstances prevailing in our experiment. We proclaim that superiority at a *significance level* of .0002, because there is only that much of a probability that our proclamation is wrong—that we are rejecting a true null hypothesis.

It may seem to you that the level of significance in this instance should be 1.000 − .0002, or .9998. But the level must be expressed as *the probability that a true null hypothesis is being rejected*. That means that the *lower* the significance level, the *higher* is our confidence that the effect we have observed is real—that it is *reliable*.[4]

Tradition also provides us with two levels recognized as significant and very significant, respectively. A *significant* difference is one that would occur only *five times* (or fewer) in 100 comparisons if every sample were taken at random from the same population; a *very significant* difference is one that would occur only *once* in 100 comparisons. These are sometimes referred to as the .05 and .01 levels of significance, or as $p \leq .05$ and $p \leq .01$, respectively. (The p stands for "probability.") In tables and other places where the briefest possible abbreviations are used, a significant difference sometimes is designated simply by "S" and a very significant one by "VS." In practice, those two levels are often used but also often ignored. For example, a researcher who obtained as impressive a difference as that between our control and experimental groups in the language-testing study ($p \leq .0002$) would be unlikely to hide his or her light under the bushel "less than .01"!

A COMMON MISINTERPRETATION

Given the information that $p \leq .01$, you may be tempted to say that because there is at most a .01 probability that a difference as large as the obtained one occurred by *chance*, there is a probability of at least .99 that the obtained difference is *real*. But the significance test is concerned only with the null hypothesis, and the null hypothesis asserts that the real difference is *zero*. A p of .01 tells us only that if the two measured groups are random samples from a single population—that is, if the null hypothesis is true—then either

1. the probability of getting a difference as large as this (the two-tail test) is no greater than .01, or
2. the probability of getting a difference as large as this *in the expected direction* (the one-tail test) is no greater than .01.

ONE- VERSUS TWO-TAIL TEST

I have just used the terms *one-tail test* and *two-tail test*. The term "tail" refers to the upper or lower end of a normal frequency distribution where the curve is close to the baseline. Figure 9-4 is a diagram of a one-tail test. The shaded portion *in the right tail* is very small: specifically, it represents a .0002 probability that a difference as large as 10 points *and in the expected direction* would occur by chance.

The alternative is a two-tail test, which is appropriate whenever *we have not predicted the direction* of the difference. If you will look back at Figure 9-3 for a moment, you will see that the probability of a difference of 10 in *either* direction is double that of a difference in *one* direction. That means that a particular difference can fail the significance test if there has been no directional prediction and pass it if a prediction has been made (provided, of course, that the outcome matches the prediction). We are not free to select either test arbitrarily; if we have no reason to expect a difference in one direction rather than another, we are obliged to use the two-tail test. Only when we do have such a reason, *when we can make a prediction*, and when the results confirm the prediction are we justified in using the one-tail test.[5]

STATISTICAL VERSUS PRACTICAL SIGNIFICANCE

The length of this section barely exceeds that of a long footnote, but it contains an important notion that might easily be overlooked if it were not emphasized. The point is that even though a difference is shown to be statistically significant, it may not have any *practical* significance.

The result of our language-teaching experiment was very convincing; there can be almost no doubt that the difference between the two groups at the end of the teaching period was real. But how does that information affect the decisions of a school administrator who must decide whether to adopt the new method? That depends on many considerations, some of which have nothing to do with statistics. It depends to a great extent on how costly the new teaching method is to implement —whether it requires expensive equipment or specially trained teachers. The administrator's decision may also depend on the absolute size of the difference. That may seem a contradiction of everything we've been saying up until now, and it is—unless we keep in mind the distinction between statistical and practical significance.

The joker is in that standard error term—more specifically, the importance of n in determining the size of that term. If you analyze either Figures 9-1 to 9-8 or Formula (9-1), you will find that the size of the standard error of the difference is proportional to the standard errors of the means of the two samples. The size of a standard error of the mean [Equation (8-1)] is determined partly by the *variability* in the sample and partly by its *size*:

$$s_{\overline{X}} = \frac{s}{\sqrt{n}}$$

Now, the significance of an obtained difference depends upon the ratio between that difference and the standard error of the difference:

$$z_{\overline{X}_e - \overline{X}_c} = \frac{\overline{X}_e - \overline{X}_c}{s_{\overline{X}_e - \overline{X}_c}} \tag{9-3}$$

where $z_{\overline{X}_e - \overline{X}_c}$ is the difference between two means in standard error units, $\overline{X}_e - \overline{X}_c$ is the difference between two means in raw score units, and $s_{\overline{X}_e - \overline{X}_c}$ is the standard error of the difference. The presence of n in the denominator of the $s_{\overline{X}}$ formula means that a large n yields a small $s_{\overline{X}}$. That, we know, reduces the standard error of the difference, and because z increases as the standard error decreases, the ultimate result of a large n is high significance. That is why if our samples are large enough, *any* difference will be statistically significant, no matter how small it may be. Thus, although a z ratio of 3.53 indicates that the difference, no matter how small, is probably genuine, a school administrator might decide that a very *small* difference would not be worth its additional cost, even if it were *absolutely* reliable—that is, even if $s_{\overline{X}_e - \overline{X}_c}$ were zero.

In short, the size of a difference is in no way affected by its reliability. The temptation to infer that it is should be firmly suppressed.

SUMMARY

Sometimes it is important to know whether two groups are different from each other—or rather, *whether they represent populations that are different from each other*. In this chapter, we have compared the performance of two groups of students exposed to two different teaching methods. One group, taught by the traditional method, was called the *control group*; the other, taught by a new method, was the *experimental group*. The experimental group outscored the control.

But two groups given the *same* treatment could have scored differently by chance, and in such a case we would have to admit that the obtained difference was a *sampling error*—that the two groups are really only two samples from the same population with respect to the performance being tested. We must consider the possibility that our control and experimental groups are like that, too—that the difference we obtained was due entirely to sampling error and that the two groups are really only two samples from a single population. That possibility is known as the hypothesis of no difference, or *null hypothesis*.

The general question, then, is really "How large does an obtained difference have to be before we are justified in rejecting the null hypothesis?" The procedure by which that question is answered is known as a *test of significance*. In the case at hand, we ask a more specific form of the question, namely, "Is our obtained difference large enough to justify a rejection of the null hypothesis?"

The answer will not be a simple yes or no; it must be given in terms of *probability*. What is the probability that a difference as large as ours would occur if there were no real difference in the effectiveness of the two teaching methods?

Acceptable levels of probability are somewhat arbitrary, but two such *levels of confidence* have traditionally been set at .05 and .01. However, a researcher who obtains a null hypothesis probability much lower than .01 may report it exactly, because the lower it is, the more impressive is the outcome of the study.

Is it possible to ask all of the above questions about any measurement of any two groups of subjects? We could use the test of significance simply to explore—to search for differences worthy of further investigation. But in the case of the two teaching methods described above, we were not exploring; we *expected* the experimental group to be better than the control. That is an important distinction, because if we can state our expectancy in advance, we are privileged to use a *one-tail test* instead of a *two-tail test* of significance. If we are exploring, we must ask, "What is the probability of a sampling error this large *in either direction*?", whereas if we have stated our expectancy, we may ask instead, "What is the probability of obtaining this large an error *in favor of the experimental group*?" Since the probability of a difference in *one specified* direction is just half that of the same amount of difference in *either* direction, the one-tail test is twice as sensitive as the one we would be obliged to use if we were merely exploring; a difference half as large will qualify at whatever level of significance we have prescribed. Finally, a test of statistical significance tells us how *reliable* our difference is, but that is only one factor in making practical decisions.

Sample Applications

EDUCATION

You are principal of a high school. The teachers, counselors, and administrators of the school have developed a one-semester group-counseling program for students who are disrupting class to help those students learn more appropriate ways of resolving conflicts and participating in classroom activities. To find out whether the program helps reduce student disruptiveness, you randomly assign half of the 50 most disruptive students to the group-counseling program. At the end of the semester, the teachers rate the disruptiveness of all 50 students. When the scores have been collected, what do you do with them?

POLITICAL SCIENCE

Increasingly in recent years, political scientists have used statistical methods to evaluate the effectiveness of public programs. An example of this kind of research can be seen in the following case involving a crime control program. The citizens of Gritty City have demanded that local officials do something to resolve the problem of burglaries. You are chief of police. As a first step, you propose that the city council establish a special task force that will give presentations in each neighborhood on how to prevent burglary. Since the council is hesitant to fund the task force on a permanent basis without any evidence of its effectiveness in reducing burglaries, its members agree to make a decision after analyzing data from a pilot program. To set up such a program, you draw two random samples from the population comprised of all of the city's precincts. The residents in one group then receive the task force presentations; residents in the other group do not. After three months, you compare the mean number of burglaries in the precincts receiving the presentations with that in the cops-only precincts. What is an appropriate statistic for making such a comparison?

PSYCHOLOGY

As a clinical researcher, you are interested in ascertaining how a period of training in muscle relaxation will compare with the use of stimulant drugs in reducing hyperactivity in young children. You assign half of the children medically classified as hyperactive to the relaxation program and half to the drug program. Following a 30-day period of intervention, the activity level of all children in the study will be assessed by means of a rating scale. There will almost certainly be some difference between the two groups. How can you tell whether the difference is significant?

SOCIAL WORK

As a social worker in a senior citizens' center, you are concerned about the health and vitality of the seniors who frequent the center. It is your assessment that the center's current program of bingo, pool, backgammon, quilting, films, and occasional field trips is not sufficient to keep seniors active and alert, for they experience numerous health problems, including strokes, heart attacks, upper respiratory illnesses, and emotional illnesses involving depression and anxiety. After attending a workshop on services for the elderly, you plan to implement a new program that includes group discussion, meditation, and physical exercises. It is a structured program; seniors meet for two hours twice weekly. To ascertain the effect of this new program, you randomly select half of the seniors and engage them in the program for a year. You then plan to compare this group to the half of the membership that has continued in the regular center activities. After collecting health data on all of your subjects, you find a difference in favor of the experimental group. How can you estimate the probability that this difference arose by chance?

SOCIOLOGY

A family planning agency has asked you whether Catholic families are larger than non-Catholic families in your state. You draw a random sample from census data and find that the mean size of Catholic families is indeed larger than that of non-Catholic families. How might you test the significance of this difference?

10

More on the Testing of Hypotheses

A chapter with this title could easily occupy as much space as all the rest of this book put together. It will not, because I have selected just two tests of significance as illustrations. The others will be mentioned only briefly, if at all. For our purposes, it is not necessary to describe those others. In fact, probably the most important contribution that I can make to your understanding of all of them can be stated without *any* illustrations. The fundamental idea is this: Each is conceived as the *testing of a null hypothesis*—a hypothesis of no difference.

The difference that emerged from our experiment with two methods of teaching French grammar was a difference in *amount*. We asked whether an innovative method resulted in students who were more able than students taught by the more traditional method with which it was being compared. More technically, we asked whether the difference we obtained was attributable to the difference in teaching methods or merely to sampling errors—that is, we tested the null hypothesis. But we could have asked a different question: Given some clear criterion of passing, does the innovative method produce fewer failures than the traditional? That question is posed in terms of *frequencies*. Indeed, frequencies are often used to indicate magnitudes, as when number of words correct indicates amount of typing skill or number of strokes (or rather its inverse) indicates amount of golfing skill. The significance tests described in Chapter 9 can be used in such cases, and they can be modified to deal with the categorical cases just mentioned (pass–fail, yes–no,

innovative–traditional, etc.). More often, however, such cases are analyzed in a different way. The first section of this chapter will examine a test of the null hypothesis in which all of the data are in the form of frequencies.

There is a feature of the example used in Chapter 9 that is not characteristic of all experiments: Only two "treatments" were compared. What should we do if we wanted to assess the relative effectiveness of several teaching methods? Or what if we suspected that one treatment method might be effective in the hands of one teacher and another might be superior when used by a different teacher? The remaining sections of the chapter will describe a kind of analysis by which both of those questions can be answered.

Notice that every test of significance is a test of a null hypothesis; conversely, any test of the null hypothesis is a test of significance. Since any observed difference can result from sampling errors, any difference can be put to the test of significance.

COMPARISON OF FREQUENCIES: CHI-SQUARE

In the introduction of this chapter, I mentioned that the scale of scores used earlier in our investigation of teaching methods could have been reduced to only two scores: passing and failing. I pointed out that under those circumstances, the data would probably be in the form of frequencies indicating how many subjects passed under each of the two conditions.

In practice, however, scales of test scores are seldom reduced to two class intervals; to do so would be to throw away information. Significance tests designed to deal with frequencies, proportions, and probabilities are usually applied in situations that do not produce finely divided scales in the first place. In a public opinion poll, for example, respondents are usually asked a question to which they are expected to answer either yes or no, for or against, and so forth. There are ways of obtaining more finely graded responses, but they are more cumbersome to administer than the single question. Furthermore, a relatively crude index is often the most appropriate to the circumstance. For example, in predicting the outcome of an election, every mark that a voter makes on his ballot represents what is essentially a yes–no decision.

Let us consider briefly the reasoning behind a significance test of a set of data from an election. Imagine that a particular male candidate for public office has considerably more sex appeal than another but seems to differ very little from his only opponent (another male) on any other dimension. He wins, and we wonder whether his sex appeal played a part in his victory.

If we assume that the voters of both sexes have a basically heterosexual orientation, we may get an informative answer by rephrasing the question: "Did women vote for Mr. Sex in significantly larger numbers than did men?" The data can be arranged in a 2×2 table (Table 10-1). If there are in this election 20,000 voters—10,000 male and 10,000 female—of which we have a 1 percent representative sample, if Mr. Sex has received 60 percent of all the votes cast, and if there is no difference between men and women with respect to voting behavior in this election, the table will look like Table 10-2.

TABLE 10-1 Arrangement of data in election problem

		Votes for Mr. Sex: Yes	Votes for Mr. Sex: No
Sex of voters	M		
	F		

You should recognize that last "if" as the null hypothesis; our test of significance will be an attempt to show that it is untenable. To do so, we shall have to demonstrate (1) that our *obtained* frequencies differ from those that would be *expected* if there were no real difference between the voting behaviors of men and women and (2) that the difference is larger than we should be likely to get through sampling errors.

The division of votes in our sample turns out to be as in Table 10-3. In Table 10-4 we'll put expected and obtained frequencies together, the better to compare them. It is clear that our candidate received a higher proportion of the female vote than of the male. But is the difference *significant*? How likely is it that a difference (actually a set of four differences in Table 10-4) as large as that obtained would occur in a random sample if there were no difference at all within the population of voters? (The null hypothesis here is that the two variables—"sex of voters" and "votes for Mr. Sex"—are *independent* of each other.)

To find out, we may use a statistic called *chi-square* (χ^2). The general formula is

$$\chi^2 = \Sigma \frac{(f_o - f_e)^2}{f_e} \tag{10-1}$$

TABLE 10-2 Expected frequencies under null hypothesis

		Votes for Mr. Sex: Yes	No	Total
Sex of voters	M	60	40	100
	F	60	40	100
	Total	120	80	200

TABLE 10-3 Obtained frequencies from actual sample

		Votes for Mr. Sex		
		Yes	No	Total
Sex of voters	M	50	50	100
	F	70	30	100
	Total	120	80	200

where χ^2 is chi square, f_o is the obtained frequency, and f_e is the expected frequency.[1] Focus on one part of that formula for a moment

$$f_o - f_e$$

and you will grasp the fundamental nature of chi-square. In each cell of the table, the greater the deviation from the expected frequency, the larger chi-square is likely to be. Notice in Formula (10-1) that squaring each difference allows negative differences to augment rather than reduce the total.

Since the expected frequency is derived from the hypothesis of no difference, chi-square is an index of deviation from that hypothesis. Tables are available in which one can find the significance level of any given chi-square. In the present example, chi-square is 8.34, which is significant at the .01 level; the probability that our obtained difference was due to sampling error is less than 1 percent.

Calculation of chi-square from the defining equation (10-1) is demonstrated for the Mr. Sex election problem in Box 10-1.

