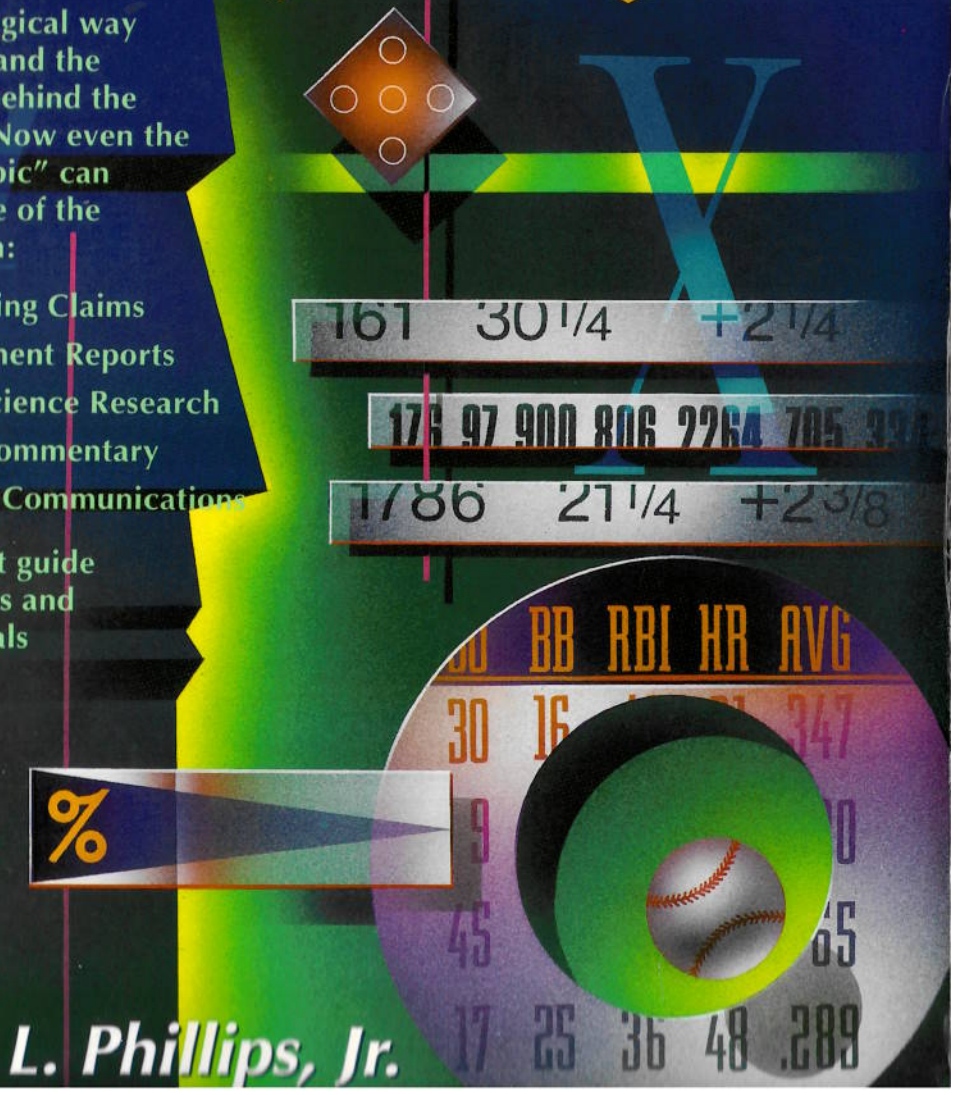
Fifth Edition

HOW TO THINK ABOUT STATISTICS

An easy, logical way to understand the concepts behind the numbers. Now even the "mathphobic" can make sense of the statistics in:

- Advertising Claims
- Government Reports
- Social Science Research
- Media Commentary
- Business Communications

The perfect guide for students and professionals



John L. Phillips, Jr.

Contents

Preface

While I was teaching a class in educational psychology, it occurred to me that it was my students' lack of background in statistics that was blocking their comprehension of some of the concepts I was presenting to them. There was no room in their curriculum for a course in statistics, so I decided to devote two weeks of my course to the concepts that my students needed. I wrote a small book to help them master those concepts.

When *Statistical Thinking* was published, other people began using it, and I discovered that the concepts needed by my students were the same ones that were needed by students in many courses other than mine, and not only in my discipline but in the social/behavioral sciences generally and in various professional curricula—business, education, and social work, for example—that are related to those sciences.

The culture of any industrialized society is suffused with quantitative information. Some quantitative messages are simple and direct; others involve a relatively complicated process of inference. Knowing how to think statistically makes possible the comprehension of both kinds of messages.

College Use

One of two ways in which students can use this new book is as a supplementary text in a course that demands some statistical thinking but does not focus on statistics. The other use is as a self-teaching preparation for a course that does focus on statistics. It has been my observation, and that of my colleagues, that it is possible for a student to complete such a course without ever really *thinking* about statistics. Many students learn to do the required calculations but have only the foggiest

conception of what the calculations mean. Although it is true that statistical methods courses discuss the logic of statistics, that is not their focus. This book emphasizes the logical structure of statistical thinking and deemphasizes techniques of data manipulation. If you are planning to enroll in a statistics course, you should read this book as soon as possible after enrolling in the course, before you begin reading the course text. Different books use different symbols to refer to the same concepts, but that will present no difficulty once you have mastered the concepts here. (It might confuse you while you are *learning* them.) I recommend that you finish this book before you begin another, but if that is not feasible, do it as soon as possible after enrolling in the course. In either event you will find that course both easier and more rewarding as a result of this relatively small extra effort at the beginning.

Everyday Use

If you are engaged in business or a profession, you probably encounter quantitative information frequently in your work. If you have no training in statistics and lack the time, inclination, or opportunity to take a course, and if your need is for the consumption (as distinguished from production) of statistical information, then *How to Think about Statistics* can provide you with the necessary background.

But it is not only business and professional people who have access to statistical information and the motivation to interpret it intelligently. If you are a consumer in a competitive economy, the interpretations you make of statistical information will affect your economic well-being. As a citizen, your interpretations will help determine your support of candidates and your position on important political and economic issues, regardless of the kind of work you do.

Imagine that one such issue is currently being discussed by your state legislature: To compete with private industry for competent employees, should the state raise the salaries of its employees? The Penny Pinchers faction in the legislature cites an average salary of state employees that is impressively high already, but the Spectacular Spenders faction says the figures are misleading. How could that be? The two factions agree on the correctness of the basic data from which the average was calculated, and an average is an average, isn't it? Not really, and different types of average are appropriate in different kinds of situations; Chapter 3 will show you why.

Or consider this advertisement for a particular make of automobile: "A survey of owner satisfaction reveals that 31 percent more owners of Cadmobiles than of any other car in its class say they expect to choose the same brand the next time they buy a car." The survey actually was done, and it did indeed turn out as reported. Could its implication nevertheless be deceptive? It could unless you can think statistically. After you have read Chapter 8, you will be able to show how this advertiser could deceive persons who don't know how to think about statistics.

In these examples, the data themselves were solid enough; it was the implications drawn from them that were deceptive. This book is mainly about extracting

implications from data. There is, however, another source of error in quantitative information. The data themselves can be faulty. There is a virtually unlimited set of possibilities here, but one can serve to illustrate the genre: Conclusions based upon anecdotal data should be approached cautiously no matter what statistical operations have been performed on them. (Chapters 6 and 11 address problems of validity.) Some anecdotal data are less trustworthy than others: The exploits of golfers and fishermen, for example, are suspect when reported by the persons involved, and young children's reports of events in the real world are likely to be liberally laced with fantasy. Statistical thinking cannot protect you from *all* possible mistaken conclusions; as computer people say, "garbage in, garbage out."

On the other hand, if you want to draw useful implications from data, many pitfalls remain even when the data themselves are sound. It is in surmounting those pitfalls that statistical thinking can be exceedingly helpful.

SAMPLE CALCULATIONS

I have stressed the importance of focusing on logic rather than the manipulation of numbers; for most readers manipulations are best reserved for a course in statistical methods. For others, however, it is reassuring to follow each process through numerically. I have, therefore, introduced several sample calculations. Each calculation and a parallel verbal account is sequestered in a box near the corresponding discussion in the text. It will be easy for you either to follow the calculation or not, as you please.

SAMPLE APPLICATIONS

The logic of statistics is fascinating in itself, but most people want to *use* that logic to solve problems, to follow solutions proposed by others, and if need be to criticize those solutions. With that in mind I have included at the ends of most chapters some opportunities to test those skills. (Suggested solutions are at the back of the book.) As you do so, you may also be testing your depth of comprehension with respect to the logical principles presented in the main text. Failure to apply a principle *may* indicate shallow comprehension.

The Revision

The most important change from the previous edition has resulted from my desire to clarify the distinction between description and inference and to demonstrate the relation of that distinction to the sample-population dichotomy. In the service of that objective I decided to mention this distinction early, to present descriptive operations first and the inferential ones later, and to include a special chapter emphasizing the transition from one to the other.

Except for that new chapter, the most noticeable change in the table of contents is the chapter entitled "Correlation." In the last edition it had been moved almost to the end of the book so that it could be contiguous with another chapter that also dealt with correlation. The present emphasis on the description–inference distinction necessitates placing those two chapters on opposite sides of the new transition chapter. Chapter 11, "Correlation, Causations, and Effect Size," involves statistical inference; Chapter 6, "Correlation," does not.

Acknowledgments

I am indebted to Doctors Mark Snow and Steven Thurber for their support and assistance. Our discussions of matters related to the writing of this book have been both stimulating and edifying, and I am deeply grateful. Dr. Snow is Professor and Chairman of the Department of Psychology at Boise State University, and Dr. Thurber also teaches there.

That kind of two- and three-way interaction was not feasible with reviewers engaged by the publisher, but I gleaned many useful ideas from the comments of Catherine Renner, Ph.D.; Barbara E. Reynolds, Cardinal Stritch College; and Kay B. Somers, Moravian College.

John L. Phillips, Jr.
August, 1995

How to Think about Statistics

1

Introduction

You may be planning to study statistics not because you want to but because you have to. If so, I know how you feel. I went through the same experience years ago; if I could have avoided statistics, I probably would have. However, my attitude changed after I began to study it, for I discovered in it a new way of thinking that was truly fascinating.

But your present task may be even more challenging than mine was. You won't have to do the computations that I did, but you are about to acquire within a very short time (and possibly by yourself) the same grasp of the underlying structure of statistics that I acquired in two full semesters under an excellent teacher.

The Task

Your plight and your prospects are well illustrated, I think, by the experience of a student who was asked to evaluate the prototype of this book:

> It was the most difficult book I have ever read. It was foreign to me since I had had no background knowledge of the things talked about. . . . If I was to understand it, I realized I would need to outline the book chapter by chapter. I did so, and to my amazement, it followed a very orderly pattern after all. It really did present what the author had stated he hoped to present. If you understood Chapter 1, you could see the logic of Chapter 2, and so on through each chapter. I feel I learned a great deal about measures in a relatively short period of time.

This quotation contains some important advice on how to use the book. I would only add that even though you may have mastered the ideas preceding the one you are working on at any given moment, you should be prepared to go back to those ideas from time to time and consider their relation to the new one being presented. I have tried, by providing frequent cross-references, to help you do just that. (You may wish to keep a couple of bookmarks handy for that purpose.) When you are finished, you should have in mind a hierarchical structure, with each new idea related to one or more of the ideas that have preceded it.