TABLE 10-4 Obtained minus expected frequencies

		Votes for Mr. Sex	
		Yes	No
Sex of voters	M	50 −60 −10	50 −40 +10
	F	70 −60 +10	30 −40 −10

BOX 10-1 Calculation of a Chi-Square [see Table 10-4 and Formula (10-1)]

$$\chi^2 = \Sigma \frac{(f_o - f_e)^2}{f_e} \tag{1}$$

$$= \frac{(50 - 60)^2}{60} + \frac{(50 - 40)^2}{40} + \frac{(70 - 60)^2}{60} + \frac{(30 - 40)^2}{40} \tag{2}$$

$$= \frac{(-10)^2}{60} + \frac{10^2}{40} + \frac{10^2}{60} + \frac{(-10)^2}{40} \tag{3}$$

$$= \frac{100}{60} + \frac{100}{40} + \frac{100}{60} + \frac{100}{40} \tag{4}$$

$$= 1.67 + 2.5 + 1.67 + 2.5 \tag{5}$$

$$= 8.34 \ (p \leq .01) \tag{6}$$

Row 1: This is Formula (10-1).

Row 2: Numerator of each of the four ratios in this row shows the difference between the observed frequency and the expected. (See the four cells of Table 10-4.)

Row 3: Difference from row 2 is squared and divided by the expected frequency, as specified in Formula (10-1).

Row 4: Same as row 3, but notice that there are no longer any negative numbers in the equation. That is appropriate, because *any* deviation of an observed frequency (f_o) from the null hypothesis prediction (f_e) *increases* our confidence that the "votes for Mr. Sex" is related to "sex of voters." The null hypothesis is that the two are *not* related.

Row 5: Each ratio is in decimal form.

Row 6: The sum of the four ratios is the chi-square.

MULTIMEAN COMPARISONS: ANALYSIS OF VARIANCE

In an experiment, an *independent variable*—for example, a teaching method—is a variable that is manipulated by the experimenter. (Mathematically, it is independent; in an experiment, it is really a *manipulated* variable. But the mathematical term is frequently used whether the context is mathematical or experimental.) For that reason, it is sometimes called a *treatment* variable. The *dependent variable*—for example, student response to teaching—is a variable whose values are deter-

mined by those of the independent variable(s). (A more extended discussion of the variables in an experiment begins on page 141.)

In our evaluation of teaching methods in Chapter 9, we applied a significance test to a difference between two means. That technique is frequently needed in behavioral research, but it is certainly not the only possibility. An experimental design may call for more than one independent variable—for example, several teaching methods. It may also assess the effects of those variables in various combinations—for example, combinations of methods and teachers. Tests based on the z and t ratios described in Chapter 9 cannot be used to assess significance in either of those situations. However, one technique is appropriate to both of them. It is called *analysis of variance* (abbreviated ANOVA).

One-Way Analysis of Variance

In this section, we shall be concerned only with the design mentioned above in which there is one treatment variable but more than two categories of that variable. Our example will be a simple extension of the experiment we used to illustrate the application of the z (or t) ratio. (See pages 104ff and 109ff.) There we had two groups, which we called control and experimental; here, we shall have six groups, to which we shall refer simply as I, II, III, IV, V, and VI. There we were comparing a new method of teaching French grammar with a traditional method; here we have six different methods, one or more of which may be traditional. Let us say that the control and experimental groups of the earlier experiment are groups I and IV of this one.

Eventually, we shall want to know which of the six methods is (are) the most effective. We could proceed by comparing all possible combinations of groups, but that would be tedious since it would require, in this problem, 15 such tests to check all the possibilities. What we need is some kind of survey test that will tell us whether there is any significant difference *anywhere* in an array of categories. If it tells us no, there will be no point in searching further.

There are other reasons for using an overall test of significance that are more important in the long run than the saving of labor. First, any statistic based on *all* the evidence will be more stable (see "Effect of n on Standard Error," pages 99ff) than one based on only part of it, as would be the case if only two of the six methods were compared. Second, there are so many comparisons that some will be significant by chance. If there were a hundred such comparisons, five probably would show significance at the .05 level and one at the .01 level, even if there were no real differences at all. So whenever we are dealing with several categories of the treatment variable, we need an overall test of significance.

Such a test does exist. It is called the *F test* or *F ratio*. F is a ratio of two variances. In Chapter 4, variance was defined as the square of the standard deviation:

$$s^2 = \frac{\Sigma x^2}{n}$$

Then in Chapter 7 (pages 86–89) you learned that

1. the sample statistic s (and hence s^2) is only a means to an end,
2. the end to be approached is the population parameter σ, and
3. when only a sample is available, the parameter can be approximated by a statistic s, the standard deviation of the population as estimated from sample data (or more simply, the "estimated standard deviation of the population.")

Now, since the variance of any distribution is the square of its standard deviation, our best estimate of the variance of a population is

$$\text{estimated population variance} = s^2 = \frac{\Sigma x^2_{\text{sample}}}{n - 1} \qquad (10\text{-}2)$$

where s^2 is the square of the estimated standard deviation of the population, $\Sigma x^2_{\text{sample}}$ is the sum of the squared deviation scores in a sample, and $n - 1$ is degrees of freedom in this calculation.

The F test is a ratio between two variance estimates. But before we analyze the logic of the F test, let's take just a moment to consider in very general terms what it is intended to accomplish. We have six groups. The mean of each group differs by some amount from that of every other group. The question is "Are those *significant* differences?" (or more precisely, "Is there at least one significant difference among them?"). We want to know whether the observed variability of the means is greater than could be expected by chance. Look at Figures 10-1 and 10-2. Which of these diagrams represents the more reliable (significant) differences among means?

Of course the smaller the variability of individual scores within each group, the more confident we can be that we are really dealing with different groups—or, more precisely, with samples drawn from different populations. It is relatively difficult to imagine that the six groups in Figure 10-1 are random samples taken *from a single population*. That is the null hypothesis in analysis of variance—that all of the groups being compared are samples taken from the same population. The

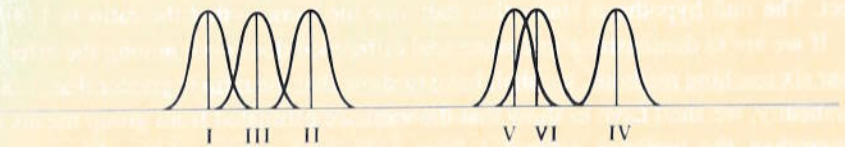


FIGURE 10-1 Groups I through VI with small variability within groups.

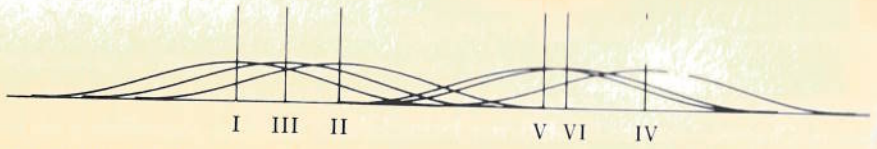


FIGURE 10-2 The same means as in Figure 10-1 but with more variability within groups.

null hypothesis is intuitively more credible in Figure 10-2 than it is in Figure 10-1. But we must not rely on intuition; we need a quantitative test of the null hypothesis. For analysis of variance, that test is in a *ratio* known as F.

One important application of the F ratio is to an overall test of significance. As always in a significance test, the null hypothesis states that all samples are random samples from a single population; as always, we shall attempt to disprove that statement. Again, as in the z and t tests, it is a *ratio* that is being evaluated. But this time the critical ratio is not of a difference to its standard error; this time it is a ratio of two estimates of the population variance—one estimate from differences among the means of the categories being studied, the other from differences among individual scores *within* categories:

$$F = \frac{s_b^2}{s_w^2} \tag{10-3}$$

where F is a universally recognized symbol for the ratio at the right of the equation sign, s_b^2 is the population variance estimated from observed variability among the means of the groups, and s_w^2 is the population variance estimated from observed variability *within* the groups.[2]

You will remember that back in Chapter 8 (especially pages 91–93) we used the variability of individual scores to estimate the variability of means. Well, it can work the other way, too; we can use the variability of obtained means to estimate that of the individual scores in a population. The numerator in the F ratio is just such an estimate: the population variance estimated from the variability of *means* of the groups being studied. The denominator is the population variance estimated from the variability of *individual scores* within those groups. So we have in this ratio two estimates of population variance. You could say that the numerator is an estimate of population variance with the effect of the treatment variable included, while the denominator excludes that effect. If the two estimates are the same, there *is* no effect. The null hypothesis states that they *are* the same—that the ratio is 1.00.

If we are to demonstrate that some real difference does exist among the effects of our six teaching methods, we shall have to show that the ratio is greater than 1.00. Specifically, we shall have to show that the variance estimated from group means is greater than the variance estimated from individual scores within the groups. Understanding why that is so requires an analysis of the logic behind the F test.

Every estimate we make of population variance is an estimate of the variability of individual scores around a true mean. The closest thing we have to a true mean here is the *grand mean*. Every individual deviation from the grand mean can be thought of as two components: (1) the deviation of the individual score from its group mean and (2) the deviation of that group mean from the grand mean. Now, if we could somehow eliminate the means component, we should have an estimate of *what the population variance would have been* if there had been *no differences among the group means*. If all of the group means are alike, then we can measure the deviation of each individual score from its own group mean and get the same result as if we had measured the deviation of each individual score from the grand mean. From such measures we can estimate population variance. The result is the denominator of the F ratio.

If you will compare Figures 10-2 and 10-3, you will see immediately that the total variability of the six groups is smaller when the variability of their means is eliminated. (You see only *one* mean in Figure 10-3, because all six are in the same place.) That is why the denominator of the F ratio is smaller than the numerator if there are any differences at all among the group means.

Those group means do vary almost always; the question is whether they vary enough to be significant in the sense that was developed in Chapter 9. To find out, we compute a ratio. To provide the *denominator* of that ratio, we discover how much *individuals* deviate from their own *group* means; whereas the *numerator* also takes into account the deviations of those *group* means from the *grand* mean.

The null hypothesis is that there is no variance among the means: Hence it predicts that an estimate of the population variance that *includes* that variance (the numerator of the F ratio) will be the same as one that *excludes* it (the denominator); if there are no differences among the means, the population variance estimate that takes those differences into account will be no larger than the one that does not, and the F ratio will be 1.00. Ratios larger than 1.00 can be evaluated by tables that give the probability that any given F occurs entirely by chance. That, of course, is the

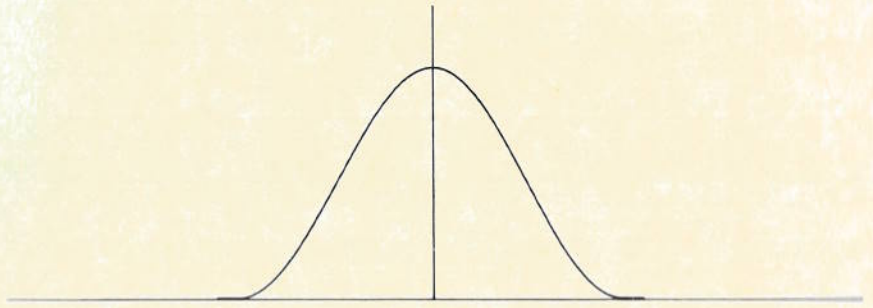


FIGURE 10-3 The six distributions in Figure 10-2 superimposed so that differences between means are eliminated.

significance level of the ratio, and it is interpreted in the same way as one derived from a z or a t. We reject the null hypothesis if the probability of its occurrence by chance is sufficiently low; usually $p \leq .05$ will suffice.

Box 10-2 was designed to illustrate the calculation, from Formula (10-4), of an F ratio for the six groups depicted in Figure 10-1. The illustration is not very realistic, however, because in order to make it as easy as possible for you to follow all the manipulations, it was necessary to make the n of each group very small. (A classroom group is much more likely to include 30 students than the 3 that make up each group in Box 10-2.) If you could see the calculations for six groups of 30, however, you would be grateful that these groups are so small.

One feature of Box 10-2 that may puzzle you is that the formula for F, instead of the familiar

$$F = \frac{s_b^2}{s_w^2}$$

Box 10-2 Calculation of one-way analysis of variance (see Figure 10-1)

Data table

	Group I	III	II	V	VI	IV
Individual within groups	9	10	12	15	17	17
	11	12	13	16	15	19
	10	11	11	14	16	18
ΣX	30	33	36	45	48	54
$\bar{X}$	10	11	12	15	16	18

Grand $\bar{X} = 13.7$

Source table

Source	df	SS	MS	F
Between	5	148	29.6	29.6
Within	12	12	1	
Total	17	160		

$$F = \frac{MS_b}{MS_w}$$

$$= \frac{29.6}{1}$$

$$= 29.6 \qquad (p \leq .01)$$

The data table organizes the data of the experiment in the same way that the subjects were organized into six groups. Notice that:

1. individuals vary around their group mean (*within-group variance*), and
2. group means vary around the grand mean (*between-groups variance*).

The source table is the analysis of variance table that you are most likely to see in a research report. Two of the terms in this table, "SS" and "MS," are new. The concepts to which they refer are familiar, however:

SS = Sum of squares, i.e., sum of the squared deviations from the mean = Σx^2

MS = mean square, i.e., mean of the squared deviations from the mean = $\Sigma x^2/df$, which is the *variance* (s^2); see the text discussion on pages 128ff

The variance term (MS) in the "between" row of the table is of group means around the grand mean; the variance term (MS) in the "within" row is of individuals around their group means. (The "within" row is often labeled "error.") The F ratio is thus

$$\frac{s_b^2}{s_w^2} = \frac{MS_b}{MS_w}$$

The term "df" refers to *degrees of freedom*, which is the denominator of the formula for the estimated population variance, whether the estimate is based on between- or within-group deviations:

Between df = number of group means minus the 1 degree of freedom lost by computing the grand mean: $6 - 1 = 5$

Within df = number of individuals in one group minus the degree of freedom lost by computing that group mean multiplied by the number of groups: $6(3 - 1) = 12$

Don't forget that a significant F reveals only that there is at least one difference between groups somewhere in the table of data. (See the discussion in text on pages 122ff.)

appears there as

$$F = \frac{\text{MS}_b}{\text{MS}_w}$$

Be assured that the two expressions are really equivalent. MS is a symbol for *mean square*. Now recall what is under the square root radical in the formula for the estimated standard deviation of the population:

$$s = \sqrt{\frac{\Sigma x^2_{\text{sample}}}{n - 1}}$$

Remember, too, that what is under that radical is the estimated *variance* of the population:

$$s^2 = \frac{\Sigma x^2_{\text{sample}}}{n - 1} = \text{estimated population variance}$$

Now notice that s^2 is a *mean of the squares* of individual deviations from the population mean. When calculating the F ratio, statisticians nearly always use MS (read "mean square") instead of s^2 to refer to variance; thus,

$$F = \frac{s^2_b}{s^2_w} = \frac{\text{MS}_b}{\text{MS}_w}$$

where s^2_b/s^2_w is the familiar F ratio of two variance estimates [Formula (10-3)]. Since the mean of the squared deviations (MS) *is* the variance, the ratio of mean squares is a ratio of variances—that is, it is F.

After the *F* Test

When an F test turns out to be significant, we know (with some specified degree of confidence) that there is a real difference somewhere among our means. But we don't know *where* it is.

The most obvious approach to this problem is to perform a t test on each difference, beginning with the largest and continuing until one test fails to achieve significance. That is not acceptable, however, for some of the same reasons that prevented us from using t in the first place. Take another look at the third paragraph of the preceding section; the second reason cited there is the most cogent one for not using t after F: "There are so many comparisons that some will be significant by chance." Some statisticians will approve the use of t in analysis of variance designs, either before or after analysis of variance, but only if the particular comparisons are selected on a rational basis *before the data are collected*. It is the same requirement

mentioned earlier in relation to the use of a one-tail test of significance (page 112); the reason for it is essentially the same, and the same controversy obtains.

Other tests have been devised for use in the post-F situation. All are attempts to disprove the null hypothesis, and all have been made more difficult to pass than t in order to compensate for the number of comparisons and the concomitant increase in the probability that some will show significant differences by chance.

More Complex Designs

We have examined the logic of the F test using a one-way analysis of variance as an example. Others are possible: two-way, three-way, four-way, and so forth. My first impulse was to illustrate a two-way analysis by expanding our one-way example into a methods-by-teachers design; each method would have been used by five teachers, making 6 methods × 5 teachers = 30 treatment conditions. But that would have gotten us into problems inappropriate to a discussion of this kind. For example, the performance of a teacher using any one method would probably be affected by whatever experience he or she had had with the other methods; that would have taken us into the problem of counterbalancing the design. Also, the larger number of cells in such a matrix would have made your comprehension of interaction effects more difficult than I believe is necessary. So let's leave that one, with the passing comment that once the design problems have been solved and the treatments applied, data from such an experiment can be evaluated by analysis of variance techniques.