The conceptualization of such a structure is highly satisfying in itself, but there are many other reasons for making the effort. It is true that an understanding of statistical concepts will not be of critical assistance to you in gathering economic data, in conducting interviews for a political or sociological opinion or attitude study, in uncovering archeological artifacts, or in teaching children. But often people who do economic, political, sociological, anthropological, archeological, or educational studies (to name but a few) report their findings in statistical terms. If you are planning a career in any of the several professions to which those studies are relevant, it is important that you be able to read them with understanding. A continuing awareness of developments in one's field is the mark of the true professional, and this book will help you maintain that awareness.

The Basic Ideas

To understand the meaning of any measurement in the social sciences, you must know at least two things about it. First, you must be able to describe the operations by which it was obtained, and second, you must be able to compare it with other measurements that have been obtained in the same way.

This book is concerned primarily with the second kind of knowledge. Statistical thinking deals with multiple measurements. It analyzes the relation of how many to how much—of frequencies to scores. If the basic element in measurement itself is a *score*, the corresponding concept in statistics is a *distribution* that includes many scores—in short, a *frequency distribution*.

Such a distribution can be described by drawing a picture of it—and indeed in the early stages of your learning about distributions, that is the method I shall use to describe them. The method is cumbersome, however; so ways have been devised to achieve roughly the same result through the use of numbers rather than diagrams. The most important advantage of numbers over diagrams is that they can be manipulated in a way that diagrams cannot.

Imagine that (for reasons known only to your psychoanalyst) you have just had a pile of gravel dumped on your front lawn and that (for similar reasons) you want to describe the result to me over the telephone. To do that successfully, you will have to tell me at least three things about the pile: (1) its general *configuration*—

that is, whether it is shaped like a cone, a pancake, or perhaps your garage roof; (2) its *location*—that is, how far and in what direction it is from some reference point familiar to me; and (3) its *dispersion*—the extent to which it is spread out—for example, if it is a cone, is it a steep-sided one that covers only a small area, or is it a low-profile one that covers most of the lawn?

That pile of pebbles is analogous to a *frequency distribution* of scores, and the same kinds of information are needed to describe either adequately. Concerning configuration, we have certain names for frequency distributions that convey information to anyone who is familiar with them—names such as "normal," "symmetrical," "positively skewed," and "bimodal." Concerning location, the procedure is pretty much the same as it is for a pile of gravel: A dimension is identified and a reference point is chosen, and measurement of distance to the center of the distribution is from that reference point. The result is a measure of *central tendency*. Finally, to communicate information about the amount of dispersion, a new reference point is used—namely, the center of the distribution—and the needed information may consist of the average distance of individuals (like that of the individual pebbles in a pile of gravel) from the central point. That distance is a measure of *variability* (dispersion).

But just describing a distribution is not always enough. Often you will be interested in *two* distributions and in the relationship that exists between them. Consider a single variable, "IQ in the general population," and a few other variables with which it might be paired: family income, some index of health care, a general index of socioeconomic status, race, or place of residence. Other interesting relationships could be investigated among the IQs of various subgroups of the general population: between parents and their children, between identical twins, between fraternal twins, between non-twin siblings, and between pairs of unrelated children. These are just a few of the relationships that come to mind at the moment. Others would occur to you if you were making a study of intelligence and its correlates, and in every case you would need a way of communicating your findings to others; in short, you would need a measure of *correlation*.

Whether or not you wish to relate one set of scores to another, you will surely want to be able to tell what each individual measure means. If you are a teacher who has given my child a test, and I ask you how well he did on it, you might try to put me off by saying his score was "high" or "low" and go on to talk about something else. But if I want to know *how* high, you are in trouble. You may answer that he got 90 percent of the test items right. You think you are off the hook, but I persist: "How *hard* is that test? Ninety percent is very impressive if the items are all difficult, but not if most of the other kids score above 95!"

Our conversation has been concerned entirely with the *interpretation of individual measures*, and I'm sure you'll agree that all of my questions are pertinent. Without answers to them and others like them, one really cannot know the meaning of a score.

On the other hand, you must be careful to avoid overinterpreting measures, whether they be of individuals or of groups. If, for example, you were to weigh a *random sample**·† of fifty 10-year-old boys, could you easily compute a measure of central tendency for the sample that is identical to the central tendency of a population that includes *all* 10-year-old boys? How different might the obtained weight be if it were computed from a different random sample of that same population? Questions like these have to do with *precision of inference*, which is a kind of *reliability*, and we do have ways of dealing with them.

Assuming for the moment that you do know how to deal with questions about precision, consider this one additional question: Upon further analysis of your sample of 10-year-old boys, you discover a marked difference between the weights of boys living in one area of the country and those living in another. You suspect that the difference is due to diet, and you subsequently narrow that hypothesis down to a single vitamin that seems to be more plentiful in one of the areas than in the other. One way to test your hypothesis would be to select two samples of male infants who are living in the area in which the vitamin is less plentiful, introduce the suspected vitamin into the diet of one group, and then after a period of nine years weigh the subjects in both samples again. Let's imagine that there is a difference between the two. Is it large enough that you can be reasonably sure that it did not occur by chance—that a replication of the same study would not turn up a difference of zero, or even a difference in the other direction? To put it another way, how *significant* is the difference you obtained? Again, there are ways of dealing with such questions.

In the chapters that follow, each of the above ideas will be developed further, but always in the manner that you have seen here. The discussion is aimed directly at the underlying *logic* of statistical thinking, with an absolute minimum of arithmetical and algebraic manipulations. You will find the logic similar in many ways to common sense. The main difference is that the logic presented here is rigorously systematic, and like any system, its parts are interdependent. So the book cannot be studied piecemeal; the ordering of the chapters is deliberate and necessary. Once

* A random sample is one in which (1) every member of the population has an equal chance of being included in the sample and (2) each selection is made independently of all the others.

† The notes in this book are of two kinds and are denoted by two kinds of superscript. Footnotes are indicated by conventional footnote symbols (*, †, etc.); other notes, which appear in the back of the book, are indicated by numbers.

The footnotes are intended to supplement the central discussion at any given time and are important to a full understanding of that discussion. A footnote may refine an idea by restricting or extending its implication; it may explain a pedagogical technique by commenting on the relative importance of its various attributes; or it may supply cross-references that are not essential but are enriching. In fact, the objective in every case is enrichment. The way to use the footnotes, then, is (1) *to ignore them* as you work through a section for the first time and (2) *to study them carefully* the second time through. Every chapter has a few footnotes.

The notes in the back of the book are more technical and go beyond the scope of the text. Ignore them at least until you have mastered the chapter to which they refer.

you have worked through it, however, the book will serve as a convenient reference for you in your professional life. It has been organized with that in mind.

DESCRIPTION OF DATA VERSUS INFERENCE TO POPULATION

A reader trained in statistics would have noticed in the preceding section a subtle change between the eighth and ninth paragraphs. Every comment before that point was concerned with the description of a set of data—its configuration, average value, dispersion, and relation to other data sets. But when we began to consider taking a sample from a population and estimating properties of that population from what we know about the sample, we were shifting from description to inference.

That is an important shift—so important that I have devoted a special chapter to it. Chapter 7 is entitled "Description to Inference: A Transition."

FACING MATHPHOBIA

Many persons who are intelligent and who perform many other tasks well find themselves frozen in fear when confronted by any mathematical problem beyond the level of basic arithmetic. If you are not afflicted with such mathphobia, skip this entire section and proceed to Chapter 2. If you do have such a phobia, I am not going to attempt to rid you of it; no book is likely to accomplish that. What I believe I *can* do in this section is show you that your phobia need not be aroused by the contents of this volume, because there is very little mathematics here beyond basic arithmetic—that is, the four fundamental processes of addition, subtraction, multiplication, and division. By "very little" I mean (1) formulas, (2) some arithmetic that is more advanced than the four fundamental processes, and (3) graphic representations of data.

Because most of the concepts in these three categories are or have at some time in the past been familiar to you, the remainder of this section is mostly a very brief review. The one concept that may be new to you (the frequency distribution) is probably the easiest of them all; nevertheless, because it is new, it is treated separately in Chapter 2.

Formulas

You may have thought of formulas as guides to computations: You plug numbers into a formula, follow the rules of algebra, and out comes an answer. Formulas can indeed be very useful in that way, but the emphasis in this book is on the *concepts*—the logical structures—that underlie the computations.

For that reason, the only formulas in the book are definitional. (A *definitional formula* is an equation that defines a concept mathematically.) There are no calculating formulas even where there are calculations, because the computations are

there only to illustrate the concepts, and the concepts are illumined by the definitional formulas.

So when you see a new formula, try to discover what it means; look for the relations that it specifies. For our purposes, that specification usually need not be very precise; you may think, for example, of one term in an equation as being "larger" or "smaller" than another instead of "3.14 times as large" or "$\frac{1}{3.14}$th as large." Or you may note that as one term becomes larger, another becomes smaller. For our purposes, that frequently is the most important observation you can make. For example, take the equation

$$v = \frac{d}{t}$$

where v is velocity, d is distance, and t is time. It says that you can find v by dividing the numerator (d) of the expression d/t by the denominator (t). A body that moves a long distance in a given amount of time is moving faster than one that moves a short distance in the same amount of time. Conversely, one that takes a long time to cover a particular distance is moving *less* rapidly than one that takes a short time to cover the same distance. The two sides of any equation are equal by definition; so if there is a change in one side, the other must change, too, in a way that restores equilibrium. Thus, increasing a numerator in a right-hand term has the effect of *increasing* the left-hand term, while increasing the denominator makes the left term *smaller*. To understand basic relationships, this kind of knowledge is really all you need.

If, after you have analyzed a formula in this way, you want to pursue its implications for manipulating data, refer to the "calculation box" that you will find near the text where the formula is introduced. If you want to investigate still further, consult a text on statistical methods. But calculation is not our main concern; for us, it is but another way of illustrating a concept. Our major concern is the comprehension of the concept rather than the calculation of an answer that is numerically correct.

Arithmetic

Concerning the issue of more advanced arithmetic, three concepts will suffice for comprehending the ideas presented in this book: (1) the square, (2) the square root, and (3) negative numbers. If you do suffer from mathphobia but do not have any difficulty with these three concepts, skip the rest of this subsection.

The *square* of a given number is that number multiplied by itself. If you square 3, you multiply 3 by itself; the result is 9.*

* Probably the most common symbol for multiplication is an x interposed between the values that are to be multiplied. When x is also being used to represent a *value*, however, using it to signify an operation (multiplication) can be confusing. One way to avoid the confusion is to let *parentheses* signify multiplication—that is, any value bounded by parentheses is to be multiplied by whatever value is represented immediately adjacent to either parenthesis. Thus $2(2) = 2 \times 2 = 4$, and $4(3 + 17 + 5) = 4(25) = 4 \times 25 = 100$. Also $(2)2$ means the same as $2(2)$, and $(3 + 17 + 5)4$ is the equivalent of $4(3 + 17 + 5)$.

3^2 = 3 times 3 = 3(3) = 9

10^2 = 10 times 10 = 10(10) = 100

100^2 = 100 times 100 = 100(100) = 10,000

The *square root* of a given number is the number that when multiplied by itself produces the given number. After you have squared a number (e.g., $3^2 = 9$), taking a square root of the result ($\sqrt{9} = 3$) gets you back to where you started, namely, 3.