What we want now is the simplest design that can be used to illustrate the basic principles of two-way analysis of variance. Such a design is called a *2 × 2 factorial* design. Also, we'll select treatment variables that do not require a counterbalanced design. As in the preceding illustration, calculations will be extremely simple.

The dependent variable in this experiment (the one affected by the treatment) is "persistence in the face of failure"—specifically, the amount of time spent on an insoluble problem. The two treatment variables are stress and self-confidence. The design is shown in Table 10-5 as a 2 × 2 matrix in which there are four cells, each of which represents one treatment condition—(1) low stress with low confidence, (2) low stress with high confidence, (3) high stress with low confidence, and (4)

TABLE 10-5 A 2 × 2 factorial design with index numbers of subgroups of subjects as cell entries

		Confidence: Low	Confidence: High	
Stress	High	3	4	(Group 3-4)
	Low	1	2	(Group 1-2)
		(Group 1-3)	(Group 2-4)	

high stress with high confidence. (Find the corresponding index numbers in the table). There are 25 subjects in each treatment condition; each subgroup is a random sample of a population of 10-year-old American males, and the question to be answered is whether they will *remain* random samples of a single population, with respect to persistence, at the end of the experiment. The null hypothesis says they will.

Several hours before the experiment begins, all subjects receive a painful electric shock "accidentally" while playing with some laboratory equipment. (As will become apparent later, it is important that they know what it is like to be shocked.) Then, just a few minutes before the persistence task is introduced, they all take a short paper-and-pencil test. The test is also the same for all subjects; however, half of them (group 2-4) are told that their performances were successful and the other half (group 1-3) that their performances were not. We are manipulating their self-confidence.

Immediately after that experience, they are all introduced to the persistence task. The problem is insoluble, but the subjects don't know that. They are all told that they should do the best they can, but that they may leave at any time they wish. Then half of them (group 3-4) are told that if they fail the test they will receive several shocks of the kind they had experienced earlier, whereas the other half (group 1-2) are told nothing of the shocks; thus, we are manipulating stress as a second independent variable.

Now, analysis of variance offers three advantages over the z or t type of significance test: (1) It can compare the effects of more than two categories of a treatment variable;[3] (2) it can compare the simultaneous but separate effects of two or more variables; and (3) it can assess the interaction effects of two or more variables. The first of those advantages was illustrated in the preceding section. The second is attained by computing an F ratio for each treatment variable. In the present case, we would do one F test for the "stress" *main effect* (group 1-2 versus group 3-4) and another for the "confidence" main effect (group 1-3 versus group 2-4). The most interesting of the three features of analysis of variance, but also the most difficult to understand, is the last named: its ability to identify *interaction effects*.[4]

The best way to explain interaction is to cite an example, so let's return to our experiment. Table 10-6 is like Table 10-5 except that the index number of each subgroup has been moved to the upper left-hand corner of that subgroup's cell, and the number in the middle of the cell is the subgroup's mean score (time spent before quitting). The data in the table show what look like two substantial main effects. (Whether they are significant depends also on the amount of variance *within* the two groups that are being compared in each case; let us assume a sufficiently small amount.) Group 3-4 (50 minutes) is more persistent than group 1-2 (30 minutes), and group 2-4 (50 minutes) is more persistent than group 1-3 (30 minutes). High stress produces more persistence than low stress, and high confidence produces more persistence than low confidence. But there is no interaction.

TABLE 10-6 Group means (minutes at task) in two-way analysis: outcome 1

Stress	Confidence: Low	Confidence: High	
High	[3] 20	[4] 30	50
Low	[1] 10	[2] 20	30
	30	50	

By contrast, imagine that the outcome of the experiment is as depicted in Table 10-7. There is one main effect (high versus low confidence), and there is interaction because increasing stress (from low to high) has a different effect on subjects in the low-confidence condition (down 10 minutes) than it does on confident subjects (up 10). Similarly, confidence has a different effect in the high-stress condition (up 20) than it does in the low-stress condition (no change). The interaction effect may be seen more clearly in the graphic representation of the two outcomes shown in Figures 10-4 and 10-5.

Very different patterns, aren't they? In outcome 1, there is a very simple main effect of stress shown in the left diagram and of confidence on the right. Outcome 2, however, shows that the effect of high stress—as opposed to low—is to *increase* persistence in subjects who have recently experienced success and to *decrease* it in those recently subjected to failure. It also shows that the effect of an experience of success (high-confidence condition) is to increase persistence dramatically in the high-stress group but to make no change at all in the low-stress group.[5]

We might be tempted to speculate about possible explanations for those results. However, this is a treatise on statistics, not psychology; in addition, these data were not derived from any real experiment. The main point is that interactions do occur, and that when they do, they can be detected by analysis of variance.

The calculation of three F ratios (stress main effect, confidence main effect, and stress-by-confidence interaction) for outcome 1 is illustrated in Box 10-3. The sizes of the groups here are more realistic than those in Box 10-2, but they are still very small. That should help you to follow the calculations.

TABLE 10-7 Group means (minutes at task) in two-way analysis: outcome 2

Stress	Confidence: Low	Confidence: High	
High	[3] 10	[4] 30	40
Low	[1] 20	[2] 20	40
	30	50	

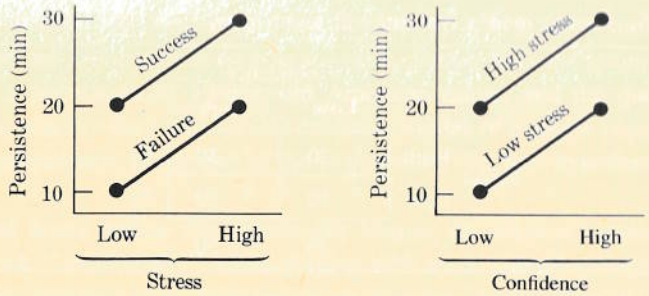


FIGURE 10-4 Graphic representation of outcome 1, showing two main effects and no interaction.

BOX 10-3 Sample Calculation of a Two-Way Analysis of Variance, Outcome 1 (see Table 10-6 and Figure 10-4)

Data table

Stress	Confidence: Low	Confidence: High
High	3: 17, 23, 21, 19 $\Sigma = 80$ $N = 4$ $\bar{X} = 20$	4: 29, 27, 33, 31 $\Sigma = 120$ $N = 4$ $\bar{X} = 30$
Low	1: 9, 7, 13, 11 $\Sigma = 40$ $N = 4$ $\bar{X} = 10$	2: 19, 17, 23, 21 $\Sigma = 80$ $N = 4$ $\bar{X} = 20$

Source table

Source	df	SS	MS	F
Stress	1	400	400	59.97
Experience	1	400	400	59.97
Stress × experience	1	0	0	0.00
Within	12	80	6.67	
Total	15	880		

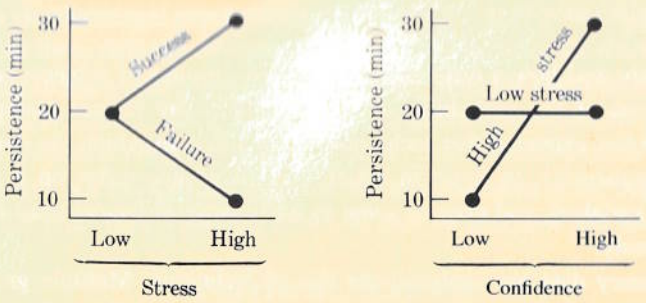


FIGURE 10-5 Graphic representation of outcome 2, showing interaction and one main effect.

$$F = \frac{MS_b}{MS_w}$$

$$F_S = \frac{MS_{b_S}}{MS_w} = \frac{400}{6.67} = 59.97 \qquad (p \leq .01)$$

$$F_C = \frac{MS_{b_C}}{MS_w} = \frac{400}{6.67} = 59.97 \qquad (p \leq .01)$$

$$F_{S\times C} = \frac{MS_{S\times C}}{MS_w} = \frac{0}{6.67} = 0$$

where b_S indicates "between groups that differ in amount of stress," b_C indicates "between groups that differ in self-confidence," and $S \times C$ indicates "interaction of stress with self-confidence. You already know the meaning of MS_w."

The data table displays the raw scores of all of the subjects, how the experiment was organized, and the means of all of the subgroups. The dependent variable throughout is "persistence," as measured by time on task.

The first two sources listed in the source table are between-groups variation. The persistence of the subjects may differ as a function of the amount of stress to which they were exposed or as a function of their recent experience of success or failure. There is also the possibility that the factors of stress and confidence interact. That possibility is represented as "stress × confidence." The final source, "within," comprises the deviations of individuals from their various group means. (The "within" row is often labeled "error" or "residual.")

The sums of squares (SS) and the mean squares (MS) are calculated as they were in the one-way analysis of variance illustrated in Box 10-2. In this case, however, there are three calculated F's rather than just one: with one F specifying the significance of the variation caused by each of the independent variables and a third concerning their interaction. It should be apparent that in this experiment both stress and confidence were associated with significant variation between groups but that there was no interaction between them.

SUMMARY

Chapter 9 showed how a difference between two groups of subjects can be evaluated in terms of the probability of its occurrence by chance (sampling error). Chapter 10 has extended that technique to situations in which the data are in the form of *frequencies*, to those in which *several* groups differ on a given factor (treatment variable), and even to some in which several factors must be evaluated simultaneously.

Frequency data are analyzed via the χ^2 technique. Multiple groups are addressed via *one-way analysis of variance* and the calculation of an *F ratio*. An example is also given of a more complex design to which analysis of variance might be applied—a two-way analysis in a 2×2 *factorial design*. A distinction is drawn between *main effects* and *interaction effects*. Because *F* functions initially as a kind of survey test, the follow-up problem—the identification of the precise source of these effects when *F* has proven significant—is also discussed.

Sample Applications

EDUCATION

1. In an attempt to decrease the number of high school dropouts, a school system develops a vocational training program for students who are at high risk for becoming dropouts. During the first several years of the program, the school staff wants to determine whether the program is effective in reducing the number of students who drop out, and you are asked to help. You select at random the 50 students that the program can accommodate from all students who apply and are at risk. Then you keep track of which students eventually graduate from high school. What statistic will help you decide whether the special program is effective?

2. You are a school psychologist. You have developed a program to help students systematically think about and plan solutions to social problems with which they may be faced in school, in dealing with friends, and in job situations. To try it out, you train half of the counselors in the school system to use the program. Teachers first identify students who are having difficulty solving social problems. Then one-third of those students are assigned randomly to trained counselors and another third to untrained counselors. The remaining third are provided no counseling at all. At the end of the semester, all students are tested to determine their ability to solve a number of social problems. How will you interpret the resulting data?

3. You are an educational psychologist. You and a group of your colleagues have developed three different educational programs to enhance the abstract reasoning ability of sixth- and seventh-grade students. With the cooperation of a large urban school district, 200 sixth- and seventh-grade classrooms are assigned to the four treatment programs (one of them a control condition), 50 classrooms per program. All students are tested on abstract reasoning at the beginning and end of the year. How could the resulting data be analyzed and interpreted?

POLITICAL SCIENCE

1. You are interpreting the results of an opinion poll. In a random sample of 100 individuals, you find that of the 40 Republican respondents a total of 30 favored a proposal to cut the capital gains tax, and that of the 60 Democratic respondents 20 support it. How can you calculate the probability that this association is due to chance?

2. Once again you are studying the incidence of military coups in Latin America. This time you want to know whether there are significant differences in the frequency with which coups occur, and you have reason to believe that the type of legitimacy upon which the regime rests (i.e., traditional, legal-rational, or charismatic) is the most important factor in accounting for the incidence of coups. In this situation you have a treatment variable (type of regime legitimacy) and a dependent variable (coup frequency). How can you demonstrate that types of regimes probably do (or do not) differ with respect to coups?

3. One of the most frequently studied questions in the field of international relations pertains to the relationship between domestic politics and foreign policy behavior. You are in that field and you want to test the idea that a nation's form of government (democratic, authoritarian, or totalitarian) and type of leadership (unitary, collective, or fragmented) affects the number of aggressive actions it initiates (the dependent variable). What kind of analysis is appropriate?

PSYCHOLOGY

1. You are the director of a clinic dealing solely with phobic behaviors (unusual fears) of children. Over a two-month period, 100 children are referred to your clinic for the treatment of agoraphobia (fear of large open spaces). After hypnosis therapy, you find that 55 of the 100 children are able to walk in a large open field and report little or no fear. These are fairly impressive results, but are they statistically significant? How can you tell?

2. You are a child clinical psychologist interested in which of three approaches is most effective in controlling the pain experienced by children in a hospital burn unit. You assign, at random, equal numbers of children to (1) a mental distrac-

tion condition, where the participants attempt to keep their minds off the pain by doing mental arithmetic; (2) a self-reward condition, where participants give themselves positive self-statements (e.g., "I'm a brave person") for enduring pain; and (3) imagination exercises, where the participants are taught to imagine situations in which they experience pleasure and satisfaction. How might you go about analyzing the results?

3. You are a psychotherapist at a child guidance center. You are interested in the possible interaction between personality factors and the effectiveness of psychotherapy. You assign each of 20 introverted (shy and socially withdrawn) children to either individual or group therapy. Next, you assign 20 extroverted (socially outgoing) children to the same conditions. Your prediction is that the introverted children, because of their social fears and shyness, will benefit most from individual therapy and that the extroverts will work best in a group setting. Presuming that you have a generally accepted criterion of effectiveness of therapy, how might you analyze the data from this study?

SOCIAL WORK

1. You are a social worker in a small rural community. You conclude that the existing manner of dealing with juvenile offenders through the courts is of limited efficiency and effectiveness. The magistrate is present only one and one-half days a week, and the court calendar is always too full to handle the number of youth being petitioned into court.

   You develop an alternative to this process by convincing local citizens and the judge to establish a juvenile review board. The board, composed of community people, will decide cases and use measures such as community work service, restitution, and active parental involvement instead of the jail time and probation that have been the remedies usually prescribed by the courts. Part of the agreement in establishing the alternative program is an evaluation of its effectiveness, at the end of one year, in comparison with the effectiveness of the courts traditional procedures. Low recidivism is selected as the major indicator of success, and youth will be assigned to the alternative program or to the courts on a random basis, with violent offenders excluded from both groups.

   When the one-year period is over and all the data are in, how do you evaluate the results?

2. In your child guidance clinic there is disagreement concerning how to treat children who display hyperactive behavior. One social worker with a psychoanalytical background favors play therapy with two sessions a week for a minimum of one and one-half years. The agency psychologist favors a behavioral model and the use of operant conditioning. A consulting psychiatrist views hyperactivity as the result of immature cortical development, and she recommends the drug methylphenidate to stimulate the cortex.

   Since hyperactivity is so frequent and so troublesome, you recognize it as a major problem. You decide to conduct a study to see which is indeed the most effective of the three treatment methods; you also decide to add a fourth group of children who will receive no treatment. (The clinic has a waiting list.) Children will be assigned randomly to those four groups. The social worker, the psychologist, and the psychiatrist all agree on a single test score as the criterion of effectiveness. Once you have obtained all your data, how do you evaluate it?

3. As a social worker in a foster care agency, you are interested in how the age of children and youth affects their adjustment to either a foster family home or a staffed community group home.

   The independent variables are age and type of foster care arrangement. Age is defined by two categories: "children" (5 though 12 years) and "youth" (13 through 18 years). You obtain a standardized instrument to measure the dependent variable, adjustment to foster care.

   Twenty children are randomly assigned to either a foster family home (10 children) or a group home (10 children), and 20 youth are randomly assigned to either a family home (10 youth) or a group home (10 youth).

   Are there any reliable differences in adjustment among those four groups? If so, can they be traced to age differences, the types of foster care, or both? Is there an interaction between the two? Those are the questions that your design is supposed to address; how do you treat the data statistically to obtain the answers?

SOCIOLOGY

1. You are still a consultant to the family planning group described in the sociology exercise for Chapter 9. This time you are asked about the antecedents of different attitudes toward abortion. You cannot answer such a broad question all at once, but one of your many hypotheses is that a person's belief about when human life begins influences his attitude toward abortion. Some questionnaire data are available from another study, and you are able to find two questions pertinent to your hypotheses: (1) "Do you believe that human life begins before or after the 90th day following conception?" and (2) "Do you approve of abortion on demand?" The respondents' answers to these questions are your data. What do you do with them?

2. That family planning agency is intrigued by your answer to its earlier question about the effect of religious belief on family size (see the sociology exercise in Chapter 9). The agency wants to broaden the investigation by comparing the mean sizes of Catholic, Mormon, and Mennonite families with each other and with that of the general population. You gather the data and find that there are differences among the means. How can you tell whether these differences are significant?

3. You are asked to identify some of the critical variables that inhibit or enhance communication of sex information to children by their parents. "Amount of sex information communicated" is therefore the dependent variable in your study. You construct an instrument that yields a sex information score and administer it to parents of 10-year-old children. (The score represents the amount of information that parents believe they would give to their 10-year-old if the child were to ask "Where do babies come from?") You hypothesize that the amount of information contained in the parents' answers to this question is (1) inhibited by church involvement and by lack of education and (2) enhanced by little or no church involvement and high education. You define "church attendance" as attendance four times a year or more, "nonattendance" as attendance of three times a year or less. One or more years of college defines the "high education" group, and less than a year of college identifies the "low education" group. How will you analyze the data that emerge from this study?

11

Correlation, Causation, and Effect Size

Before we can be certain that we have found an instance of *causation*, we must observe a reliable relationship between the purported causative agent and the event being caused. If every time you push a switch in your bedroom to the up position a light goes on, you come to believe that the latter event is caused by the former, even in the absence of wiring diagrams, electrical theory, and so forth. If your dog becomes wildly aggressive whenever the garbage man approaches, you say that the man's approach is causing the dog to be upset. If in studying a large sample of schoolchildren you find that those with larger vocabularies have higher IQs, you assume that the vocabulary is a cause of the intelligence.

But should you? Is it legitimate for you to assume that when two events occur together one is caused by the other? Consider the following case: What size of coefficient would emerge if we were to compute a correlation between the revolutions per minute of the left and right front wheels of your car (Figure 11-1)? The correlation would be pretty high, wouldn't it? If the car were on a curving road and turns in one direction were predominant throughout any given minute, the two scores would be different but correlated for that minute, and with varying speeds on a straight road, the correlation would be almost perfect.

But does the speed of one wheel *cause* that of the other? Only to the extent that each is a functioning part of a total system in which the two events occur. It would

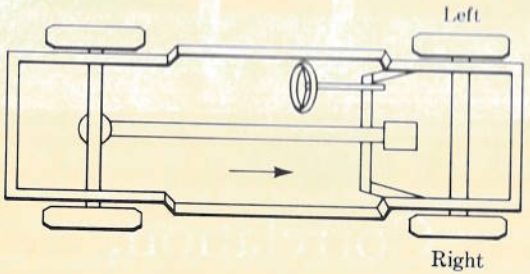


FIGURE 11-1 Top view of an automobile chassis.

be more accurate to say that the functioning system—the car in motion—causes the revolutions of both wheels.

I don't suppose anyone will ever bother to find the correlation between the two front wheels of an automobile, but the same principles apply to many situations in which correlation coefficients *are* commonly computed. For example, if we were to study a large sample of elementary school children, we might find a tendency for those children who display extensive vocabularies to be superior students in arithmetic. We might be tempted to conclude that a good vocabulary causes competence in arithmetic. But then we find that chronological age correlates with both variables. Might it not be better to say that achievements in both vocabulary and arithmetic are caused by (are parts of the functioning system of) the child's developing intelligence?

The caution that I am advocating here has important practical implications; without it we might easily be persuaded, for example, to set up elaborate vocabulary-building programs for children who show weakness in quantitative thinking. Some wag once noted that there is a substantial correlation between the intelligence of boys (as indicated by their mental ages) and the length of their trousers. He suggested that a relatively inexpensive way to increase the intelligence of boys would be to increase the length of their trousers!

So although correlation is a *necessary* feature of a causal relation, it is not *sufficient* to prove that a causal relation exists.[1] Whether the relation is interpreted as a causal one should depend not just on the correlation of two variables but also on some rational link between them—on the extent to which the relationship makes sense within some sort of conceptual framework (within a wiring diagram, for example, or a sociological theory) or, better yet, on the elimination of alternative possibilities, as elaborated below.

Correlation is not causation.

CORRELATIONAL VERSUS EXPERIMENTAL STUDIES

People who design research studies usually distinguish between correlational and experimental designs. Most *correlational* studies are concerned directly with relationships among variables that occur naturally—that is, without the intervention of the investigator. An example might be the relationship of poverty to intelligence. If there is a unique entity called "poverty" that can be measured, and if there is another unique entity called "intelligence" that also can be measured,* then a researcher can discover the relationship between poverty and intelligence and report that relationship numerically—probably with a Pearson *r*. If the *r* is significant, the researcher can make a direct inference of *relationship* but no inference of a causality in the relation between the variables. (Does poverty cause low IQ, does low IQ cause poverty, or does yet another causal structure underlie the relation?)

It is this conventional use of correlational information that was described in Chapter 6. However, unconventional uses have become more common in recent years—uses that may eventually redefine what constitutes conventional use. This chapter describes some of those new ways to interpret correlations.

Traditionally, *experimental* investigations are designed to identify *causes*. If it were possible for an investigator to *manipulate* poverty, it could be determined whether its relation to intelligence is causal. In this design the researcher would manipulate certain variables instead of merely observing them. In psychology, the simplest such case would have all relevant treatment variables held constant except one. That one would be manipulated by the experimenter, who could then make a direct inference that any change in the behavior of the subjects (in this case, their performance on an IQ test) is caused by a change in the manipulated variable (in this case, their economic well-being). The manipulated variable is called *independent*,** variables that are held constant are *controlled*, and the behavior of the subjects is the *dependent* variable (because their behavior depends on what happens in the independent variable).

I have just said that variables that are held constant in an experiment are known as controlled variables. And so they are, but a so-called controlled variable can also vary at random (be entirely *un*controlled); the important thing is that its variance be unrelated to that of the independent variable. In the example, the investigator would manipulate the environments of the subjects; if the children who were low on a scale of economic level later turned out to be low on a scale of intelligence, then it would be tempting to infer that poverty causes low intelligence. But if infants placed in poverty were *genetically* inferior to those assigned to other environments, no such

* To simplify the discussion and keep it focused on statistics, we shall assume the validity of both of these propositions.

**In many applications, the term *treatment* can be used in place of *independent*.

inferences could be made because there would be no way of knowing whether the genetic difference or the environmental difference had caused the observed difference in intelligence.

To make legitimate the inference that poverty causes low intelligence, the experimenter would have to make sure that genetic variance (a controlled variable in this experiment) is unrelated to variance in economic status (the independent variable). If genetic potential varies at random (e.g., high potential is just as likely to be found in the poor environment as in the good one), then a difference of intelligence in favor of the good environment would imply that good environment is a *cause* of high intelligence. The purpose of randomizing a "controlled" variable is the same as that of holding it constant: to prevent any systematic relation between that variable and the one that is being manipulated (the independent variable).

The purpose of all this planning and manipulation is to exclude other possible explanations of whatever change is observed in the dependent variable. If all but one of the potential independent variables are controlled, then that one must be responsible for any observed change in the dependent variable. That is an ideal to be approximated as closely as possible; an explanation is said to have *internal validity* to the extent that alternative explanations can be excluded.

On the other hand, the attempt to achieve internal validity may threaten *external validity*. For example, if a laboratory experiment is so well controlled that its subjects experience it as distinctly different from circumstances outside of the lab, its conclusions may not be valid in any other situation. External validity is the extent to which the results of a study can be generalized.

If your main purpose in reading a research report is to understand the *underlying structure* of a particular class of events, then your primary concern will be with *internal* validity. If your intent is to *apply* the results, then you want *external* validity. If you are concerned about both, you may have to accept some kind of trade-off, for internal and external validity are sometimes inversely proportional to each other.

Traditionally, the Pearson *r* has been used in correlational studies and only in correlational studies. It is, after all, a coefficient of correlation. But if scores on a dependent variable can be shown to be correlated with scores on a manipulated (independent) variable with all other variables held constant (or otherwise dissociated from the independent variable), that correlation *can* be used as evidence of a causal relationship.

It is true, however, that whereas the Pearson *r* was invented to deal with refined measures of continuous variables, many experimental studies use rather crude measurements (e.g., the categories "low" and "high" or "low," "medium," and "high" for the variable "economic level"); so the issue of continuity must be addressed.

CONTINUOUS VERSUS DISCONTINUOUS VARIABLES AND MEASUREMENTS

The Pearson *r* is an index of relationship between two continuous variables. (Presumably those variables are derived from parameters that also vary continuously; we shall make that assumption in the discussion that follows.) Many variables are not continuous, however, and even those that *are* continuous may be measured in such a way that their data are not. An example of a dichotomous variable is gender. Virtually all humans are either male or female; there is no continuous dimension underlying the two categories, for the difference between them is *qualitative*. Conversely, although the evaluation of college students in a pass–fail course also produces a report containing just two classes of data (Figure 11-2), each class represents a much larger number of categories that *could* have been reported, and the differences are *quantitative*. If the professor uses a point system to determine who fits into which category (and if there are many students in the course), the distribution of students on scores will be long and nearly continu-

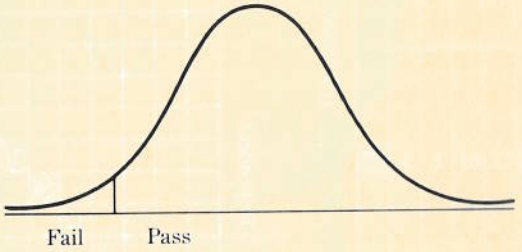


FIGURE 11-2 A possible distribution of point scores in a pass–fail class. The lowest score is less than 50 points, the highest several hundred; but only two grades are reported.

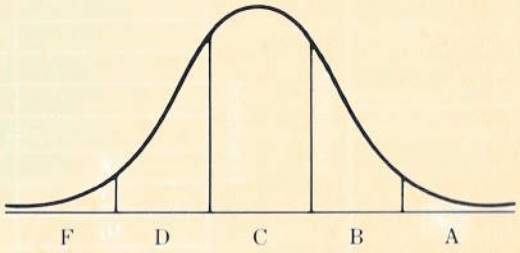
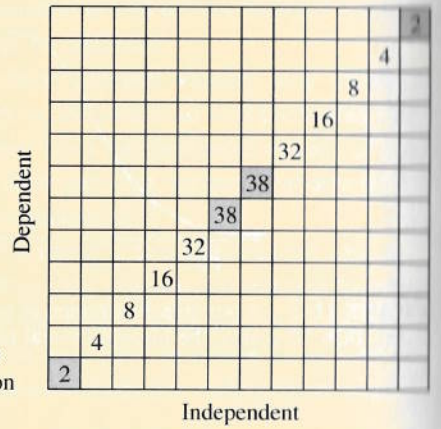


FIGURE 11-3 The same distribution of point scores as in Figure 11-2 but with five grades reported instead of two. There are still many different scores within each class interval.

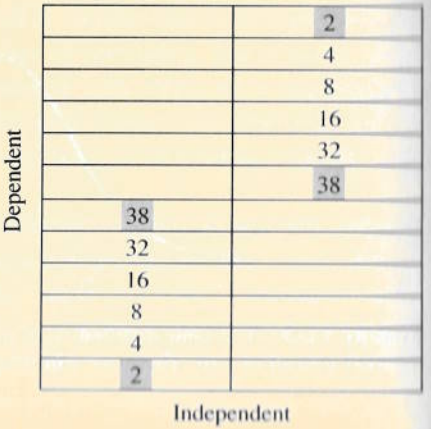
ous. No matter where the professor draws the line between "pass" and "fail," there will be students very close to the line on either side, but there will only be two categories in the professor's report to the registrar. Even a report of letter grades will comprise only five categories (Figure 11-3), although there *could* be many more (up to a maximum of one for each score). Theoretically, if the process of measurement were continued indefinitely, there would be an *infinite* number of scores and potentially an infinite number of categories. Such a variable is continuous.[2]

As you have known since Chapter 2, when a continuous variable is represented by data in a large number of categories, those categories are called *class intervals*. When a continuous variable is represented by only a few categories of data, it might be helpful to think of the categories as extremely large class intervals. (Of course, when the variable is by its very nature *not* continuous, as in the gender example cited earlier, the investigator has no choice but to limit the number of categories.)

Table 11-1 Four possible displays of data on two continuous variables ($r = 1.00$)



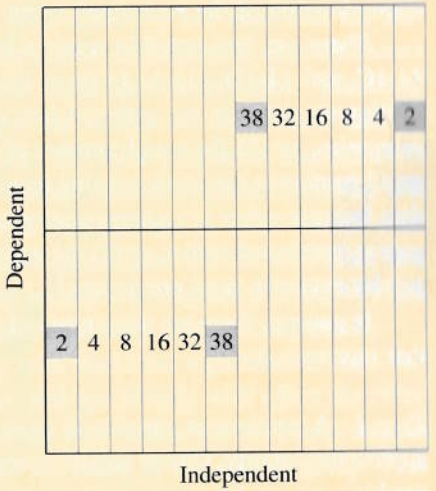
A Frequencies are distributed across 12 class intervals—a nearly continuous distribution—on each of the two variables.



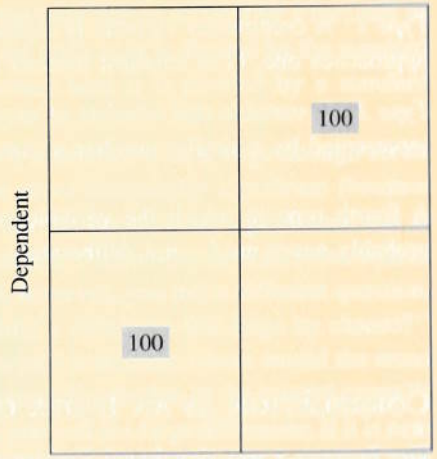
B There are two distinct groups on the independent variable, while the distribution remains nearly continuous on the dependent.

Table 11-1 features four data displays. Each display is itself a table, but it also has graphic qualities. For example, compare the spacing of scores on the independent variable in Table 11-1A with that in Table 11-1B: in A the spacing is nearly continuous, whereas in B it is dichotomous. Be aware of those special relations as you examine the rest of Table 11-1.

Table 11-1 shows four possible ways of organizing measurements taken on two continuous variables from 200 subjects. Independent measures that vary in *amount* get larger from left to right in each table. Dependent measures that vary in amount get larger from the bottom to the top of the graph. (There is no convention to guide the investigator whose independent or dependent variables are *qualitative*).



C The independent variable remains nearly continous, while the dependent has been dichotomized.



D Both variables have been dichotomized.

In Table 11-1A, both data sets are continuous; in Tables 11-1B and 11-1C, one is continuous, the other dichotomous; in Table 11-1D, both are dichotomous. Table 11-1A carries the most information and Table 11-1D the least, as an examination of the tables will reveal. (Besides examining these tables analytically, you can make an intuitive comparison by holding the page nearly parallel to your line of sight, first looking at it from the bottom, then from the left side of the page. The numbers will be illegible, but the contrast between continuous and discontinuous distributions will be enhanced.) In Table 11-1B, the 2 lowest scores on the independent variable are treated exactly the same as the 38 that are almost as high as the median, and the 2 highest are treated the same as the 38 that are barely above the median. In Table 11-1C, the same is true of the dependent variable; in Table 11-1D, it is true of both. An r computed for continuous variables from tables in which at least one variable was treated dichotomously (as in Tables 11-1B, 11-1C, and 11-1D) is only an approximation of the r obtained when the data are arranged as in Table 11-1A.

There are two possible reasons for data reports like those in Tables 11-1B, 11-1C, and 11-1D: (1) As in these illustrations, data from a continuous or nearly continuous variable are gathered into a few—here two on each variable—broad categories, or (2) the small number of categories in the data represent an equally small number in the variable being measured. So a dichotomy could occur in a set of data because either (1) administrative concerns are compelling, as in the pass–fail course mentioned earlier, or (2) the variable naturally breaks into two parts, as in the gender example cited previously.[3]

In general, there are three types of relationship between a variable and the data that represent it:

Type 1: A continuous variable is represented by a number of data categories that approaches infinity. (There are, of course, practical limits to the number that can be used.)

Type 2: A continuous variable is represented by a number of data categories that approaches one. (The smallest number that can be used is two.)

Type 3: A variable that consists of only a few categories (commonly two) can be represented by a similar number of data categories.

A fourth type in which the variable is dichotomous and the data continuous is probably never used—not deliberately, anyway.

CORRELATION AS AN INDEX OF CAUSATION

The Pearson r was developed for situations in which the relationship among the correlated variables and their data is best described as type 1. However—and right now this is the main point—correlation estimates have since been developed that can be used even when that relationship is better described as type 2 or 3.