$3^2 = 3(3) = 9$ so $\sqrt{9} = 3$

$10^2 = 10(10) = 100$ so $\sqrt{100} = 10$

$100^2 = 100(100) = 10{,}000$, so $\sqrt{10{,}000} = 100$

Conversely, after you have taken the square root of a number (e.g., $\sqrt{9} = 3$), squaring the result gets you back to where you started (9).

$\sqrt{9} = 3$ so $3^2 = 3(3) = 9$

$\sqrt{100} = 10$ so $10^2 = 10(10) = 100$

$\sqrt{10{,}000} = 100$ so $100^2 = 100(100) = 10{,}000$

In general, since the operations of squaring and taking the square root are precisely the inverse of each other, squaring the square root of a quantity yields the quantity that you started with:

$$(\sqrt{x})^2 = (\sqrt{x})(\sqrt{x}) = x$$

Similarly, taking the square root of the square of a quantity also yields the quantity that you started with:*

$$\sqrt{x^2} = \sqrt{xx} = x$$

To put it another way, *the square of $\sqrt{x}$ is x, and the square root of x^2 is also x.*

A *negative number* is opposite in sign to a positive number. When you add such a number to a positive number of the same magnitude, the result is 0. If in Figure 1-1 you think of the horizontal line (the X-axis) as a balance beam with its fulcrum at 0, you can see that a -2 (i.e., negative 2) will balance a $+2$ (i.e., positive 2), a -3 will balance a $+3$, and so on. If you turn the page clockwise 90 degrees, the same will be true of the other axis (the Y-axis).

* When the terms that constitute an expression are *letters* rather than numbers, the multiplying operation is implied by the mere juxtaposition of terms: xx means "x multiplied by x," xy means "x multiplied by y," ab means "a multiplied by b," and so on.

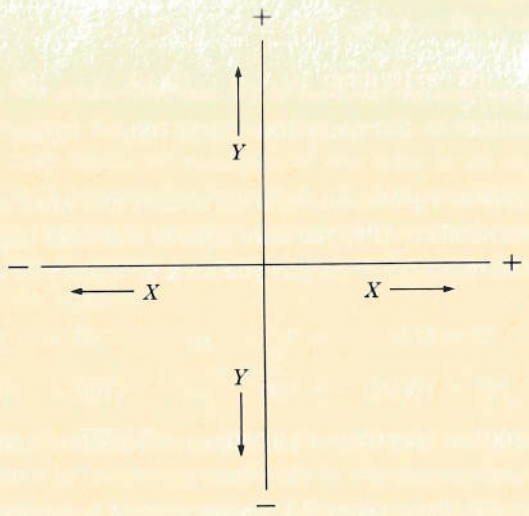


FIGURE 1-1 Two measures can be represented on a two-dimensional surface by a single point (often called a *data point*). If both of the measures are 0, that point will be precisely where the axis lines cross at the center of the graph.

If the data to be represented do not include any negative numbers—a frequent occurrence—a graph will consist of only the upper right-hand quadrant of Figure 1-1. (That's the part above the horizontal line and to the right of the vertical line.) But if there are negative numbers in the data, a surface similar to Figure 1-1 will be necessary to represent all the data graphically.

Graphs

Figure 1-1 lays the foundation for the remainder of this chapter, which is concerned entirely with *graphs*. Read the following descriptions rather quickly now; then return to them after you have finished the chapter.

You will find two kinds of graphs in this book:

1. When the relation between two quantities (variables*) is plotted, the horizontal axis (X) represents changes in one of these quantities, and the vertical axis (Y) represents changes in the other. This is the classical meaning of the term *graph*.
2. When many objects are measured on one dimension X, the numbers of objects at various values of X can be represented as a stack of units at each of the many points of the X-axis. In this usage, Y is the height of the stack at each of those points. This is called a *frequency distribution* (see Chapter 2).

* A *variable* is a quantity that under varying conditions may have varying values.

Some mathematicians prefer to reserve the term *graph* for the first of these two kinds of representation, but we shall accept a broader (and more common) meaning and refer to both kinds as graphs.

An example of the first kind of graph is one that represents the relationship between X, the temperature of the air in a room, and Y, the amount of moisture the air will hold—that is, how much water can be added before some of it condenses out of the air. The curve looks something like the one in Figure 1-2. Figure 1-3 shows how performance on a task varies with the performer's level of arousal; there is no zero on either axis. In each case, a dot represents a data point, and the simple curve that has been drawn is the one that best fits the points.

The second of the two kinds of graph described above is the frequency distribution. It is the focus of our next chapter.

But before we leave the first kind of graph, I want to warn you against drawing precipitous conclusions from your reading of *any* kind of graph. The two arrows in Figures 1-2 and 1-3 indicate the directions in which two variables, X and Y, *increase*. It is possible, however, for the larger magnitudes of a variable to be represented by the smaller number on the graph. If, for example, the Y variable is *skill at golf*, the lower scores (strokes per 18 holes) denote greater skill than do the higher ones. In a case like that, increase in skill could be represented by a falling rather than a rising curve. (The "curve" *could* be a straight line.) For example, if the X variable were *amount of practice* (measured in hours), the graph would look

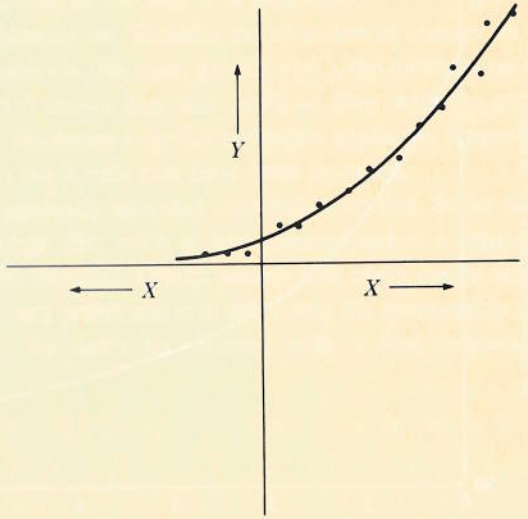


FIGURE 1-2 Graph of the relation between the temperature of a body of air (X) and the amount of water it will hold (Y). On the X-axis, the zero point is the temperature at which water freezes.

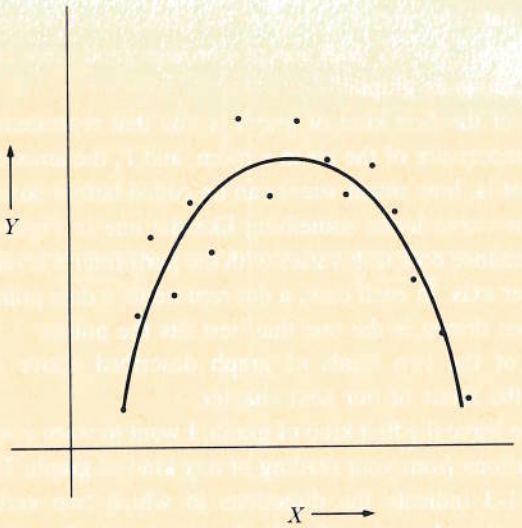


FIGURE 1-3 Graph showing the relation between level of arousal (X) and efficiency of performance (Y). There are no negative measures on either axis, and this graph uses only one quadrant of the surface displayed in Figure 1-1.

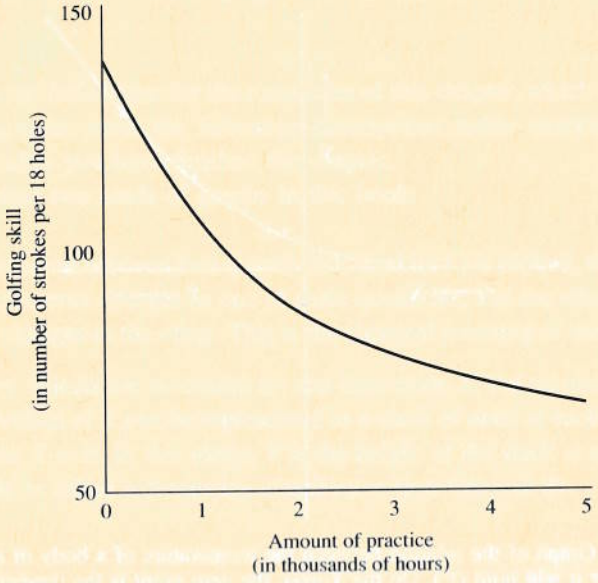


FIGURE 1-4 Graph of the relation between amount of practice (X) and golfing skill (Y).

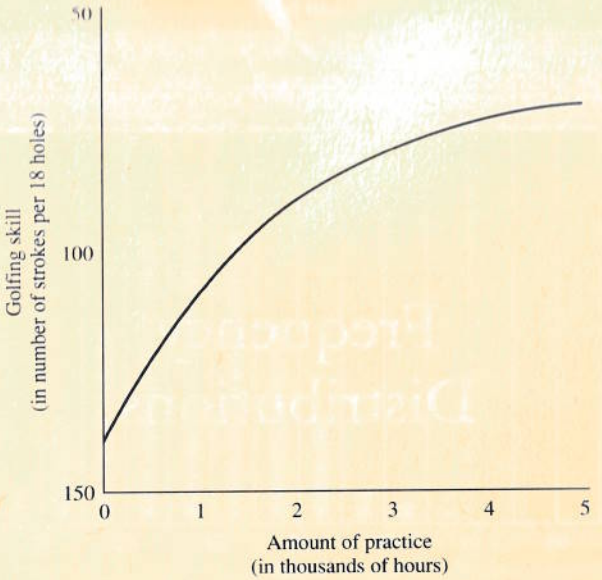


FIGURE 1-5 Graph of the relation between amount of practice (X) and golfing skill (Y). In order to show golfing skill rising instead of falling with an increasing amount of practice, the golf scores had to be plotted upside down.

something like Figure 1-4. The relation between X and Y is positive, but it appears to be negative because of the way Y is measured.

There is an alternative. To make the direction of change on the *graph* reflect that of the *Y variable* (in this case, golfing *skill*), the numbers on the Y-axis can be reversed. Figure 1-5 shows how that could be done with the golfing data. Figure 2-11, on page 22, illustrates a similar technique applied to *frequency* data. There, it is the numbers on the X-axis that have been reversed.

There is no generally accepted convention on this. Some writers are uncomfortable with a graph that even at first glance implies a relationship that is the opposite of the one intended. Others assume that every reader will carefully examine each axis of every graph and will infer only the relationship that is justified by that examination. So don't be satisfied with a first impression; it could be misleading.

2

Frequency Distributions

Every year, Central University administers a scholastic aptitude test to each of its incoming freshmen. After the tests have all been scored, each freshman is handed a small card with a score on it; then all the freshmen are herded into the football stadium.

A university official stands on the sideline of the field with a microphone in hand. She points to the space between the west goal and the adjacent 5-yard line and announces that anyone with a score below 5 should come down and stand in the middle of that space. Nobody moves, so she walks 5 yards to the east and repeats the instructions for that space. There is a long pause, then one miserable soul slinks down to the field; he is the only one in the *class interval* 5 through 9. Another call (10 through 14) yields 2 students, a fourth (15 through 19) gets 4, a fifth (20 through 24) produces 13, and so on,* with increasing frequencies through the middle scores followed by decreasing ones after that. Figure 2-1 is an aerial view of the field after all the students have assumed their positions. Notice that all of the students whose scores are in a given interval (e.g., 15–19) have been placed at the midpoint of that

* Technically these are all examples of the kind of class interval known as a *score interval*. The section on "Grouped-Frequency Distributions" will show you that the *exact intervals* here are 4.5 to 9.5, 9.5 to 14.5, 14.5 to 19.5, and so on up through the highest interval that contains any scores: 64.5 to 69.5.