Those estimates have made it possible to use correlation in experimental studies—that is, those that manipulate one or more independent variables—and to observe whether one or more dependent variables change. This section is concerned with the use of r in such studies. We shall focus on its use in the simplest kind of experiment: assessing the effect of a single variable on another single variable.

You will recall that kind of study as the very one with which we were concerned in Chapter 9, which was mainly about the statistical significance of an obtained difference between two means. You may also recall, however, that one section of that chapter was entitled "Statistical versus Practical Significance." *Statistical significance* is concerned with the *reliability* of an effect (e.g., of the difference between control and experimental groups), regardless of the size of the effect; *practical significance* is concerned with reliability but also with the *size* (either *amount*, as in difference between mean scores, or *frequency*, as in difference between numbers of votes cast). That section could have been called "Statistical Significance versus Effect Size," although that would not have been entirely accurate because practical significance includes the value judgments that still have to be made even after the size of an effect is known. (Given a substantial effect, is it worth the cost of attaining it?) The "versus" in the title is there to call your attention to the difference between statistical significance on the one hand and practical significance on the other; it is *not* intended to suggest that the two are alternatives. One contributes to the other, and ultimately both are important.

But Chapter 9 was concerned mostly with statistical significance; here our concern is mainly with practical significance and specifically with *effect size*. One index of effect size is the difference between two means (of a control and an experimental group) divided by the standard deviation of the control group. That *looks* very much like the significance test that we discussed in Chapter 9. Like the significance test, it consists of a difference between means divided by a measure of variability. In a significance test, however, the difference is divided by the standard error of a distribution of differences, whereas here it is divided by a standard deviation of individual raw scores.

Because the standard error shrinks as n increases, with a large enough set of measures even a very small difference can prove statistically significant (because the divisor is so small). That is as it should be, because the significance test is concerned with the *reliability* of the difference, and reliability does increase with n. If you are interested in the *size* of an effect, however, you ask a different question. Instead of "What is the probability of getting a difference this large by chance?" you ask, "Assuming that my obtained difference is reliable, where would the mean score of the treatment group fit into the obtained distribution of the control group?" If it is in one of the tails of that distribution, you call it a large difference; if it is near the mean, you have to admit that it is rather small. In neither case do you know how reliable it is.

There are now several commonly used indices of effect size. The one just discussed is a ratio between an observed difference and a standard deviation, but most of them are variants of the Pearson *r*. In the remainder of this section, we shall limit our discussion to those variants; we shall emphasize their similarities (indeed, we shall treat them all alike), and we shall symbolize each of them with a simple *r*.

Table 11-2, in conjunction with Table 11-1, should help you to understand how *r* can be used to interpret the results of an experimental study. In Table 11-2 two experiments are represented in which the success rate (dependent variable) resulting from treatment by therapeutic intervention (independent variable) was recorded. There are 200 subjects in each study, but notice how differently the subjects are distributed. In the table for study A the success rates for the control and experimental groups are the same. In study B, all of the 100 subjects in the control group are in the low-success-rate cell of the table, while the entire experimental group is in the high-success cell.

You might think of the tables in Table 11-2 as scatterplots (pages 64ff) in which the number of class intervals in each variable has been reduced to two. You may get a better intuitive feel from this report of an experimental study than you did from Table 11-1, because Table 11-2 includes a zero correlation for comparison with a perfect one. Study A of Table 11-2 indicates that the therapeutic intervention had no effect at all; in study B, its effect was maximal. Of course, most correlations computed from actual data fall somewhere between those extremes, but the extremes serve better than actual data to illustrate the meaning of correlation—in this instance, of a correlation computed from dichotomous data. (Both sets of *data* are

TABLE 11-2 Zero and perfect correlations as indicators of the effect size of a therapeutic intervention

Study A: No correlation ($r = .00$)

		Treatment	
		Control	Experimental
Success rate	Substantial improvement	50	50
	Little or no improvement	50	50

Study B: Perfect correlation ($r = 1.00$)

		Treatment	
		Control	Experimental
Success rate	Substantial improvement	0	100
	Little or no improvement	100	0

clearly dichotomous, but the structure of the *underlying variables* is not so obvious. One variable—"treatment"—can be conceived as dichotomous, but the other—"success rate"—is essentially continuous.)

Even though most correlations fall far short of perfection, they can yield information that is of both scientific and social importance. The psychotherapeutic intervention suggested above would probably produce an *r* somewhere between the two extremes illustrated in Table 11-2. Table 11-3 lists 18 independent-to-dependent-variable correlation coefficients and shows the difference between the success rates of the control and experimental groups that correspond to each *r*. For example, a very modest $r = .30$ indicates a change in success rate from .35 (35 of the 100 *untreated* patients improved during the period under study) to .65 (65 of the *treated* ones did); an *r* of .50 corresponds to an increase from 25 untreated to 75 treated successes; when *r* is .70, the increase is from 15 in the untreated group to 85 in the treated—a difference of 70 percent! Notice that *in every case, the difference in the success rates is identical to r*. Somebody still has to decide whether the payoff is worth the effort, but there should no longer be a question that there *is* a substantial payoff, even when *r* is as low as .30.[4]

TABLE 11-3 Eighteen levels of correlation as indicators of the effect size of a therapeutic intervention

	Success rate		
r	Control group	Experimental group	Difference
.00	.50	.50	.00
.02	.49	.51	.02
.04	.48	.52	.04
.06	.47	.53	.06
.08	.46	.54	.08
.10	.45	.55	.10
.12	.44	.56	.12
.16	.42	.58	.16
.20	.40	.60	.20
.24	.38	.62	.24
.30	.35	.65	.30
.40	.30	.70	.40
.50	.25	.75	.50
.60	.20	.80	.60
.70	.15	.85	.70
.80	.10	.90	.80
.90	.05	.95	.90
1.00	.00	1.00	1.00

Source: Adapted from R. Rosenthal, "Assessing the Statistical and Social Importance of the Effects of Psychotherapy," *Journal of Consulting and Clinical Psychology* 51 (1983): 12.

"Correlation is not causation." I said that on page 140, and I repeat it here. From the information in this section, however, it should be clear to you that under certain circumstances a coefficient of correlation *can* serve as an index of causation. And if Table 11-3 is any indication, in those special circumstances (those that constitute an *experiment*) the coefficient can be more clearly interpreted than it typically is in more traditional correlational studies. Moreover, Table 11-3 suggests that the practical implications of many experiments can be revealed more clearly if the results are expressed as correlations rather than differences between means.

This use of correlations is not likely to replace the use of differences between means in experimental studies. But even if it never does, you can use the information in this section to interpret in concrete terms studies of traditional *correlational* design. Table 11-3 shows success rates and corresponding correlation coefficients that might be obtained from an *experiment* (here a study of the effects of psychotherapy), but it could be used to interpret *any* correlation. The interpretation would be of the form "If I were to split the distribution on each of the two correlated variables into halves, the relations given in the table would obtain" (but without any inferences as to causes).

Two examples are given below. As you read them, look at the corresponding table. Each cell in the four-cell table gives the proportions not of the entire sample but of those subjects who are in the upper or lower *half* of the sample of a given variable.

In Table 11-4, an r of .40 between sugar intake and hyperactivity means that as compared to children in the lower half of the sugar-intake dimension, 40 percent more of those in the upper half are also in the upper half on a scale of hyperactivity. Because this is not an experimental study, you cannot be sure what causes what, but the table can help you judge the importance of the relation.

In Table 11-5, an r of $-.50$ between women's self-esteem and the duration of child-support payments received means that 50 percent more of the women in the upper half of the distribution on self-esteem are in the *lower* half than are in the

TABLE 11-4 Median split of both distributions (X and Y) when $r_{xy} = .40$

Hyperactivity (Y)	Lower half of sugar intake (X)	Upper half of sugar intake (X)
Upper half ↑	.30	.70
Lower half ↓	.70	.30

TABLE 11-5 Median split of both distributions (X and Y) when $r_{xy} = -.50$

Child-support payments (Y)	Lower half of self-esteem (X)	Upper half of self-esteem (X)
Upper half ↑	.75	.25
Lower half ↓	.25	.75

upper half of the distribution of duration of child-support payments. Again, the causal structure is not clear.

In general, the correlation coefficient gives you in a single number all the information you need to construct a 2×2 matrix of proportions, four entries in all, as illustrated in Tables 11-4 and 11-5. The proportions in any such table constitute an interpretation of a correlation coefficient.

Correlation is not causation, but with proper precautions it can serve as an index of the strength of a causal relationship. And whether causal or not, a relationship between two variables now can be interpreted in more concrete terms than had previously been possible.

SUMMARY

Traditionally, quantitative research has been done in either a correlational or an experimental mode. *Correlational* designs warrant direct inferences only about the relationships among variables, not about the *nature* of those relationships. *Experimental* studies, on the other hand, are designed so that inferences can be made about the *causality* of the relationships identified by the investigators. In experimental studies, *independent variables* are manipulated while *controlled variables* are held constant (or varied at random) and *dependent variables* are observed. If all relevant variables (other than the independent) are held constant, the changes in the observed variable are said to *depend* on (that is, to be *caused* by), those in the independent variable.

The Pearson r is an index of relationship between two continuous variables. Many variables, however, are not continuous, and those that *are* continuous can be represented by data that are not. Indeed, continuous variables nearly always are broken into class intervals for convenience in handling data. If there are many class intervals, the deviation of the data from continuity is trivial; however, if there are

only a few class intervals—the lower limit is two—that deviation is more substantial. Even so, very useful approximations of the Pearson r can now be extracted from noncontinuous data.

An even more recent development is a set of techniques for extracting *effect size* (as distinguished from reliability). Those new techniques include the use of correlation coefficients in interpreting experimental studies. Used in this way, correlation coefficients can be indicators of causal relations: Correlation is not causation, but causation is one kind of relationship between variables. Thus a measure of relationship (the correlation coefficient) can in the right circumstances be a measure of *causal* relationship. Moreover, once *any* relationship has been quantified as a correlation coefficient, it can be interpreted in a new way that significantly clarifies its meaning.

Sample Applications

See pages 82–83, and specify as many plausible causal structures as you can.

12

Summary

The central objective of this book has been to convey an understanding of the basic concepts and their interrelations—the "big ideas," so to speak—in statistical thinking. In Chapters 9 and 10, for example, the big idea was the logic of testing for the significance of differences, especially in experimental settings; in Chapter 6, it was correlation between variables in settings other than experimental; and in Chapter 11 it was a comparison of experimental with correlational strategies.

The first four chapters established a necessary foundation for everything that was to follow. However, in Chapters 2, 3, and 4 it became apparent that big ideas—for example, the idea of distribution, the importance of specifying the kind of average, the concept of variability—are involved even in the description of a single sample. Chapter 5 stressed the importance, when interpreting any measurement of a given subject's behavior, of comparing it with similar measurements that have been made of some identifiable reference group. Chapters 7 and 8 demonstrate that the description of a sample is not an end in itself—that inferences can be made from what you know about a particular sample to what you would like to know about the population whence it came. Be aware, however, that description and inference are independent functions. It is true that inference, when it occurs, is always from sample to population; but description may, in principle, be applied directly to a population, and in many institutional settings, that is exactly what happens.

only a few class intervals—the lower limit is two—that deviation is more substantial. Even so, very useful approximations of the Pearson r can now be extracted from noncontinuous data.

An even more recent development is a set of techniques for extracting *effect size* (as distinguished from reliability). Those new techniques include the use of correlation coefficients in interpreting experimental studies. Used in this way, correlation coefficients can be indicators of causal relations: Correlation is not causation, but causation is one kind of relationship between variables. Thus a measure of relationship (the correlation coefficient) can in the right circumstances be a measure of *causal* relationship. Moreover, once *any* relationship has been quantified as a correlation coefficient, it can be interpreted in a new way that significantly clarifies its meaning.

Sample Applications

See pages 82–83, and specify as many plausible causal structures as you can.

12

Summary

The central objective of this book has been to convey an understanding of the basic concepts and their interrelations—the "big ideas," so to speak—in statistical thinking. In Chapters 9 and 10, for example, the big idea was the logic of testing for the significance of differences, especially in experimental settings; in Chapter 6, it was correlation between variables in settings other than experimental; and in Chapter 11 it was a comparison of experimental with correlational strategies.

The first four chapters established a necessary foundation for everything that was to follow. However, in Chapters 2, 3, and 4 it became apparent that big ideas—for example, the idea of distribution, the importance of specifying the kind of average, the concept of variability—are involved even in the description of a single sample. Chapter 5 stressed the importance, when interpreting any measurement of a given subject's behavior, of comparing it with similar measurements that have been made of some identifiable reference group. Chapters 7 and 8 demonstrate that the description of a sample is not an end in itself—that inferences can be made from what you know about a particular sample to what you would like to know about the population whence it came. Be aware, however, that description and inference are independent functions. It is true that inference, when it occurs, is always from sample to population; but description may, in principle, be applied directly to a population, and in many institutional settings, that is exactly what happens.

Those are all important ideas. They won't enable you to *do* social science research, but they will enable you to understand research done by others. That is clearly a high enough payoff to have justified all the effort you have expended in mastering those ideas, but there is an additional reward that, though incidental to the stated objectives of the book, could well prove most important of all in the end. While learning to think better statistically, you may have learned to think better generally!

You may not have occasion to think statistically every day, every week, or even every month; but whenever the occasion does arise, you will be ready for it. After an especially long hiatus, you may not be able to deal immediately with every concept that you have mastered while studying the book, but you will find that a quick reference to its discussion of a particular concept will revive your understanding not only of that concept, but of many related ones as well. In fact, the habit of reviewing relevant sections of this book is a good one even while you are reading another. Books on statistical methods, for example, are based primarily on concepts presented here.

General agreement is so far lacking on a single set of symbols to represent those concepts. For that reason, I have provided on pages 155–156 a list of commonly used statistical symbols that you may encounter in your reading.

Good luck! And remember what Sir Francis Galton used to say: "Whenever you can, count."

List of Symbols

Entries in this table are listed in two categories:

1. Every "symbol used here" (i.e., used in this book) is included unless, like AD (average deviation) and ρ (rank-difference correlation), it has been described in the text only as an introduction to a more useful but more complicated concept (such as the standard deviation or the "Pearson r").

2. The list of "Other symbols" that you might encounter elsewhere may include the most common alternatives, but it is not exhaustive. There is no standard usage either accepted by authors or endorsed by the American Statistical Association.

Sample		Population		Sampling Distribution		
Symbol used here	Other symbol(s)	Symbol used here	Other symbol(s)	Symbol used here	Other symbol(s)	Verbal description
X	x	X	x			Raw score
n	N	N	n			Size (number of observations)
$\overline{X}$	$\bar{x}$, M	μ	$\overline{M}$, $\hat{M}$			Mean
Mdn						Median
x_{sample}	$X - \overline{X}$, d	$x_{\text{pop.}}$	$X - \overline{X}$, $\overline{d}$, $\hat{d}$,			Deviation score
S	s, SD, σ	σ	$\overline{\sigma}$, $\hat{\sigma}$			Standard deviation
S^2	V					Variance
		s	$\bar{s}$, $\hat{s}$			Estimated standard deviation
		s^2, MS				Estimated variance, mean square
				$s_{\bar{x}}$	$\bar{s}_{\bar{x}}$, $\hat{s}_{\bar{x}}$	Estimated standard error of the mean
				$s_{\bar{x}_1 - \bar{x}_2}$	$\bar{s}_{\bar{x}_1 - \bar{x}_2}$, $\hat{s}_{\bar{x}_1 - \bar{x}_2}$	Estimate standard error of the difference between means
		z				z ratio
				t		t ratio
				F		F ratio
				χ^2		Chi-square
r		r				Pearson product-moment coefficient of correlation

Notes

CHAPTER 3: MEASURES OF CENTRAL TENDENCY

1. If you should ever have to do the computations yourself, you would find that the median is often *within* a class interval rather than precisely between two intervals as in the above illustrations. In that situation, it is necessary to interpolate; almost any text on statistical methods will quickly tell you how.

CHAPTER 4: MEASURES OF VARIABILITY

1. The interquartile range actually extends from $\frac{1}{2}$ unit below the score at Q_1 to $\frac{1}{2}$ unit above the one at Q_3, and as with the median, interpolation is necessary if either quartile falls within a class interval.
2. With respect to the additional $\frac{1}{2}$ unit on either end, the same is true of the total range as of the interquartile range (see Note 1).