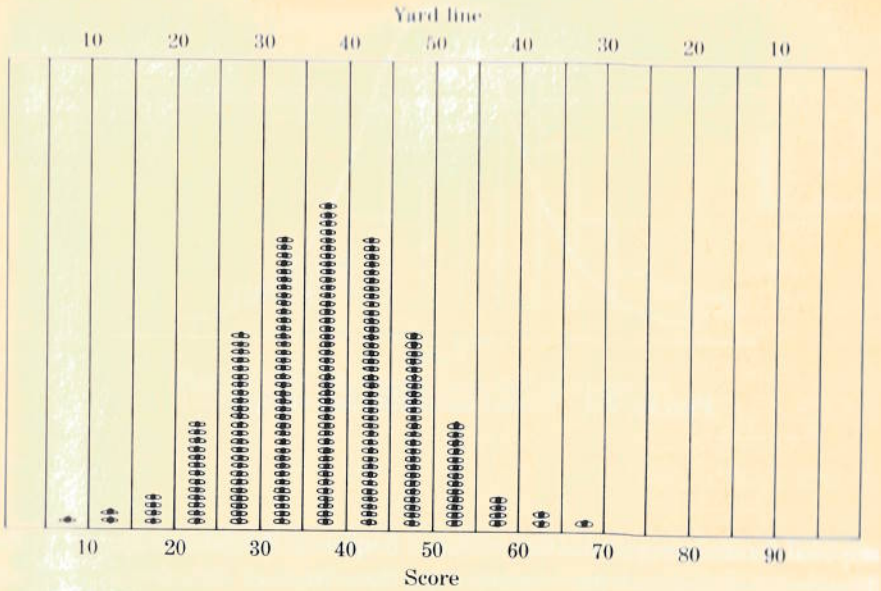


FIGURE 2-1 Students standing on football field in columns determined by scores on an aptitude test—a frequency distribution.

interval (e.g., four students at 17). For many purposes we may choose to treat all of those scores the same (e.g., as 17s). I shall have more to say about that very soon.

NORMAL DISTRIBUTIONS

Central U. lacks the budget for it, but if our official had a rope long enough to run from the sideline out around the entire freshman class and back to the sideline, the rope would form what is known as a *normal curve*. A normal curve encloses a *normal distribution*. The "norm" here is a mathematical ideal that is frequently approximated by actual measurements such as the ones taken of CU freshmen. It is *a mathematical model of randomness*. The normal curve is sometimes called a "bell curve" because of the bell-like shape that is clearly discernable in Figure 2-1.

If instead of draping the rope loosely in a smooth curve, our official were to instruct the end student in each column to grasp the rope firmly and pull it taut, the rope would form a series of straight lines and angles known as a *frequency polygon*. Alternatively, we might enclose each column in a rectangle. The resulting figure would be a series of rectangles, each with a width equal to that of the class interval and a height determined by the number of students within that interval—e.g., 2 within the interval 10 through 14, 4 within the interval 15 through 19, and so on. Such a series of rectangles is called a *histogram*.

Whenever measures are arranged in order of magnitude and a frequency is

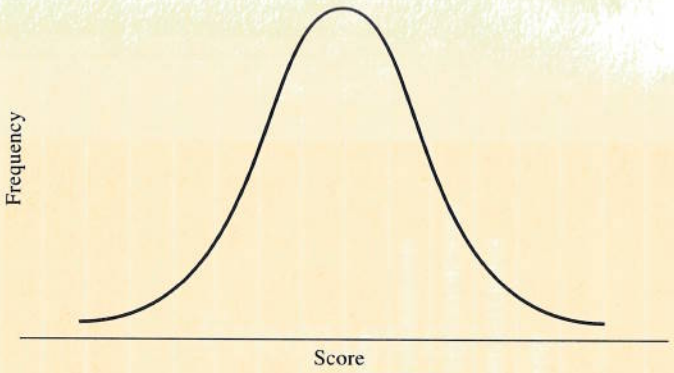


FIGURE 2-2 Normal curve from data of Figure 2-1.

recorded for each magnitude, the result is a frequency distribution. The *curve*, the *frequency polygon*, and the *histogram* are three ways of depicting a frequency distribution graphically (see Figures 2-2 through 2-4).

Quantities that are complexly determined do tend to form normal distributions. Scholastic aptitude measures show that tendency. Such measures have many determiners, both hereditary and environmental; some of those determiners influence an individual's score in an upward direction, some in a downward. In a few individuals, there is a preponderance of positive determiners; theirs are the scores in the upper (right-hand) *tail* of the distribution. In a few others, negative determiners predominate; their scores form the lower tail. Most scores, however, represent more balanced combinations of determiners; they form the large middle area of the distribution.

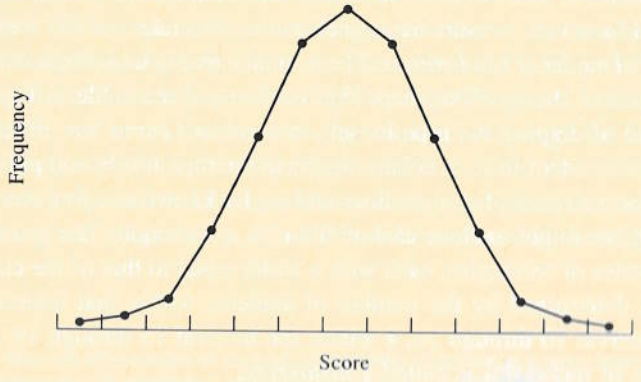


FIGURE 2-3 Frequency polygon from data of Figure 2-1.

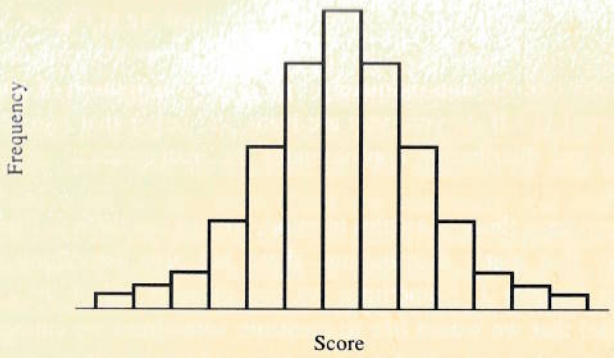


FIGURE 2-4 Histogram from data of Figure 2-1.

Each combination of determiners is assumed to be a matter of chance, and you can readily see that the probability of, say, 100 determiners being all positive (or all negative) in any individual would be vastly smaller than the probability of there being approximately half and half. If you *don't* readily see that, toss a mere three coins just eight times and plot the number of heads from each toss (0, 1, 2, and 3 are all possibilities).* You should come up with a distribution similar to the one in Figure 2-5 (although with such a small number of observations your distribution might differ markedly from that one). "All heads" (score 3) does not occur often by chance, and neither does all tails (score 0). The middle scores are much more frequent.

* In this example, the three coins correspond to 3 determiners (instead of the 100 cited in the example we just considered). The eight tosses correspond to 8 students instead of the 200 CU freshmen who took the aptitude test.

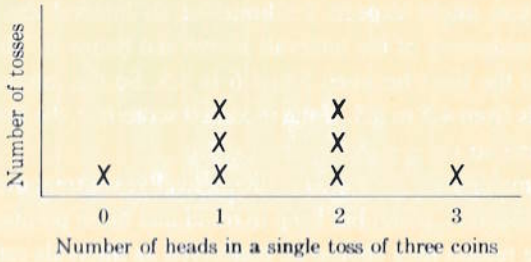


FIGURE 2-5 Distribution of eight tosses of three coins.

By the way, do you get the impression when looking at Figure 2-5 that if there were more determiners and more observations the distribution of coin tosses would look very much like the approximately "normal" distribution of students that we got from Central U.? If do you, you are right; the larger that number is, the more closely the actual distribution approximates the mathematical model we call the normal curve.

While you have these examples in mind, we can use them to illustrate not only the normality that many distributions share but also something that was first mentioned in Chapter 1: Sometimes we can measure *all* of the objects (for us usually people) that we would *like* to measure; sometimes we cannot. All of those objects taken together are called the *population* of interest; that part of the population that we study is called a *sample*.

The symbol for the size of a sample is n; for the *size* of the population it is N. In the case of the CU freshman, n is 200 if you intend to apply the results of your study to persons outside that group. If you do *not* plan to generalize, that particular class is the population of interest, and N is 200. If you *do* generalize beyond that group—say, to college freshman in general—you will be unable to count or measure the entire population; so anything you say about that population must be inferred from what you know about the sample.

The other example at hand is the coin experiment. There, n is 8, your inference is to *all* coin tosses, and again, N is unknown to you, as is everything else about the population. That is, you cannot observe and describe *all* coin tosses; you can only *infer* properties of such a population from those of the sample that you *do* observe. Our focus in this chapter is on description, not inference; I mention inference here because this section emphasizes normal distributions, and most of the inferential statistics in this book assume the normality of distributions.

GROUPED-FREQUENCY DISTRIBUTIONS AND THE MEANINGS OF SCORES

One meaning of "score" is a *point* on a scale. But that is an ideal. In practice a measuring operation places an observation within the limits of an *interval*, and thereafter that observation is treated as if it were at the *midpoint* of that interval even if it is not. As you might expect, the limits of an interval are halfway from its midpoint to the midpoints of the intervals above and below it: The limit between 5 and 4 is 4.5, and the limit between 5 and 6 is 5.5. So the interval indexed by the number 5 extends from 4.5 to 5.5. If the recorded score is 9, the limits of the interval are 8.5 to 9.5, and so on.

In order to manipulate measures mathematically, we treat them as though they are located at points on a scale; but keep in mind that those points are the midpoints of *intervals*. One reason they should be conceived as intervals rather than points is that many—probably most—of the variables we measure are *continuous* rather

than *discrete*. Time passes gradually; the numbers on your digital wristwatch change abruptly. (Time is a continuous variable; the watch's measurement of it is not.) Driving speed changes gradually ("continuously"); "number of speeding tickets" progresses in discrete units. "Scholastic aptitude" varies continuously; scholastic aptitude test scores are discrete, like the numbers on your digital watch.

I shall have more to say later about the continuity and discontinuity of variables. Right now all you need to know is that in many applications, a reported score represents not a *point* on a scale of discrete units but an *interval* on a continuous scale and that it is nevertheless treated as though it were at the midpoint of that interval.

The conceptually simplest and at the same time most precisely accurate frequency distribution results from merely listing every score that is represented on the baseline and then counting the number of times each of them actually occurs. But doing it that way is like reconnoitering hilly terrain by hiking through its rocks and trees: You get a maximal amount of information, but it is not sufficiently organized to form the kind of pattern that you could see easily from an airplane.

Data generated by Central University's freshmen (Figure 2-1) have been organized into class intervals. To see how different the display might have looked if the data had *not* been "grouped," compare Figure 2-1 with Figure 2-6 and Figure 2-7A with Figure 2-7B. Figure 2-6 shows the entire 200 student scores ungrouped. In Figure 2-7A I have magnified a portion of Figure 2-3 (a frequency polygon that corresponds to the distribution in Figure 2-1) in order to emphasize the details exposed by the ungrouped configuration (Figure 2-7B).

If you count the number of students whose scores are in each of the class intervals of Figure 2-7A, you will find just 1 in the class interval 5 through 9. In the 10 through 14 interval there are 2; in the 15 through 19 category, 4; 20 through 24 holds 13 scores; and 25 through 29 has 24. Notice that corresponding intervals of

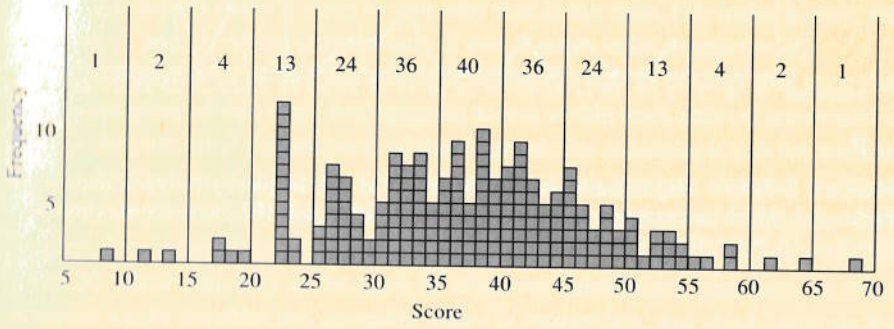


FIGURE 2-6 Ungrouped scores from which Figure 2-1 was constructed.

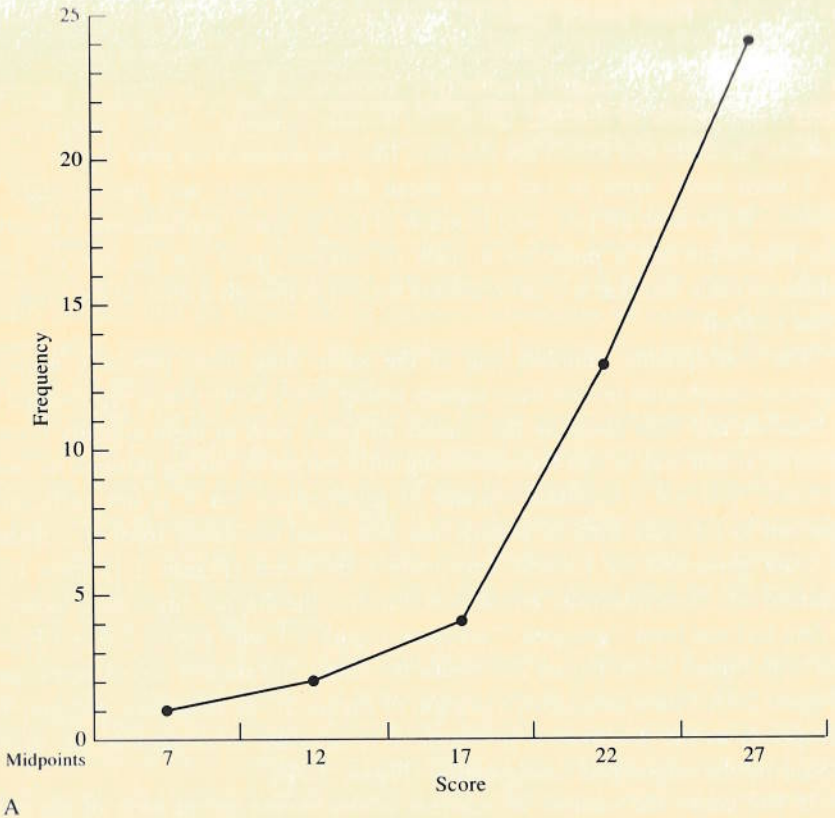


FIGURE 2-7(A) Magnification of the lower tail of the frequency polygon depicted in Figure 2-3, which corresponds to the distribution in Figure 2-1. **(B)** Same set of scores as in Figure 2-7A but without grouping.

Figure 2-7B contain exactly those same scores. The only difference is that in Figure 2-7A the scores have been "grouped"—i.e., *all* the scores in each interval have been moved to the midpoint of that interval, whereas Figure 2-7B gives you the precise location of each score. As you can see, that additional information obscures the general trend of the data.

You may have noticed also that the *numbers* denoting the scores 10, 20, and 30 have been shifted to the right of the *lines* (yard lines in Figure 2-1, mere ticks below the baseline in Figures 2-6, 2-7A, and 2-7B) that separate one class interval from another. The adjustment is precisely one-half of a unit. That is a refinement of Figure 2-1, because from the beginning I defined the lowest interval as "scores of 5 through 9," the next "10 through 14," and so on. Given that definition, the two limits of each score interval must be (1) below its lowest score and (2) above its highest score.

For example, a score of 10 is within the interval "10 through 14," *not* on the line *between* the intervals "5 through 9" and "10 through 14"; so the "10-yard line" in Figure 2-1 lies *below* a score of 10. That line is halfway between the highest score in the lower interval (in this case 9) and the lowest score in the higher interval (in this case 10); the line is therefore actually at 9.5. Accordingly, any score of 10 will be placed *above* the lower boundary of an interval that extends from 9.5 to 14.5, and all other intervals are marked off in the same way (e.g., 4.5 to 9.5 . . . 14.5 to 19.5 . . . 64.5 to 69.5). These are the exact class intervals that I promised on the first page of this chapter.

Five illustrations of the distinction between the score limits and the exact limits of a class interval are given in Figure 2-8. Score limits are inscribed above the baseline, exact limits below it.

This diagram also identifies (by means of a small arrow) the *midpoint* of each interval. Notice that the only midpoints that are whole numbers (41, 42) are in the two intervals (B and C) that are 3 and 5 score units long, respectively; 3 and 5 are odd numbers. Every interval that subtends an *even* number of units (2 in A, 10 in D, 20 in E) has a midpoint that falls *between* the middle *two* scores (40 and 41, 44 and 45, and 49 and 50, respectively). The midpoints of these three pairs of scores are 40.5, 44.5, and 49.5, none of which is a whole number. A close examination of Figure 2-8 should make it clear to you why that is the case. (To make it easy, just compare A and B, which contain only 2 and 3 score units, respectively.)

The importance of this discussion of midpoints lies in the fact that once data have been grouped into class intervals, the scores within each interval are treated as though they are all at the midpoint of that interval. For example, if you are working with a distribution of 30 scores grouped into the class intervals depicted in Figure 2-8A, subsequent calculations will be done with 30 scores none of which is a whole number. The same 30 scores deployed on the baseline pictured in Figure 2-8B will yield 30 whole numbers to be used in whatever calculations you have in mind. Because whole numbers are generally easier to manipulate, and since intervals that

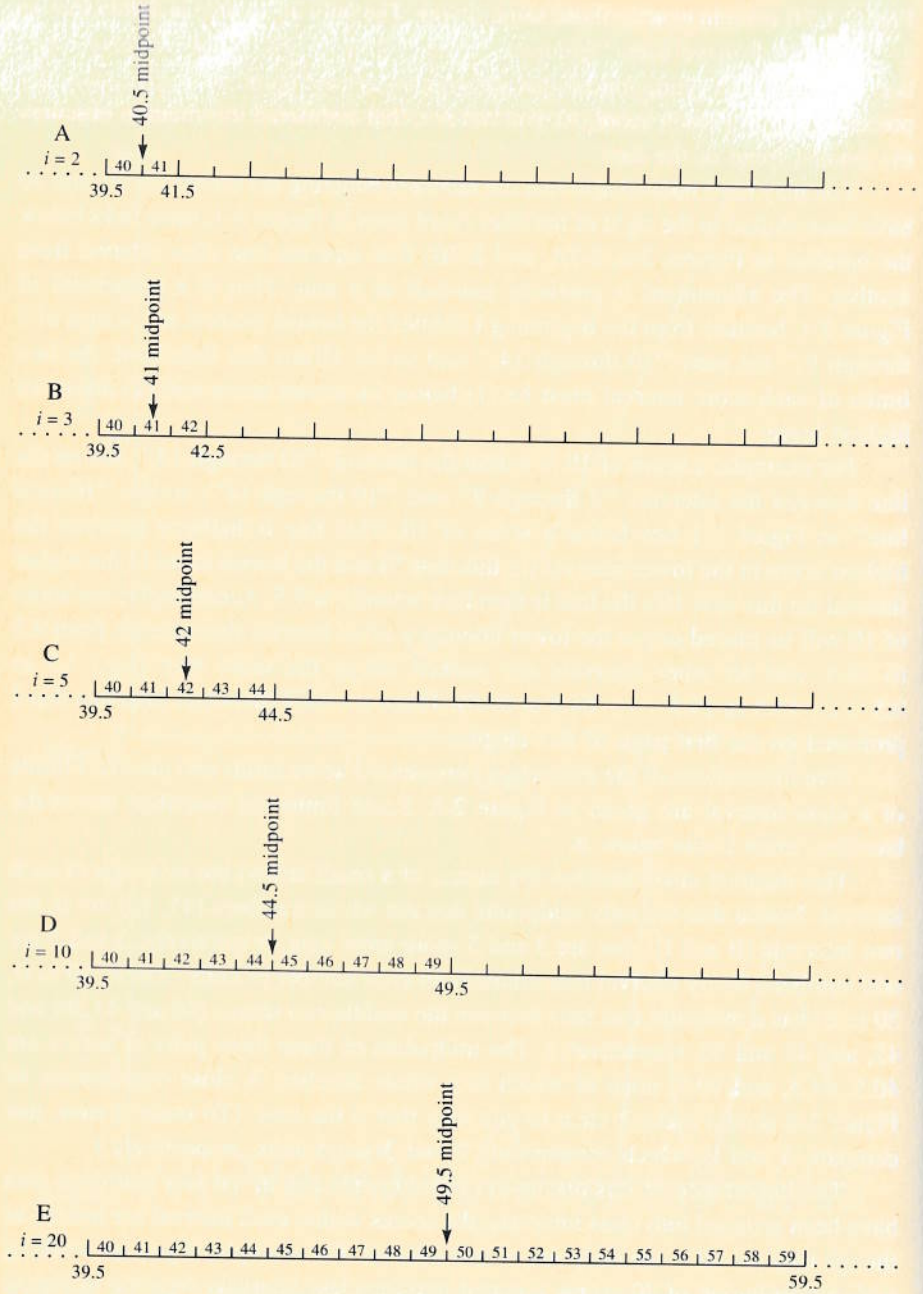


FIGURE 2-8 Class intervals of 2, 3, 5, 10, and 20. The numbers arranged sequentially from left to right above each baseline are scores. The *score* limits of intervals, reading top to bottom, are 40–41, 40–42, 40–44, 40–49, and 40–59. Corresponding exact intervals are 39.5–41.5, 39.5–42.5, 39.5–44.5, 39.5–49.5, and 39.5–59.5 *Midpoints* are 40.5, 41, 42, 4.5, and 49.5.

subtend odd numbers of score units have midpoints that are whole numbers, that kind of interval is frequently preferable to one containing an even number of units.*

SKEWED DISTRIBUTIONS

If the tail of a distribution is extended abnormally, that distribution is said to be *skewed*. When the direction of the extension is downward (i.e., toward the lower values), it is further categorized as *negatively* skewed. If your instructor in this course should give you an especially easy test, the distribution of your scores might look rather like Figure 2-9, which illustrates a negative skew.