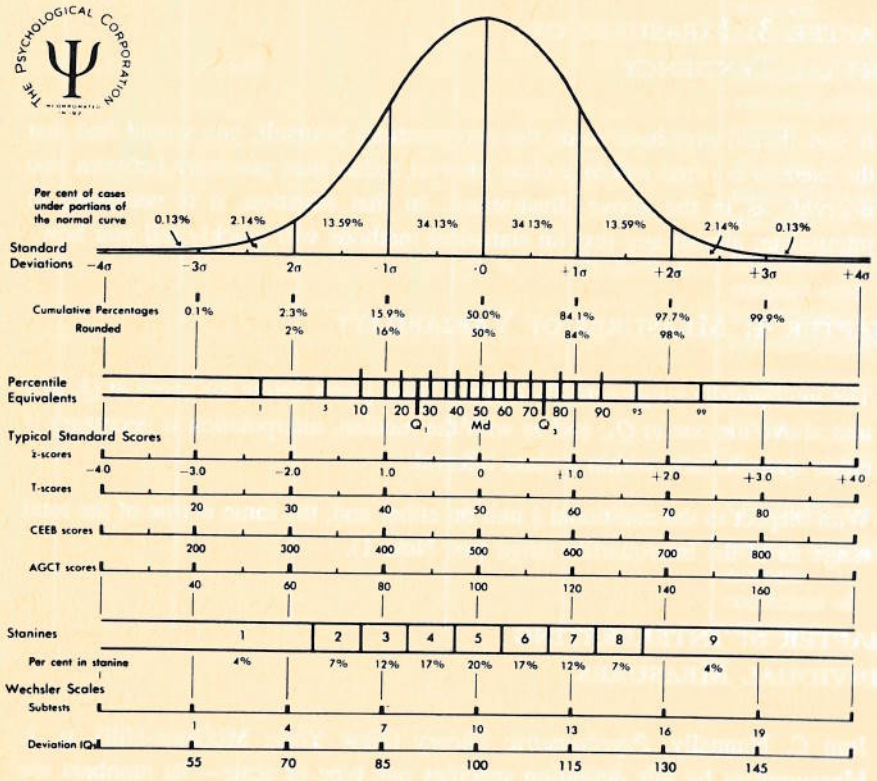
CHAPTER 5: INTERPRETING INDIVIDUAL MEASURES

1. Jum C. Nunnally, *Psychometric Theory* (New York: McGraw-Hill), p. 2. Measurement by this definition specifies one type of scale—its numbers are called *cardinal* because they are generally preferred, but most authorities recognize two other types: *nominal*, in which data are merely classified, and *ordinal*, in which the classes are placed in order of size, ability, prestige, or some other attribute that can be ordered. Some classes, like genders or countries

of origin, cannot be ordered. Others, like military ranks or beauty contest awards, can be not only classified but ordered as well. Still others, like grade-point or batting average, can be not only ordered but placed on a scale of constant units, like grams, centimeters, number of hits in a baseball season, and so forth. We prefer this last type—a *quantitative* scale—but we have ways of dealing with the other two.

2. Originally a *T* score was any standard score other than a *z* score, but usage has gradually changed its meaning to "a standard score in a distribution that has a mean of 50 and a standard deviation of 10."

3. The numbers above the baseline in the figure below indicate the precise percentages that are only approximated in Figure 5-4. The alternative scales below the baseline include the standard and derived scales that are described in the text as well as some that are not. The long caption is quoted from the original Psychological Corporation bulletin:



The normal curve, percentiles, and standard scores. Distributions on many standardized educational and psychological tests approximate the form of the *normal curve* shown at the top of this chart. Below it are some of the systems that have been developed to facilitate the interpretation of scores by converting them into numbers which indicate the examinee's relative status in a group.

The zero (0) at the center of the baseline shows the location of the mean (average) raw score on a test, and the symbol σ (sigma) marks off the scale of raw scores in *standard deviation* units.

The cumulative percentages are the basis of the *percentile equivalent* scale.

Several systems are based on the standard deviation unit. Among these *standard score* scales, the *z* score, the *T* score, and the stanine are general systems which have been applied to a variety of tests. The others are special variants used in connection with tests of the *College Entrance Examination Board*, the World War II *Army General Classification Test*, and the *Wechsler Intelligence Scales*.

Tables of norms, whether in percentile or standard score form, have meaning only with reference to a specified test applied to a specified population. The chart does not permit one to conclude, for instance, that a percentile rank of 84 on one test necessarily is equivalent to a *z* score of +1.0 on another; this is true only when each test yields essentially a normal distribution of scores and when both scales are based on identical or very similar groups of people.

The scales on this chart are discussed in greater detail in *Test Service Bulletin No. 48*, which also includes the chart itself. . . . Copies of this bulletin are available from The Psychological Corporation, 304 East 45th Street, New York, N.Y. 10017. [Psychological Corporation, "Methods of Expressing Test Scores," *Test Service Bulletin No. 48*, September 1954.]

4. In the former case (mental age = 6), the child's IQ is 100. This is obtained by dividing the mental age (MA) by chronological age (CA):

$$\frac{\text{MA}}{\text{CA}} = \frac{6}{6} = 1.00$$

and then multiplying by 100 to get rid of the decimal point. In the latter case, the child's IQ is 133:

$$\frac{\text{MA}}{\text{CA}} = \frac{8}{6} = 1.33$$

The IQ defined above was for many years *the* IQ. Now, however, it is known as the "ratio IQ" to distinguish it from the newer "deviation IQ" (see "Deviation IQs" scale in the figure accompanying note 3). A deviation IQ is a standard score. In a distribution of general mental ability test scores, the obtained mean is converted to 100, because the mean ratio IQ is 100. Then the standard deviation of the raw scores is changed to match that of ratio IQs on the same test. The result is a set of standard scores that look almost the same as scores obtained by the old method. However, the new method is easier to use and has other technical advantages, the explanation of which would require a more extensive digression into psychometrics than would be appropriate here.

CHAPTER 6: CORRELATION

1. Actually, a Pearson *r* coefficient on these data would be somewhat less than 1.00. The relationship is perfect, but *r* is accurate only in rectilinear (straight-

line) regressions, and this one is really curved. (For every height increase of a uniform amount, the corresponding weight increase is systematically greater as you move from left to right in Figure 6-1.) Another coefficient can be used in such cases; look for "eta (η) coefficient" or "correlation ratio" in any text on statistical methods.

2. If you *should* ever need a quick estimate of correlation, rho is the one to use. Consult any standard text on statistical methods about what to do with tied ranks, and within a few minutes you'll be ready to use Formula (6-1).

3. In practice, the errors caused by such a loss of information tend to cancel each other, and when both calculations are made on the same data, rho and r are almost always nearly identical. Current practice is to use rho only when r is too difficult to obtain. The probable reason is that there is no way to get a standard error of rho [see J. P. Guilford and B. Fruchter, *Fundamental Statistics in Psychology and Education*, 5th ed. (New York: McGraw-Hill, 1973), pp. 144–146; see also J. C. Nunnally, *Psychometric Theory* (New York: McGraw-Hill, 1967), p. 25]. The concept of the standard error is developed in Chapter 8.

4. Those assumptions are (1) that each of the two distributions is unimodal and symmetrical and (2) that the line best representing the relation between them is straight rather than curved. (See the diagrams in the subsection on scatterplots in this chapter for examples of such *rectilinear* relationships. Also see note 1, above, concerning curved ones.) Actually, computer simulations [L. L. Havlicek and L. Peterson, "Effect of the Violation of Assumptions upon Significance Levels of the Pearson r," *Psychological Bulletin* 84, no. 2 (1977): 373–377] have shown that violation of the traditional assumptions does not affect r very much. It is therefore said to be a *robust* statistic.

CHAPTER 7: DESCRIPTION TO INFERENCE: A TRANSITION

1. Earlier editions have included the standard explanation in which a degree of freedom is lost whenever you calculate a mean. I have deleted the standard explanation because I seem to have discovered a flaw in it. A mathematical rationale for the standard explanation is nicely stated by Helen M. Walker, "Degrees of Freedom." *Journal of Educational Psychology*, 31(1940): 253–260.

CHAPTER 8: PRECISION OF INFERENCE

1. It is possible conceptually to separate errors of sampling from those that inhere in the process of taking the measurement, such as allowing distracting sounds to enter the examining room. Our concern here is strictly with sampling error.

2. Even if the population is not normal, the distribution of sample means will approach normality as n increases. This is the *central limit theorem* [E. Mansfield, *Basic Statistics with Applications* (New York: W. W. Norton, 1986), pp. 241ff]. Because so much statistical thinking assumes that measures are normally distributed, this theorem opens doors that otherwise would be closed.

3. The number of individuals in an actual population is frequently unknown, and the N of a hypothetical distribution of means is infinite; there is thus no way in which either can be represented accurately in a drawing. But because I believe that the relations among samples, populations, and hypothetical distributions are best understood when presented graphically, I have drawn them anyway. The drawing of the population is larger than that of a sample because the sample is a part of the whole population, but the *amount* of the difference between them can never be known because we can't know the size of the population; my rendition is arbitrary in that respect. In Figure 8-1 the width of the drawings of sample means is smaller than that of the population because means vary less than individual scores do. But the *height* of the distribution of sample means is infinite, so I have arbitrarily rendered its height equal to that of the population.

CHAPTER 9: SIGNIFICANCE OF A DIFFERENCE BETWEEN TWO MEANS

1. This is the formula used when means are uncorrelated. If students were *matched* on some variable (for example, intelligence) related to their level of performance under both control and experimental conditions, a correlational factor would have to be introduced. But that refinement is beyond the scope of this book, and the basic idea of $s_{\bar{X}_e - \bar{X}_c}$ is better conveyed by Formula (9-1) as it stands.

2. To find how many standard errors an obtained difference is from zero, it is necessary to find the size of the standard error. In this instance, there were two such computations:

   First, there was the standard error of the difference for Figures 9-3 and 9-4, assuming that the standard errors of the means of the two distributions were as indicated in Figures 9-1 and 9-2:

$$s_{\bar{X}_c} = 2 \qquad s_{\bar{X}_e} = 2$$

$$s_{\bar{X}_e - \bar{X}_c} = \sqrt{s^2_{\bar{X}_c} + s^2_{\bar{X}_e}}$$

$$= \sqrt{2^2 + 2^2} = \sqrt{4 + 4} = \sqrt{8}$$

$$= 2.83$$

(You will find this calculation also in Box 9-1.)

Second, there was the standard error of the difference for Figures 9-7 and 9-8, assuming that the standard errors of the means of the two distributions were as indicated in Figures 9-5 and 9-6:

$$s_{\bar{X}_c} = 5 \qquad s_{\bar{X}_e} = 5$$

$$s_{\bar{X}_e - \bar{X}_c} = \sqrt{s^2_{\bar{X}_c} + s^2_{\bar{X}_e}}$$

$$= \sqrt{5^2 + 5^2} = \sqrt{25 + 25} = \sqrt{50}$$

$$= 7.07$$

3. In our examples, the no-difference, or no-deviation, hypothesis specifies no deviation from a difference of zero between (or among) treatment groups. By extension, the null hypothesis becomes *any hypothesis concerning the location of a parameter*—a hypothesis of no difference *from that hypothetical location*, whether or not the location is zero.

   A good illustration of a nonzero null hypothesis comes from quality control in industry. If a product is supposed to contain X amount of some chemical, for example, then each random sample is drawn against a null hypothesis of X, and any significant deviation from X requires remedial action. (In this example only one sample is required, and the hypothetical distribution is of means rather than differences between means.)

4. The significance level states the probability that we are making an error when we reject the null hypothesis. That kind of error is known as *Type I*. The probability of making a Type I error is sometimes called "alpha error" or simply "alpha" (α). By stating before the experiment that we will accept only a very low probability (that is, a low probability of rejecting a null hypothesis that is in fact true), we can reduce our Type I errors to a level approaching zero. Unfortunately, however, the *lower* we set *that* level, the *higher* the probability of *accepting* a null hypothesis that is in fact *false*. The latter kind of error is called *Type II*. The following table summarizes those relationships. The ability of a significance test to make the decision represented by the lower left cell of the table—that is, to reject a false null hypothesis—is called its *power*. ("Sensitivity" might have been a more descriptive term for the power to detect a difference.)

		Decision: Reject	Decision: Accept
Null hypothesis	True	Type I error	Correct
	False	Correct	Type II error

5. Many authorities will not accept a one-tail test under any circumstances. Their argument is that we cannot have a good reason for expecting one group to be superior to the other—or rather, that if we did, we wouldn't have to make the statistical test. To put it another way, the reason that we conduct the test at all is precisely that we *don't* know what the outcome will be. There is controversy in every field.

CHAPTER 10: MORE ON THE TESTING OF HYPOTHESES

1. If instead of a simple yes or no, we were to use a 10-point scale extending from "strongly approve" to "strongly disapprove," we could use this formula without alteration; as it is, accuracy demands that we correct for the crudity of our two-point scale. But my purpose in this book is to reveal the basic structures of statistics. I have therefore declined to present to you a formula that disguises the structure on which chi-square is based. If you are interested in *computing* a chi-square, you will need a "correction for continuity" (really a correction for the *dis*continuity imposed by a crude scale of measurement); you will find one in any good text on statistical methods.

2. This latter estimate is sometimes referred to as *residual* or even *error variance* (s^2_E). The term *error variance* may be somewhat confusing, because most of the variance probably stems from factors that in different circumstances might prove to be effective independent variables. They are regarded as errors in the context of a given investigation because they are random with respect to the independent variables in *that* experiment. They are the so-called controlled variables. There are two general strategies for "controlling" a variable in an experiment: (1) to hold it constant over all conditions of the independent variable(s) and (2) to randomize its variance. Both strategies are designed to achieve the same objective: prevention of any systematic (i.e., *not* random) relationships between the controlled variable and any independent variable. There usually are several controlled variables in an experiment.

3. When only two groups are being compared, the analysis of variance is really a t test; to put it another way, t is a special case of analysis of variance that can be used only when there are just two categories of the variable being examined. In such a case, $t = \sqrt{F}$; with whole populations, of course, the same is true of z—that is, $z = \sqrt{F}$.

4. The F ratio for interaction is

$$F = \frac{s^2_I}{s^2_w}$$

where s^2_I is the interaction variance and s^2_w is the within-group variance.

5. In order to maximize simplicity—and hence clarity—in this example, I have assumed the linearity of each of the functions depicted in Figures 10-4 and 10-5. A more refined experimental design might have defined those functions more precisely by including several degrees of stress (instead of just two) and several degrees of experience. If that were done in the case of stress, for example, it might turn out that medium amounts of stress produce more persistence than either extreme amount. An investigation of the effects of a single variable through many changes in its value is sometimes referred to as a *parametric* study.

CHAPTER 11: CORRELATION, CAUSATION, AND EFFECT SIZE

1. There is now a purely statistical method of assigning causes in a *matrix* of correlations. It is called *cross-lagged panel correlation* (*CLPC*) and is based on the time relations in the matrix; it is axiomatic that causes precede effects [D. A. Kenney, "Cross-Lagged Panel Correlation: A Test for Spuriousness," *Psychological Bulletin* 82 (1975): 887–903]. There is no way that such an assignment can be extracted from a single correlation coefficient, however.

2. Even if a variable is really continuous, scores from it are reported in discrete intervals; even an interval of 1 breaks the continuum into segments. (For example, a score of 6 on a test is awarded to all persons whose performances are better than 5.5, but not as good as 6.5.) But *a dimension with a large number of class intervals approximates continuity*, so I shall refer to such data as *continuous*.

3. In the case of a qualitative difference (like gender), one would expect a normal distribution *within* each group rather than the truncated configurations (half of a normal distribution in each category) shown in Tables 11-1B, C, and D.

4. Of course the r must be reliable. Its reliability is assessed by conducting a significance test. The test is similar to that of a difference between means (Chapter 9) except that the null hypothesis is different. Instead of "there is zero difference on the dependent variable between the two groups defined in the independent variable," the null hypothesis here is that "there is zero correlation between the independent and the dependent variables." In addition, instead of dividing the obtained difference between the two groups by the standard error of a difference between means ($s_{\bar{X}_e - \bar{X}_c}$), you divide the obtained r by its standard error $s_{r_{id}}$, where i is the independent variable and d is the dependent. This works for coefficients of $r = .50$ and smaller. For larger r's, the testing of the null hypothesis is more complicated because the sampling distribution is distorted by the upper limit of 1.00 that applies to any coefficient of correlation. For a discussion of the problem and a recommended solution, see J. P. Guilford and B. Fruchter, *Fundamental Statistics in Psychology and Education*, 5th ed. (New York: McGraw-Hill, 1973), pp. 144–146.