If the upper tail is similarly extended, the skew is said to be positive, and the distribution is *positively* skewed. A difficult test would produce a distribution more like Figure 2-10, because only a talented (or industrious) few deviate very far from the lowest possible score.

OTHER CONFIGURATIONS

There are other possible distributions. If we were measuring conformity behavior, like that of motorists at a busy intersection, we might get a J *curve*, like the one in Figure 2-11. The same would be true of a distribution of scores on a test that is extremely easy (so easy that most people get perfect scores) or extremely difficult (so difficult that most of the scores are zero). (See also the discussion on "Skewed Distributions.")

An entirely different configuration would emerge if we were to measure the standing height of humans, because there are two physical types of humans, male

* On the other hand, our decimal numbering system makes an interval size of 10 the most convenient, especially during the tallying process. For that reason, an interval of 10 or some multiple thereof is worth considering (even though it does not give you midpoints that are whole numbers) if it yields an appropriate number of class intervals (somewhere between 10 and 20) and if you are faced with so many scores that convenience in tallying is an issue.

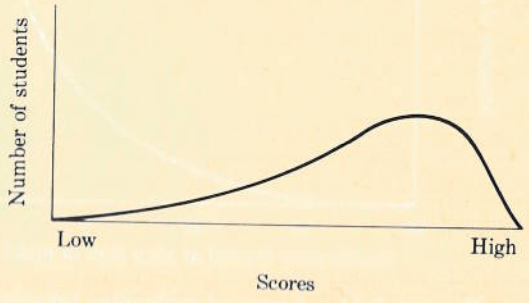


FIGURE 2-9 Negatively skewed distribution.

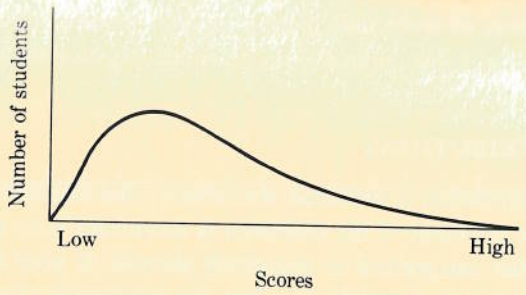


FIGURE 2-10 Positively skewed distribution.

and female. Thus, we should obtain a *bimodal* distribution, as shown in Figure 2-12. Similarly, if your instructor were to spring a pop quiz on Chapter 9 of this book at a time when only half of the class had read it, the distribution of those scores would be bimodal. (Note that in Figure 2-12 the distribution is apparently quite symmetrical; however, a bimodal distribution can be asymmetrical, as indeed this one surely would be if, say, it were composed of twice as many women as men.)

The J distribution is difficult to deal with statistically, and a bimodal distribution can be dealt with by separating the two normal distributions that are partially concealed within it. So these other configurations are of only passing interest to us here.

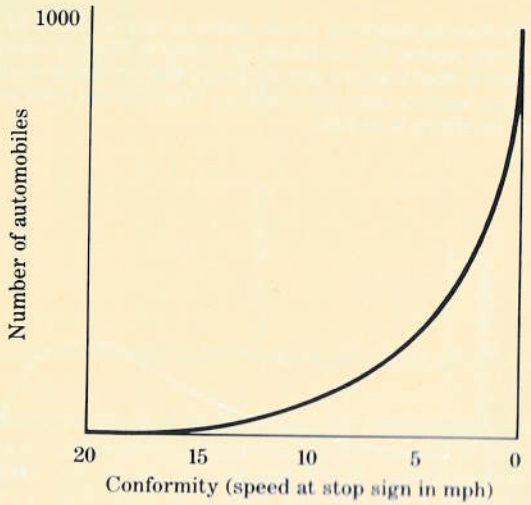


FIGURE 2-11 J curve of conforming behavior.

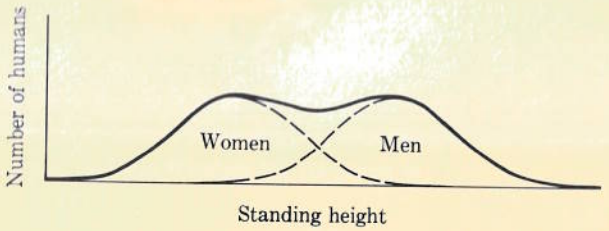


FIGURE 2-12 Bimodal distribution of humans on a scale of height.

SUMMARY

Many frequency distributions in the social sciences are approximately normal; that is, in the typical distribution there are a few very low scores and a few very high ones, but the great mass of individuals tend to pile up on the middle scores. This occurs because the probability of all the many determiners of a trait pointing in the same direction is virtually nil, and that of a balanced combination of determiners is much higher. Although other configurations do occur (positive skew, negative skew, J curve, and bimodality are mentioned), our primary concern in this book will be the normal distribution, because it serves as a good approximation to the kinds of distribution most frequently encountered in behavioral and medical investigations. Moreover, it is the configuration for which the best-known statistical treatments are available. We shall encounter some of those treatments in subsequent chapters.

There are many possible ways of organizing and displaying data. One is to "group" frequencies—to pool individual observations into class intervals. Grouping has both positive and negative consequences. A negative consequence is the loss of information that occurs when scores of varying values are assigned whatever value is at the midpoint of a given class interval. A positive consequence is that grouping tends to reveal underlying patterns by allowing many random variations to cancel each other. If you have a large number of observations to organize, another positive effect is that grouping can simplify calculations.

3

Measures of Central Tendency

Somewhere in the intermediate grades you were introduced to the concept of *average*, and you may have used it ever since on the assumption that there is only one such. Actually, there are several types of averages, of which three of the most common will be described here: the mean, the median, and the mode.

The Mean (μ and $\overline{X}$)

The average that you learned about in grade school was a *mean*. The mean is not, as you were led to believe, *the* average, but it does have characteristics that make it the best one to use in many circumstances. Whenever the frequency distribution is fairly symmetrical and a calculator is available, the mean is the statistic of choice. The computation time is necessary because all of the scores must be added together before the sum can be divided by N (the number of scores), and there are sometimes thousands of scores to add. If you have access to the entire target population, the formula for the mean is

$$\mu = \frac{\Sigma X}{N} \tag{3-1}$$

where μ is the mean of the population, Σ is a combining term meaning "sum of," and X refers to the *raw scores* that have been recorded.* The expression ΣX (read "summation ecks"), therefore, is the sum of all the raw scores in the distribution. N is the size of the population.

If you do *not* have access to the entire population, the scores that you *do* have constitute a sample, and the formula for *its* mean is

$$\overline{X} = \frac{\Sigma X}{n} \tag{3-2}$$

where $\overline{X}$ is the mean of a sample and n is the size of that sample. Notice that *the operations specified by the two formulas are exactly the same*. If you calculate what you believe to be a population mean using Formula (3-1) and subsequently learn that the scores you have comprise only a small fraction of the target population, all you need to do is to change the label from μ to $\overline{X}$; no further calculation is necessary.

Because in published research the description of samples is more common than that of populations, every reference to "the mean" after this chapter will be to a sample mean unless otherwise noted (although your own studies may frequently be of populations).

In any event, Formulas (3-1) and (3-2) say the same thing, and what they say has two aspects. The first is that the formulas *define* the mean; the second is that they specify a procedure for *calculating* it. But this book is not about calculating; our primary objective is rather to acquire an understanding of statistical concepts. The concept of the mean can best be understood through the following illustration.

Essence of the Concept

Try to imagine a long beam made of an exotic metal that is totally rigid and utterly weightless. Upon that beam we shall place 14 cubes (to represent 14 scores), each of equal weight. Figure 3-1 shows one possible distribution of cubes.

At what point on the beam would a fulcrum (support) have to be placed to establish an equilibrium? That is, what is the *balance point* of the distribution? Because this distribution is symmetrical, it should be easy to see that it would

* A *raw score* is one that does not imply a comparison with any other score; "inches," "points," "runs," "hits," "errors," and "number right," when those units are simply counted, are raw scores. Raw scores are the only kind we have dealt with so far, others will be introduced later.

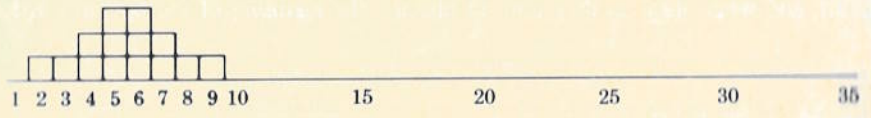


FIGURE 3-1 Fourteen cubes on a weightless beam: a symmetrical distribution.

balance if the fulcrum were placed at 5.5. But that is what all the computation is about; by using the formula, we can find the precise point we are seeking* even in cases where a diagram would be rather puzzling. (That such situations do arise will become clear in the section on "The Median.")

Another way of saying "balance point" is "the point from which all deviations sum to zero." In Figure 3-1, for example, that point is 5.5. One of the scores (the 2) deviates $3\frac{1}{2}$ units in the negative direction (left) and one (the 9) deviates $3\frac{1}{2}$ units in the positive direction (right); one is a negative and one is a positive $2\frac{1}{2}$; there are two negative and two positive deviations of $1\frac{1}{2}$; finally there are three scores $\frac{1}{2}$ unit below and three scores $\frac{1}{2}$ unit above the mean. Most distributions will not be perfectly symmetrical as this one is, but all will produce the same result: The positive and negative deviations from the mean will always cancel each other, bringing their sum to zero.

What I have been saying in different ways is that *the mean is the balance point* of any distribution of scores.

Appropriate Applications

As the balance point of a distribution, the mean is the only measure of central tendency that is sensitive to all of its scores. Probably its most important application is to other measures. Because of that balance-point feature, it is compatible with many more complex measures that you will meet later. Indeed, calculating a mean is an integral part of calculating a *standard deviation*, a product-moment *coefficient of correlation*, and all of the various *standard errors*, to name a few.

A related advantage of the mean over other measures of central tendency is its usefulness in making inferences from sample to population: The mean of a sample is the best estimate of that of the population. But even the best estimate will probably miss the mark, and is is important to know the probable extent of the error. The mean lends itself to error estimation as well, as you will see in Chapter 8.

So if you have a sample, and you want to know the population mean, you will choose the mean to represent central tendency in your sample. And if you anticipate calculating more complicated statistical measures, you will probably need to calculate your sample mean first.

As I mentioned in the preface, some people do not feel comfortable with a quantitative concept until they have followed the relevant computation. If you are one of those people, Box 3-1 provides you with an opportunity to do that. The calculation is based on the defining equation for the mean. Calculational formulas —and hence calculations—frequently differ from definitional ones (see pages 5–6), and when they do they tend to obscure the *meaning* of the operations they

* $\bar{X} = \frac{\Sigma X}{N} = \frac{\text{sum of raw scores}}{\text{number of raw scores}} = \frac{77}{14} = 5.5$

Box 3-1 Calculation of a Mean (see Figure 2-1)

(1) Class interval	(2) Midpoint X_m	(3) f	(4) fX_m
65–69	67	1	67
60–64	62	2	124
55–59	57	4	228
50–54	52	13	676
45–49	47	24	1128
40–44	42	36	1512
35–39	37	40	1480
30–34	32	36	1152
25–29	27	24	648
20–24	22	13	286
15–19	17	4	68
10–14	12	2	24
5–9	7	1	7
		$\Sigma = 200$	$\Sigma = 7400$

Column 1: Class intervals. See Figure 2-1, page 13.

Column 2: Midpoints of class intervals. When data are grouped into class intervals, all individuals (X) within each interval are treated as though they were at the midpoint of that interval (X_m). Of course, that is not strictly true, but if the class intervals are small, error is negligible.

Column 3: Frequency (f) of scores in each class interval.

Column 4: Product of midpoint and frequency (fX_m). Only one person scored in the 5–9 interval, and the midpoint of that interval is 7; so the fX_m for that interval is just 7. There are two scores in the 10–14 interval and its midpoint is 12, so its fX_m is $2 \times 12 = 24$. Four persons scored somewhere in the interval 15–19, so fX_m is $17 \times 4 = 68$, and so on through the remaining intervals.

$$\bar{X} = \frac{\Sigma X}{n}$$

It should be clear that the sum of column 3 is n. It should also be clear that simply adding all the individual X_m's will give us the same sum as that of column 4; that is, the sum of column 4 is the ΣX in the formula, give or take an error of negligible magnitude.

$$\bar{X} = \frac{7400}{200} = 37$$

represent. Since meaning is your mission here, it will be better for you to keep your calculations close to your concepts.

THE MEDIAN (Mdn)

Another way of indicating central tendency is to tell which point on the baseline divides the distribution into two equal parts. Note that I said it divides the *distribution* in half, not the *baseline*. Look back at Figure 3-1. There, the beam is the baseline; it is 35 units long, but $\frac{35}{2} = 17.5$ is not the median. Nor is the median defined as a point halfway between the lowest point (1.5) and the highest (9.5), although because of the perfect symmetry of the distribution, it happens to be there in that particular case.

Essence of the Concept

In every case, the *median* is the point between the lower and upper halves of the distribution. (Remember, the distribution is the group of individual scores comprising the sample.) In Figure 3-1, that point is 5.5, because there are 7 scores below it and 7 above. In *any* distribution with an N of 14, the median will be midway between the 7th and 8th scores (counting up from the bottom of the distribution); in any distribution of 1000 individuals, the median will be the point midway between the 500th and 501st; and so on, with as many illustrations as you care to cite.[1]

Appropriate Applications

"Who is in what half of the distribution?" If that is our question, the answer will depend upon our first finding the median. A more important characteristic of the median, however, is that although it is not sensitive to the exact location of every score in the distribution, it can be used in situations where the mean would be inappropriate.

It may not be too much of an oversimplification to say that the median should be used in preference to the mean whenever the shape of distribution departs radically from perfect *symmetry*. Consider the situation depicted in Figure 3-1. There, because the distribution is perfectly symmetrical, the mean and the median are at precisely the same place (5.5). Now look at Figure 3-2 to see what happens when that symmetry is disturbed. In that figure we have exactly the same distribution as in Figure 3-1 except that two of the scores have been shifted far to the right.

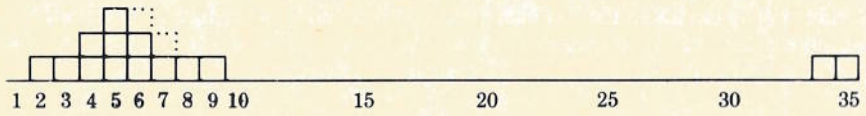


FIGURE 3-2 Fourteen cubes on a weightless beam: an asymmetrical distribution.

The balance point—that is, the mean—has shifted also; it is now 9.5, which, lying as it does above 12 of the 14 scores, is probably an inappropriate index of the central tendency of this distribution.

But what has happened to the *median* as a result of that shift of two scores? It hasn't moved at all! (Count the scores above and below it, and see for yourself.) You may say that it *should* have moved—at least a little bit—because the distribution is different from what it was before; however, even though it is insensitive to that change, I'm sure you will agree that in this distribution the median is a better measure of central tendency than the mean, because the two extreme scores have *too much* influence on the mean. Often what has happened in such cases is that scores have been forced into one category when they should have been classified into two or more. For example, if you were to plot a distribution of the annual incomes of a football coaching staff, you would probably find that most are rather close together but that the salary of the head coach is distinctly separate from the others. If you had to report a single average of the salaries of football coaches at Central University, would you use the mean or the median? If you *don't* have to confine your report to a single central tendency, it probably would be better to report the head coach separately in this case, but if salaries of the senior members of the support staff approach that of the head coach, the median of the entire staff might be an appropriate index of central tendency.

In another situation that militates against using the mean, the scale is not long enough to accommodate all of the scores at one end of the distribution. For example, whereas a long, difficult test might produce a normal distribution of student scores, a short, easy one might pile up, say, a third of the scores at the top of the scale. The true magnitude of these scores is indeterminate: There is no way to know "how far out on the bar" each weight should be placed and hence no way to locate the balance point of the distribution. But you can use the median.

Finally, there is the rather rare array of categories that are essentially nonquantitative but nevertheless appear in a universally recognized *order*. Military ranks come to mind, but any ranking structure could illustrate the genre. An especially good one, though, is the results of a race—say a marathon—in which results are recorded in two forms: (1) the precise *time* that it took each contestant to reach the finish line and (2) the *order* in which they all reached it. If you know the runners' *times*, you can calculate their mean time, but if all you have is their *ranks*, the median will have to do, because there are no *scores* to support the calculation of a mean.

THE MODE

The last of the three averages to be presented here is also the easiest. The *mode* is simply the point with the greatest frequency. In Figure 2-1, the mode is 37; in Figure 3-1, 5.5; in Figure 3-2, 5.0.

If your data are quantitative and ordered, as are nearly all of the data presented in this book, the mode's very simplicity explains one of its two main purposes: It is used when a very quick estimate is needed. It is also used specifically for identifying the typical (most common) score.

If your data are qualitative and not ordered, like sales of various colors of designer dresses or enrollments in the subject-matter categories of a college curriculum, the mode is really the only central tendency that *can* be used: You cannot count up from the bottom as you must in order to find a median, much less add scores as you must to find a mean. (There *are* no scores to add.)

The mode can therefore be used where the mean and the median cannot. Beyond that, it can supplement but not supplant either or both of those statistics.

SUMMARY

The three most common measures of central tendency are *mean*, *median*, and *mode*.

The *mean* is the balance point of a distribution. It is the only average that utilizes all available information, and when distributions are approximately normal it is the one that serves as an essential component of many of the more complex measures that you will encounter in later chapters.

The point at which a distribution can be cut in half is its *median*. It is used when precise location of the two halves is a prime concern, when data are ordered but not quantitative, and when a distribution of quantitative data is far from normal. In those circumstances the median can supplement or even supplant the mean.

The *mode* is the place where the greatest number of cases occurs; if the data are quantitative, it is the most common score. When that is exactly the information you want or when your data are neither quantitative nor ordered, the mode is the appropriate index of central tendency. In other situations it can be useful as a supplement to the median and/or the mean, but it should never supplant them.

In a skewed distribution, the arrangement (order) of the three measures along the baseline is predictable. If the skew is negative, as in Figure 3-3, the arrangement from left to right is: mean, median, and mode. If the skew is positive, as in Figure 3-4, the order is just the opposite: mode, median, and mean. Conversely, if you know the order of the three averages, you can tell the direction of the skew.

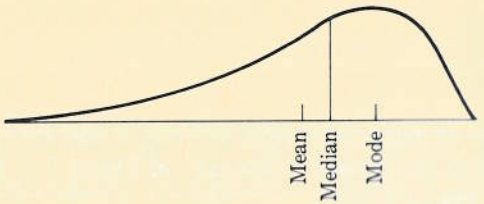


FIGURE 3-3 Measures of central tendency in a negatively skewed distribution.

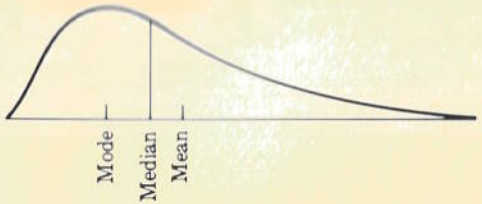


FIGURE 3-4 Measure of central tendency in a positively skewed distribution.

Because of the mean's affinity to many of the more complicated constructs that I will present later (beginning in Chapter 4), future references to measures of central tendency will be almost exclusively to the mean. For simply describing a small population, however, or for doing a preliminary exploration of data that might later be used to make inferences beyond your sample, all three measures can be helpful, as can graphic representations.

Sample Applications

The following problems require that you make some decisions about the appropriate use of statistics. In each case, you are presented with a situation that might be faced by a practitioner of the indicated discipline. Imagine yourself as that person in that situation.

Make a tentative choice without looking back into the chapter; explain it as best you can. Then review the text and refine your answer. Finally, check the back of the book (pages 165–181) for a suggested choice and brief discussion of it.

You may use that choice and that discussion to correct or further refine your response. But if you feel that yours is as good as ours, discuss it with your instructor or someone else who has a knowledge of statistics. If he or she agrees, I'd very much like to see what you have done.

Besides answering the questions in the following paragraphs, try to identify the sample and/or population that can be found in each.

EDUCATION

A cooperative vocational education program has recently opened to provide one year of training for 200 twelfth-grade students from 10 schools. The teachers are in the process of selecting and developing curriculum materials, but they are unsure of the appropriate reading level for the materials. They decide to get an estimate of the

reading skills of the students enrolled in the program, and you are called in as consultant. You suggest that the teachers administer a standardized reading comprehension test and obtain for each student a single score indicating the reading level of that student. Now what do you do?

POLITICAL SCIENCE

You are studying the domestic and military expenditures of European nations. You want to find the average amount spent by European countries on arms. You cannot afford to study all European countries, so you take a random sample. Now what do you do?

PSYCHOLOGY

You are a family counselor working with the mother of a three-day-old infant. The mother is very concerned about her child (her sister has recently given birth to an infant with birth defects) and asks you if her infant is showing normal behaviors for a newborn. You observe the infant in question, but you are not certain which behaviors are classified as normal for a neonate. Your task, therefore, is to find out how newborns tend to behave. You go to three hospitals in your city, visit the neonatal units, and observe and measure a variety of infant behaviors. For example, you dangle a large red ring in front of each infant to see whether it follows the ring visually or even attempts to grasp it. You sound a bell close to each infant's right ear and observe whether the infant turns toward the sound. You then assign points based on your observations (e.g., 1 point for following the ring visually and 2 points for grasping the ring). What single statistic would best represent all the children you have tested?

SOCIAL WORK

The director of a child welfare agency is interested in the length of time that families receive protective services. She asks you to provide information regarding the number of treatment hours received by these clients. To approach this problem you select a random sample of clients from the closed-case files. How can these data be analyzed?

SOCIOLOGY

A city council of a city of 100,000 wants to know the average income of its residents. You are asked to make an estimate but are given only a small budget for doing so. You draw a random sample of 100 residents and ascertain the average income of all residents. What computation is appropriate?

4

Measures of Variability

Different populations (and the samples extracted from them) have different central tendencies, but they differ in another significant respect as well. Consider the two curves depicted in Figure 4-1. Both represent distributions of the same area (identical N's), and both have the same central tendency; nevertheless, the two distributions are very different. In what way are they different? You can see that one is spread out more than the other. Since the baseline on which the spreading occurs is a single scale of scores, the spreading means that the scores in that distribution *vary* more than those in the "squeezed together" distribution.