Solutions to Sample Applications

CHAPTER 3: MEASURES OF CENTRAL TENDENCY

EDUCATION

A good choice. The mean of 200 reading scores will give you a reliable average of that ability in your trainees. It would also be useful later on to meet the need to estimate reading level in future 12th-graders, should the program be made permanent, thus turning your population into a sample of *all* district 12th-graders, current and future.

Possible misinterpretations. Misinterpretations include finding a single central tendency when there are really two. It is possible that students enrolled in the semiskilled trades (such as food service, construction, sewing, and horticulture) and students in the skilled trades (such as computer programming, electronics, and radio and television) have very different reading abilities. If so, the mean reading level of the total sample would fall between the two subgroups. It therefore would not represent either of them, and materials purchased on the basis of the mean would be at too high a level for one group of students and too low a level for the other. If the two groups were of approximately equal size, the resulting *bimodal* distribution could alert the investigator to that problem, but if one of the groups was much smaller than the other, the total distribution might be *unimodal* and thus give no clue to the presence of the second group. If that presence resulted in a *skewed* distribution and that distribution was treated as a single population, the appropriate measure of control tendency would be the *median*, which gives less weight to atypical scores.

5. In order to maximize simplicity—and hence clarity—in this example, I have assumed the linearity of each of the functions depicted in Figures 10-4 and 10-5. A more refined experimental design might have defined those functions more precisely by including several degrees of stress (instead of just two) and several degrees of experience. If that were done in the case of stress, for example, it might turn out that medium amounts of stress produce more persistence than either extreme amount. An investigation of the effects of a single variable through many changes in its value is sometimes referred to as a *parametric* study.

CHAPTER 11: CORRELATION, CAUSATION, AND EFFECT SIZE

1. There is now a purely statistical method of assigning causes in a *matrix* of correlations. It is called *cross-lagged panel correlation* (*CLPC*) and is based on the time relations in the matrix; it is axiomatic that causes precede effects [D. A. Kenney, "Cross-Lagged Panel Correlation: A Test for Spuriousness," *Psychological Bulletin* 82 (1975): 887–903]. There is no way that such an assignment can be extracted from a single correlation coefficient, however.

2. Even if a variable is really continuous, scores from it are reported in discrete intervals; even an interval of 1 breaks the continuum into segments. (For example, a score of 6 on a test is awarded to all persons whose performances are better than 5.5, but not as good as 6.5.) But *a dimension with a large number of class intervals approximates continuity*, so I shall refer to such data as *continuous*.

3. In the case of a qualitative difference (like gender), one would expect a normal distribution *within* each group rather than the truncated configurations (half of a normal distribution in each category) shown in Tables 11-1B, C, and D.

4. Of course the r must be reliable. Its reliability is assessed by conducting a significance test. The test is similar to that of a difference between means (Chapter 9) except that the null hypothesis is different. Instead of "there is zero difference on the dependent variable between the two groups defined in the independent variable," the null hypothesis here is that "there is zero correlation between the independent and the dependent variables." In addition, instead of dividing the obtained difference between the two groups by the standard error of a difference between means ($s_{\overline{X}_e - \overline{X}_c}$), you divide the obtained r by its standard error $s_{r_{i,d}}$, where i is the independent variable and d is the dependent. This works for coefficients of $r = .50$ and smaller. For larger r's, the testing of the null hypothesis is more complicated because the sampling distribution is distorted by the upper limit of 1.00 that applies to any coefficient of correlation. For a discussion of the problem and a recommended solution, see J. P. Guilford and B. Fruchter, *Fundamental Statistics in Psychology and Education*, 5th ed. (New York: McGraw-Hill, 1973), pp. 144–146.

Solutions to Sample Applications

CHAPTER 3: MEASURES OF CENTRAL TENDENCY

EDUCATION

A good choice. The mean of 200 reading scores will give you a reliable average of that ability in your trainees. It would also be useful later on to meet the need to estimate reading level in future 12th-graders, should the program be made permanent, thus turning your population into a sample of *all* district 12th-graders, current and future.

Possible misinterpretations. Misinterpretations include finding a single central tendency when there are really two. It is possible that students enrolled in the semiskilled trades (such as food service, construction, sewing, and horticulture) and students in the skilled trades (such as computer programming, electronics, and radio and television) have very different reading abilities. If so, the mean reading level of the total sample would fall between the two subgroups. It therefore would not represent either of them, and materials purchased on the basis of the mean would be at too high a level for one group of students and too low a level for the other. If the two groups were of approximately equal size, the resulting *bimodal* distribution could alert the investigator to that problem, but if one of the groups was much smaller than the other, the total distribution might be *unimodal* and thus give no clue to the presence of the second group. If that presence resulted in a *skewed* distribution and that distribution was treated as a single population, the appropriate measure of control tendency would be the *median*, which gives less weight to atypical scores.

POLITICAL SCIENCE

A good choice. The mean can give you an estimate of the balance point of the distribution of European military spending. Of all the measures of central tendency in your sample, the mean is the best estimate of the population average.

Possible misinterpretations. Despite the general virtues of the mean as a measure of central tendency, it is sensitive to any single deviant case. For example, if a 1980 sample had included the Soviet Union, the expenditure of one country would have far exceeded that of any of the others and would therefore have skewed the distribution. The median would have been a more appropriate centrality measure, since the median divides a distribution in half. The Soviet Union would have had no greater effect than any other nation on the location of the median.

PSYCHOLOGY

A good choice. The mean is the only measure of central tendency that includes information from all of the children you have tested. It will yield the most reliable estimate of what the typical newborn is capable of doing (at least in your particular sample). If, for example, the mean number of points assigned were 50, then 50 would be a standard, or norm, for judging the normality of a particular infant, and a child scoring far below 50 could be viewed as abnormally delayed.

Possible misinterpretations. The neonates selected from the hospitals in your city may be different from neonates in general. For example, if your sample were composed mainly of African-American infants, the mean might be too high to be representative of infants in general because African-American infants tend to develop more rapidly than infants of other races.

SOCIAL WORK

A good choice. The mean can give you a reliable estimate of the average number of hours of service that families receive from this agency.

Possible misinterpretations. A single family that required a particularly intensive treatment program with many hours of service would raise the mean substantially unless n was very large. Even with a large n, a few really extreme cases can distort the mean as a measure of central tendency. In general, the mean should be used only when the distribution is nearly normal.

SOCIOLOGY

A good choice. The mean can give you a reliable estimate of the average income of residents in the city.

Possible misinterpretations. The mean is sensitive to extreme scores. If the distribution included such scores and if they did not happen to be arranged so that they balanced each other, the mean would give you misleading information about the bulk of the population.

CHAPTER 4: MEASURES OF VARIABILITY

EDUCATION

A good choice. The standard deviation can tell you whether the variability of the ratings for the two programs is the same. One program may be fairly uniformly rated by the great majority of teachers; that is, teachers may not differ very greatly in their assessments of the usefulness and practicality of the techniques presented in the program. The other training program may be rated very high on practicality and usefulness by some teachers and very low by others, resulting in a high standard deviation. The training program with the relatively large standard deviation may be less desirable because teachers may have widely varying perceptions about the usefulness and practicality of the techniques, leading to a less uniform adoption of the techniques than if the first program were used.

Possible misinterpretations. Misinterpretations include attributing the higher variability of the second in-service training program entirely to problems inherent in the program itself. It may be due partly or even entirely to other factors—differences in the time of day the teachers were trained, for example, or differences in administrative support among schools.

POLITICAL SCIENCE

A good choice. The standard deviation can tell you how much the incidence of military coups differs among Latin American countries. If each country experienced approximately the same number of coups, there would be almost no differences and the average of their deviations from the mean would be nearly zero. But the more differences there are, and the larger they are, the larger that average would be. The standard deviation is a kind of average of the deviations of individual countries from their mean.

Possible misinterpretations. As in the case of the mean, a few extreme scores within a frequency distribution can give misleading results when the standard deviation is calculated. Therefore, if a very few Latin American countries experienced a very high number of coups, then just as the median would be a better measure of central tendency than the mean, so the interquartile range would be a better measure of variation than the standard deviation.

PSYCHOLOGY

A good choice. The standard deviation provides a measure of consistency (agreement)—or rather, a measure of inconsistency (disagreement)—among the observers. If, for instance, some observers reported as few as 2 aggressive acts per

day and others reported as many as 20, the standard deviation would be large, and the reliability of the ratings should be questioned.

Possible misinterpretations. Given a high level of variability among the five observers in our example, it would be tempting to blame one or several raters (e.g., the least experienced) for the apparent discrepancies and inaccuracies. A more likely explanation would be that the instructions to the observers failed to clarify what constitutes an aggressive act. If so, the instruments would need to be revised.

SOCIAL WORK

A good choice. The standard deviation can tell you the amount of variability in grant awards to member agencies. A high standard deviation would indicate considerable differences in the amounts of money received by the agencies. A low standard deviation would indicate that agencies were funded at approximately the same level.

Possible misinterpretations. Again, as with the mean, a few extreme values can result in a higher standard deviation; thus, although only one or two agencies might be receiving much more or much less money than all the others, a large standard deviation would make it appear that there was much variation throughout the distribution. In general, the standard deviation should be used only when the distribution is approximately normal.

SOCIOLOGY

A good choice. The standard deviation can give an estimate of that span of family sizes which includes 68 percent of all families and that which includes 96 percent of all families. Once you know the standard deviation, the regions above and below the mean can be examined separately by referring to normal distribution tables.

Possible misinterpretations. All these conclusions are based on the assumption of a normal distribution of family sizes. If that distribution was not approximately normal, you would have to find other ways of reporting your conclusions. (See the section on "The Interquartile Range.")

CHAPTER 5: INTERPRETING INDIVIDUAL MEASURES

EDUCATION

A good choice. If you transform raw scores for all the tests of the battery into standard z scores based on performances by students of the same age and grade, it will be easier for you to compare the student's performance on the various tests. For example, a z score of 0 would indicate that the student is average in that ability in comparison with his peers. A z score of $+1$ would indicate that the student is 1 standard deviation above the mean (at the 84th percentile) in that ability in comparison with fourth-grade students nationwide. A z score of -2 would indicate that he is 2 standard deviations below the mean (at the 2nd percentile) of his grade. Such comparisons can help the teacher to determine in which areas the student is having difficulty so that some specific assignments can be designed to strengthen those abilities.

Possible misinterpretations. Misinterpretations include regarding ability scores as direct indices of hereditary potential. For instance, if a student is a member of a lower-class ethnic minority, his relatively low abilities may be due to environmental rather than hereditary limitations.

POLITICAL SCIENCE

A good choice. The z score can tell you where each score fits on a scale common to all. The z scale is based on each distribution's mean and standard deviation. In this case, each of the various civil strife scores is expressed in terms of the standard deviation of its own distribution and measured from its own mean. Once all the scores have been converted to z scores, every distribution has the same mean (0) and standard deviation (1). *Now* it makes sense to combine them.

Possible misinterpretations. Don't regard standard scores as absolute in the sense that physical measures (e.g., length) are absolute. Each score tells only where a given measure lies within one distribution of measures of the same attribute. In a different distribution, the same measure might yield a different standard score.

PSYCHOLOGY

A good choice. By transforming each raw score to a new scale with a mean of 0 and a standard deviation of 1, you can ascertain whether June is developing normally in all three domains. If she earns z scores of -1 on all three tests, you can infer generally delayed development. If all three z scores are $+1$, you can infer advanced development. If she obtains z scores of, say, $+2$ on intelligence, -1 on social development, and -2 on psychomotor development, you can infer fragmented development.

Possible misinterpretations. One possible misinterpretation is the presumption that obtained measurements are stable across time. Actually, the behavior of infants and young children tends to vary considerably from one observation to the next.

SOCIAL WORK

A good choice. Someone in the personnel office can place each raw score into a distribution with a mean of 0 and a standard deviation of 1. Then all scores are on a standard scale, thus making it possible for you to compare them.

Possible misinterpretations. If the various tests have been standardized on different populations, comparisons will be risky. For example, scoring high on client assessment in a population of beginners may not be more laudable than scoring low on client treatment in a population of highly trained professionals.

SOCIOLOGY

A good choice. The z score can tell you how this professor compares with other professors in terms of standard deviation units above or below the mean.

Possible misinterpretations. Professors teaching the courses for which you are eligible may differ in some significant way from the faculty as a whole (e.g., they may have been selected for their ability to relate to lower-division students). If so, the professor's z score may well lead you to make a bad decision at registration time. For example, his z score on authoritarianism could be low in the general faculty but high among those faculty members who are teaching the course in question.

CHAPTER 6: CORRELATION

EDUCATION

A good choice. The Pearson r can tell you whether there is a relationship between self-concept and social responsibility in these children. It will also tell you whether the relationship between self-concept and social responsibility is positive or negative. The size of the coefficient, whether positive or negative, will indicate the strength of the relationship between the two measures.

Possible misinterpretations. From a strong correlation (.70 or .80), it might be erroneously concluded that a student's self-concept causes the student to be socially responsible or irresponsible. (Although the two variables do vary together, this may be due to some third variable, such as previous school or home experiences, which affects both social responsibility and self-concept.) It might also be erroneously concluded that a program effective in enhancing a person's self-concept necessarily increases social responsibility. (If self-concept is caused by social responsibility or if both are caused by some third factor or set of factors, changing self-concept will not affect social responsibility.)

POLITICAL SCIENCE

A good choice. The Pearson r can tell you the magnitude and direction of the relation between the amount of conflict within countries and the amount of foreign conflict they initiate. First, the larger the coefficient, the stronger the association. Second, a positive sign indicates a direct relationship between internal and external conflict, while a negative sign indicates an inverse relationship.

Possible misinterpretations. Normally, a strong positive value of r would be taken as evidence that domestic conflict is associated with foreign policy conflict. Remember, however, that correlations do not necessarily entail causal relations. Also be aware of the assumptions behind r. If these assumptions are violated, then the results may not be valid. The most important of these assumptions are that the relationship between the two variables yield a straight regression line, that the distributions are unimodal, and that they are fairly symmetrical. However, violation of even these assumptions seldom invalidates an r. (See note 4.)

PSYCHOLOGY

A good choice. The Pearson r will indicate the degree of association (.00 to either $+1.00$ or -1.00) between sugar intake and activity rating for the 100 children. The relationship can be positive (high sugar levels associated with excessive activity and low sugar levels associated with low activity) or negative (high sugar levels associated with low activity and low sugar levels associated with high activity).

Possible misinterpretations. A significant positive or negative relationship does not imply a cause-and-effect relationship. A positive Pearson r does not necessarily mean that high sugar levels cause hyperactivity. Hyperactivity may stem from a fundamental impulsivity in children. That is, the child's inability to control impulses may lead to both excessive activity *and* excessive ingestion of sugar. Thus a third factor (impulsivity) may be responsible for an observed relationship between blood sugar and hyperactivity.

SOCIAL WORK

A good choice. The Pearson r provides a measure of association that reveals both the strength and the direction of the relationship between the two variables. Your hypothesis is that women of high self-esteem are less dependent on welfare than are women of low self-esteem. A high negative correlation would confirm your hypothesis.

Possible misinterpretations. Even if your hypothesis is confirmed and you know self-esteem and requesting welfare are related, you don't know just *how* they are related. Two possibilities are (1) that self-esteem renders a woman more determined to make herself economically self-sufficient and (2) that an inability to make herself self-sufficient damages a woman's self-esteem.

SOCIOLOGY

A good choice. The Pearson r can tell you whether there is a relationship between the conservativism of academic disciplines and the authoritarianism of their professors.

Possible misinterpretations. Misinterpretations include the notion that the first variable causes the second, or vice versa. In fact, both variables *may* be products of an unnamed external factor or set of factors.

CHAPTER 8: PRECISION OF INFERENCE

EDUCATION

A good choice. The standard error of the mean is used to establish a confidence interval within which the population mean for each grade level lies. The band of scores described by the confidence interval provides a better index than the obtained mean itself of the achievement level of each local grade because it takes into account the errors that inevitably occur whenever measurements are made. You can give the probability that the mean really lies between two particular values.

Possible misinterpretations. If the band provided by the confidence interval around a sample mean is higher or lower than the national mean, don't assume that the difference is due solely to the school program. Other factors such as the influences of the home and the community must also be considered.

POLITICAL SCIENCE

A good choice. Precisely these limits are described by the confidence interval. There is a confidence interval for any desired probability (confidence level). For example, say you want to specify the interval within which you can place the true mean at a confidence level of 68 percent (i.e., the interval within which the probability is 68 percent that the true mean lies.) Your obtained mean is 50. If the standard error ($s_{\bar{X}}$) is 10, the interval is $2 \times 10 = 20$ points wide (see pages 94–96). At the 68 percent level of confidence, then, the confidence interval would extend from 40 to 60 points.

Possible misinterpretations. You might think it equally probable for the true mean to be at any point within the confidence interval. Actually, the probability is higher in the middle of the interval than anywhere else. Your obtained mean is, after all, your best estimate of the true mean.

PSYCHOLOGY

A good choice. The confidence interval would allow you to estimate the limits within which the boy's true inkblot test score resides. It would also provide a statement of the probability that his true score is indeed within those limits. If the limits were far apart or the probability low, you would be cautious in interpreting the boy's score.

Incidentally, the manual might include, either in place of the confidence interval or in addition to it, a correlation coefficient (r_{tt}). That statistic represents another way of looking at reliability (see Chapter 6).

Possible misinterpretations. You may be tempted to think of error as something that lowers a score. But measurement error (unreliability) can also raise a score. The confidence interval therefore extends both below and above the obtained score. (The obtained score is in the middle of the interval.)

SOCIAL WORK

A good choice. The confidence interval provides a range of scores within which the family's true FLIP rating probably resides. There are two ways to approach the question "Where is the true score?"

First, you can set limits above and below the obtained score and then find the probability that the true score is between those limits. Second, you can set an acceptable probability and then set the limits that match it. In either case, the interval between the limits is called the confidence interval, and the probability is the level of confidence in this particular case.

Possible misinterpretations. In tests of significance (see Chapter 9) the probability cited is that of an occurrence *outside* prescribed limits. Here, you report the probability that the true mean is *inside* those limits.

SOCIOLOGY

A good choice. The standard error of the mean can give you a band of scores (the confidence interval) that probably contains the true mean. It can also specify the level of that probability (the level of confidence). This information is far from "meaningless."

Possible misinterpretations. The true mean does not *have* to fall inside the intervals you have computed—thus, you shouldn't accept a confidence interval without mentioning the corresponding level of confidence.

CHAPTER 9: SIGNIFICANCE OF A DIFFERENCE BETWEEN TWO MEANS

EDUCATION

A good choice. The t ratio can tell you whether at the end of the semester the sample of students who were in the group-counseling program differ significantly in teacher-rated disruptiveness from the sample that was not in the program. Thus the staff could determine (with a given probability) whether students in the group-counseling program were less disruptive at the end of the semester than the students not in the program.

Possible misinterpretations. You can't necessarily attribute differences between the groups solely to the counseling program. Other factors may have influenced the results:

1. Even though the groups were randomly assigned, there may have been initial differences in disruptiveness at the beginning of the semester.

2. The teachers knew which students were in the treatment program, and that may have influenced them to react to those students differently or to rate them differently even though those students actually were not different from students who were not in the counseling program.

3. Just the added attention of the school counselor or just their inclusion in a new program could have lowered the students' disruptiveness, regardless of the content of the program.

POLITICAL SCIENCE

A good choice. The *t* test can be used to determine whether the observed difference between the means of the two groups is due to chance. The average difference between the two means within repeated pairs of random samples taken from the same population would be zero. If the observed difference is large enough to be statistically significant, then we can reject the hypothesis that the difference is merely a chance difference between two random samples of the same population of precincts. We conclude rather that the precincts receiving the program are truly different from those receiving the old cops-only program.

Possible misinterpretations. Although the *t* test is appropriate for analyzing the difference of means for two small samples from the same or identical populations, its use is subject to three restrictions. First, the observations in the two samples must be independent of each other. Second, the populations must not be skewed in opposite directions. Finally, if the populations do not have equal variances, adjustments are needed in the calculation of *t*.

PSYCHOLOGY

A good choice. The *t* ratio can tell you whether the relaxation treatment or the drug therapy is the more effective as determined by the mean scores of the two groups. It will give you the probability that there really is no difference between them—that they are both random samples from a single population with respect to activity level.

Possible misinterpretations. Although the results may show a significant difference between the two groups, locating the *cause* of that difference may be difficult. Suppose that the drug-therapy group appears to benefit more from treatment. That could be the result of the drug, but unless you equalized the age of the two groups, age differences could be the critical factor. Without controlling other possible causes, you cannot know whether the cause is the drug or some other factor such as the ages of the children, their intelligence, their social background, or some characteristic of their drug therapists.

SOCIAL WORK

A good choice. The *t* ratio can be used to determine whether there is a significant difference between two groups. The difference is significant if you can reject the null hypothesis that the two groups are from the same population with respect to health. You can take your *t* ratio to a table that will give you the probability that the two groups *are* from the same population. If that probability is extremely small—say, .01—you may assert with considerable confidence that the experimental program has been effective.

Possible misinterpretations. Your confidence in the program will have been misplaced if some outside factor influences one group but not the other. For example, if the experimental group rides a special bus to the center, eats its meals together, or does something else together as a by-product of the program, it could be all that "togetherness" rather than the program itself that improves the seniors' health.

SOCIOLOGY

A good choice. The *t* ratio, entered into appropriate tables, can tell you how likely it is that the obtained difference occurs by chance.

Possible misinterpretations. As in correlation, a relationship does not by itself justify an inference of causality. If most of the Catholics in your state belong to a different social class than most non-Catholics, it might not be legitimate to infer that the difference in religious belief is the cause of the difference in family size. It might turn out that with social class held constant, Catholic families are no larger than non-Catholic ones.

CHAPTER 10: MORE ON THE TESTING OF HYPOTHESES

EDUCATION

1. A good choice. You need to know whether the frequency of high school graduation is higher for students who have been through the special program than for those who have not. Chi-square can tell you that.

Possible misinterpretations. It is important that all of the students in both groups come from the at-risk category. A group taken from the entire student body would have an initial advantage over a group identified as being at risk. Also, you cannot be certain that the training program is the only cause of the observed difference. Even if the program significantly increases the percentage of at-risk students who graduate, its success may be due more to the interest of the staff in a new program than to any attribute of the program per se.

2. A good choice. One-way analysis of variance can tell you whether the differences among the three means on the social-problem-solving test is greater than would be expected by chance. If there is such a difference (and it is in favor of the clients of the trained counselors), then you may conclude that the students who went through the program are better able to solve the problems posed by the test than are students in the other two groups.

Possible misinterpretations. Don't assume that if there is a significant F ratio, each group mean differs significantly from each of the others. For the F test to be significant, it is necessary only that two of the means differ significantly. You may also erroneously conclude that the test result must be attributable to the program. Even though the students were randomly assigned, there could have been differences among the groups before the programs began. The students could have learned which groups they were in after the study began and been affected by that knowledge. Similarly, each counselor's knowledge of his or her own place in the program might affect student behavior in ways not specified by the training the counselor received.

3. *A good choice.* Two-way analysis of variance of the abstract-reasoning-ability scores from the beginning of the year can tell you whether there is any difference among the mean test scores of the students in the four groups. The same analysis can tell you whether there is a significant difference in reasoning ability between the sixth and seventh grades prior to your intervention. The analysis can also tell you whether, before the programs begin, there are significant interactions between grade level and abstract reasoning ability in the four treatment programs.

At the end of the year, all of those significance tests can be administered again. In addition, and probably more important, if the differences among the means of the four groups were not significant before the training, a significant F test of program differences afterward suggests differential effects of the four treatment conditions. If follow-up t tests show significant superiority of one or more of the experimental programs over the control, you have evidence that your programming efforts were not in vain.

Possible misinterpretations. Attributing significant differences at the end of the year to the effectiveness of one or more programs would be an error if the differences were actually there in the beginning.

Political Science

1. *A good choice.* Chi-square can tell you the probability that any deviation of the observed frequencies from a stipulated expected frequency is due to chance. (The null hypothesis here is that equal proportions of Republicans and Democrats support a cut in the capital gains tax.) Chi-square compares the observed frequencies in each cell of the contingency table with what would be expected in these cells if the two variables—party affiliation and support for the capital gains tax cut—were independent.

Possible misinterpretations. The most common misuse of chi-square is the violation of certain key assumptions. The most important of these is that the data represent a random sample of independent observations. In this case, you must be sure that you sample *all* Republicans, not just some subgroup dedicated to the passage of a tax cut, and that your sample of Democrats is similarly random.

2. *A good choice.* One-way analysis of variance can tell you whether observed differences (in coup frequency) among types of regime are likely to have occurred by chance. The analysis of variance is an extension of the difference-of-means t test that we examined earlier. Had there been only two types of regime legitimacy, a t test would have sufficed. With more than two types, however, you must compute an F ratio. If the population variance in coup frequency estimated from differences between the countries is not significantly greater than the variance as estimated from within the groups, then you cannot reject the null hypothesis. That is, you must conclude that the differences you have observed are merely random variations—that the three types of country are all one with respect to frequency of coups.

Possible misinterpretations. When the requirements of the one-way analysis of variance are met, the F ratio allows you to determine whether the observed difference among the group means is large enough to be statistically significant. However, since in this case there are more than two categories within an independent variable, the F ratio does not tell you which of the categories differ. Further tests are necessary to accomplish that.

3. *A good choice.* Two-way analysis of variance can tell you whether there is a significant "form of government" main effect and whether there is a significant "type of leadership" main effect. It can also tell you whether there are any interaction effects. For example, it might turn out that totalitarian governments are more aggressive than other forms only when their leadership is of the unitary type.

Possible misinterpretations. The two-way analysis of variance is subject to the same kinds of limitations as the one-way analysis of variance. Specifically, the F test does not tell you precisely where among the nine categories the significant differences lie. Again, further tests can be made.

Psychology

1. *A good choice.* Chi-square can tell you whether the number of children who pass the test after treatment is greater or smaller than the number that might be expected to pass without treatment. If previous research indicates that 45 percent of untreated agoraphobics recover spontaneously, this figure might be used as the expected value (the null hypothesis).

Possible misinterpretations. In this design, uncontrolled factors could account for significant findings. For example, just coming to the clinic may be sufficient to induce change, or children who come to the clinic may not be a random sample of agoraphobic children generally.

2. *A good choice.* One-way analysis of variance can tell you whether there are significant differences among the three treatments in the children's self-ratings of experienced pain.

Possible misinterpretations. You might infer from the *F* ratio that every mean differs from every other, but a significant *F* ratio tells you only that there is a difference *somewhere* between or among treatment means. You now have to scrutinize your data to ascertain where any differences are.

3. *A good choice.* A two-way analysis of variance allows you to test: (1) whether any of the four group means (introvert–individual, introvert–group, extrovert–individual, and extrovert–group) is significantly different from any other; (2) whether group therapy or individual therapy is the more effective, regardless of introversion–extroversion tendencies; (3) whether introversion–extroversion tendencies are associated with better outcome, regardless of the type of therapy; and (4) whether there is an interaction between introversion–extroversion and therapeutic modality—that is, whether individual therapy works best with introverts and group therapy with extroverts, as you had predicted.

Possible misinterpretations. In this particular example, there are only two categories of each variable (introvert–extrovert and individual–group). Whenever that is true, the *F* ratio for each variable identifies precisely the source of any difference that we find. For example, if the *F* for the introvert–extrovert variable turned out to be significant, we would know that the difference is between the group we have designated "introvert" and the one that we call "extrovert."

When there are more than two categories of a variable, however, the *F* test does *not* identify the source precisely. If, for example, we had divided our introvert–extrovert variable into three categories ("introvert," "ambivert," "extrovert") instead of two, a significant *F* would mean only that there was a difference somewhere within that variable. It would not tell us whether that difference was (1) between introverts and ambiverts, (2) between introverts and extroverts, or (3) between ambiverts and extroverts. Further tests would be necessary for more precise identification.

SOCIAL WORK

1. *A good choice.* Chi-square can tell you whether the frequency of recidivism is significantly lower in the group of youth appearing before the board than in the group who are sent to court. Chi-square is an index of deviation from the frequencies you would expect if the new board's procedures were not any more (or less) effective than the courts'. Indeed, chi-square *is* that deviation, expressed as a proportion of the expected frequencies.

Possible misinterpretations. Chi-square deals with frequency counts only. Every score is therefore required to be either 0 or 1, and information about the magnitude of offenses is lost. It is possible that the relatively few offenders who break the law in spite of the board's rehabilitation efforts are guilty of offenses more serious than those committed in larger numbers by the control group.

2. *A good choice.* One-way analysis of variance can tell you whether there is a significant difference anywhere among the four groups.

Possible misinterpretations. A significant *F* does *not* mean that every group is different from every other with respect to hyperactivity. It tells you only that there is a difference *somewhere* among the groups. If the *F* test does reveal a significant difference, you can then follow up with other tests especially designed for use after the *F* test. Those tests will locate the differences.

3. *A good choice.* Two-way analysis of variance can tell you whether there are significant differences among the four group means, that is, (1) children–foster family, (2) youth–foster family, (3) children–group home, and (4) youth–group home. Two-way analysis of variance also identifies the location of any effects—that is, whether adjustment is affected by age, type of foster care, or their interaction.

Possible misinterpretations. A 2×2 factorial design yields an *F* ratio for each main effect and the interaction, so it is not subject to the error described earlier for one-way analysis of variance with six categories of the independent variable. The location of any difference is known from the first computation.

SOCIOLOGY

1. *A good choice.* Chi-square can tell you whether the numbers in the cells of your 2×2 table (early versus late $\times$ approval versus disapproval) could have occurred on the basis of chance alone. The other possibility is that these numbers were influenced by something other than chance; in this case, the belief about when human life begins is a good candidate.

Possible misinterpretations. Possible misinterpretations include regarding the chi-square as a general, though probabilistic, answer to the question implied by your hypothesis. The question is "Does a person's belief about when human life begins affect his or her attitude toward abortion?" You might get a different answer if, for example, "early" and "late" were defined in relation to nine months instead of 90 days. Or your respondents' attitudes toward abortion might look different if question 2 were phrased ". . . under any circumstances" instead of ". . . on demand." Such wording might also change the answer to the general question.

2. *A good choice.* If "family size" is treated as a *score*—a quantitative attribute of families—one-way analysis of variance will give you the probability that all of the obtained differences occurred by chance. If that probability is very small, at least one of the differences is statistically significant.

Possible misinterpretations. A significant *F* does not mean that each sample is significantly different from every other. The *F* test indicates only that there is at least one significant difference among those identified. If the *F* test is positive, then

each mean must be compared with every other mean, and each difference must be evaluated separately from the others.

3. A good choice. Two-way analysis of variance is an appropriate method. The summary table of F ratios will tell you whether there is a main effect of church attendance, a main effect of education, and/or an interaction between the two.

Possible misinterpretations. In this particular example, there are only two categories of each variable (church attendance–nonattendance and high education–low education). Whenever that is true, the F ratio for each variable identifies precisely the source of any difference that we find. For example, if the F for the attendance–nonattendance variable turned out to be significant, we would know that the difference was between the group we have designated "church attendance" and the one we call "nonattendance."

When there are more than two categories of a variable, however, the F test does *not* identify the source precisely. If, for example, we had divided our attendance–nonattendance variable into three categories ("high," "medium," and "low" attendance) instead of two, a significant F would mean only that there was a difference somewhere within that variable. It would not tell us whether that difference was (1) between "high" and "medium," (2) between "high" and "low," or (3) between "medium" and "low" attendance. Further tests would be necessary for more precise identification.

CHAPTER 11: CORRELATION, CAUSATION, AND EFFECT SIZE

Some possible alternatives stated in general terms. When Variable X and Variable Y are substantially correlated, maybe

X is causing Y; or perhaps

Y causes X; or possibly

both are caused by a third variable or set of variables; or even

each is but a single aspect of a multifaceted but indivisible whole.

An illustrative application of those alternatives. For example, in the application to education, maybe

worthy self-concept is causing social responsibility; or perhaps

social responsibility causes worthy self-concept; or possibly

both are caused by a third variable or set of variables—maybe by midlevel socioeconomic status, with all the life experiences that attend that status; or even

worthy self-concept and social responsibility are but two of many aspects of a particular kind of personality.

A caution born of experience. Take your choice among these alternatives, but do not invest much confidence in any of them until it has been tested by some means other than correlation. (See, for example, the discussion of correlational versus experimental studies on pages 141–142.)

Index

In this index, where "ff" follows a page number, references to the listed topic can be found not only on that page but on subsequent pages within the same section. Each word or phrase signifies a topic but is not necessarily present in every place where that topic is discussed.