Diagrams provide a superior way of approaching a new concept, but even if we had a diagram of every distribution drawn to scale so that we could compare variabilities by inspection, there would still be a need (demonstrated in Chapter 5) for an index that can enter into mathematical operations that are essentially numerical; you can't multiply or divide a visual perception with an acceptable degree of precision. There are many occasions (some of which are discussed in Chapter 5) when it is important to have some kind of numerical index of the variability of a set of scores. In this chapter, we shall examine three such indices.

THE STANDARD DEVIATION (σ AND S)

Taking its importance on faith for the moment, let us consider how an index of variability might be devised if we had none already available. It may help to have a concrete example in mind during this discussion, so let us imagine that instead of

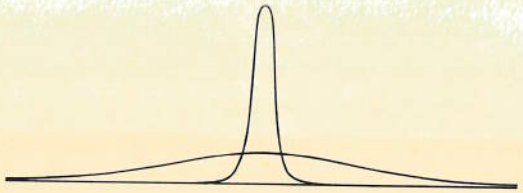


FIGURE 4-1 Two distributions with the same N's but different variabilities.

the single scholastic aptitude test mentioned in Chapter 2, those students took a *pair* of tests—one of verbal aptitude, the other of numerical. The distributions of scores on the two tests might look something like Figure 4-2.* In that diagram, the means of the two distributions have been aligned so that we may concentrate on their

* The indicated central tendencies and variabilities of the two distributions make somewhat plausible the proposition that if combined, they would form the distribution shown in Figure 2-1. However, that was not the primary consideration in their selection. Rather it was *simplicity*. We are deemphasizing computation; therefore, computations that *are* required have been made as easy as possible. It is easier to think about an interval that extends from 20 to 25, for example, than one that extends from 23 to 27, or even from 23 to 28. You will find that you can do most—possibly all—of this book's computations entirely in your head.

But do not attempt any computations unless the requisite data are readily available. For example, in Figure 4-2 do not attempt to confirm the standard deviations of 15 and 5 in the two distributions. To do so, you would need a list of the individual scores, and those have been withheld deliberately to encourage you to focus on the big picture—the configuration of the entire sample for each test and the comparison of two configurations that are very different from each other.

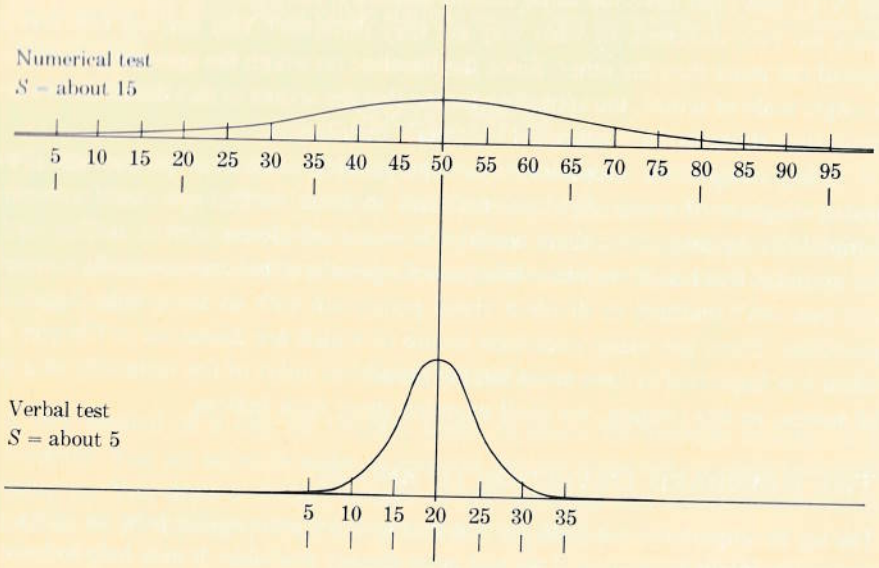


FIGURE 4-2 Two distributions with the same N's but different variabilities.

respective variabilities. As we look at the figure, we are immediately impressed by the striking difference between the dispersions of the two distributions across their baselines. But it is not enough to be impressed; we need a numerical index of dispersion (variability).

Essence of the Concept

How might such an index be devised? One way would be to compare every score in the sample to every other and to average the differences obtained thereby. But that would be entirely too cumbersome, especially with large distributions; we can get the same effect by selecting a point of reference in the middle of the distribution and measuring the distance of each individual score from that point, as in Figure 4-3. The average (mean) of those differences (without regard to sign) can also serve as an index of variability. If, for example, we were to choose the mean of our target population as our reference point, our index would be an average of the individual distances from the mean and could be obtained via the following formula:

$$\text{AD} = \frac{\Sigma|x|}{N} \tag{4-1}$$

where AD is the *average deviation*, $\Sigma|x|$ is the sum of *individual deviations* from the mean of the population, and N is the size of the population. The distance of any

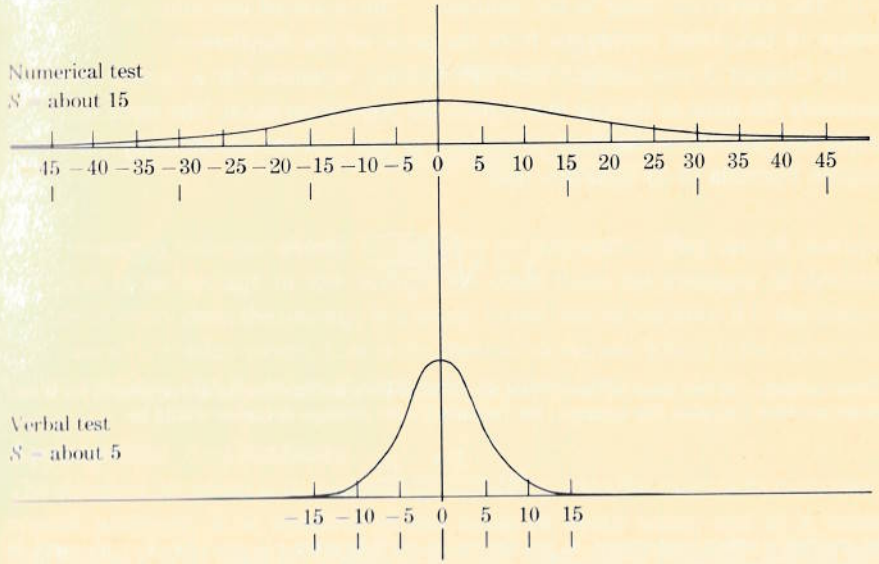


FIGURE 4-3 The distributions in Figure 4-2 with raw scores converted to deviation scores.

score from the mean ($X - \mu$) is symbolized by a lowercase x* and referred to as a *deviation score* (or sometimes simply *deviation*). The vertical bars in $|x|$ mean "without regard to sign."

The average deviation is sometimes used, but it is not common enough to be discussed here strictly for its own sake. Rather, I have included it because it is a direct expression of the basic idea that underlies the most used of all measures of variability. That idea is the concept of an *average of individual deviations* from the mean; that measure is the *standard deviation*, not the average deviation.

Now compare the little-used statistic that we have been discussing to the standard deviation (which in many applications is *the* accepted measure of variability). Here again is Formula (4-1) for the average deviation:

$$\text{AD} = \frac{\Sigma|x|}{N}$$

and here is the formula for the standard deviation:

$$\sigma = \sqrt{\frac{\Sigma x^2_{\text{pop.}}}{N}} \tag{4-2}$$

where σ is the standard deviation of the population, Σx^2 is the sum of the deviations from the population mean, and N is the size of the population. The two formulas are very similar, are they not? In fact, they are exactly alike except that for the standard deviation we take a mean of *squared* deviation scores; then we take the *square root* of that mean. But don't worry about the difference between Formulas (4-1) and (4-2). The important thing is the similarity: The standard deviation is a kind of average of individual deviations from the mean of the distribution.

In Chapter 3 you learned that the defining equation for a sample mean is essentially the same as the one that defines the population mean. The same is true of sample and population standard deviations, as you will see immediately when you compare Formula (4-2) with this one:

$$S = \sqrt{\frac{\Sigma x^2_{\text{sample}}}{n}} \tag{4-3}$$

Some authors—in fact most of them—use the *formula* for a deviation score as the *symbol* for it that appears in other formulas. For example, the formula for the average deviation would be

$$\text{AD} = \frac{\Sigma|X - \mu|}{N}$$

that notation is rather cumbersome, as you can see, so we'll be using x to stand for $(X - \mu)$—and, in the case of a sample, for $(X - \bar{X})$. Wherever it is not immediately clear from context which of these deviations is symbolized by x, be assured I shall make it so.

where S is the standard deviation of a sample, Σx^2 is the sum of the squared deviations from the sample mean, and n is the size of the sample.

As mentioned earlier, the standard deviation differs from the average deviation in that we *square* each deviation (which eliminates negative signs) and then find the square root of their mean. That mean (the mean of the squared deviations) *before* we take its square root is an entity in its own right: It is known as the *variance* of the distribution:

$$\text{Variance} = S^2 = \frac{\Sigma x^2}{n} \tag{4-4}$$

where S^2 is the variance of a sample, Σx^2 is the sum of the squared deviations from the sample mean, and n is the size of the sample.

Let us take a moment to sum up: In this and the previous chapter we have examined a sequence of concepts that share a common structure. It is important for you to be aware of that structure; not only because it will help you to understand its manifestations in concepts developed earlier but also because it will appear in other concepts later. The best way to discern the structure is to place all the items of the series in close proximity to one another. Here are the four concepts in order in which they were presented:

$$\text{Mean of the raw scores} = \frac{\Sigma X}{n}$$

$$\text{Average deviation} = \frac{\Sigma|x|}{n}$$

$$\text{Variance} = \frac{\Sigma x^2}{n}$$

$$\text{Standard deviation} = \sqrt{\frac{\Sigma x^2}{n}}$$

Their common structure should be apparent by inspection. They are all *averages*. The first is an average of raw scores; the other three are averages of deviation scores. In every case, the average is a mean. In the case of variance it is the mean of the *squared* deviation scores. Last is the standard deviation, which is the *square root* of the variance and therefore is expressed in linear units.

Appropriate Applications

The standard deviation is to measures of variability what the mean is to measures of central tendency. It is sensitive to the value of every score, and in a normal distribution it is an integral component of many other useful statistics. Among those are the product-moment coefficient of correlation and the various standard errors,

both of which will be discussed later. In scientific work, at least, the standard deviation clearly is the prevalent measure of dispersion.

As in Chapter 3 a calculation box (Box 4-1) provides you with an opportunity to try your hand at actually crunching some numbers. But here again, and for the same reason as before, the operations in the box are specified by the defining equation, not by a specialized calculating formula.

THE INTERQUARTILE RANGE (IQR)

A much easier statistic to comprehend (and to compute) is the *interquartile range* (IQR). Look at Figure 4-4. What you see is a negatively skewed distribution cut into four equal parts. At the upper end of each of those quarters is a point on the baseline called a *quartile* (Q)—the first quartile (Q_1) above the lowest quarter, the second (Q_2) above the lowest two quarters,* and the third (Q_3) above the lowest three quarters. The point just above the highest score in the distribution would logically be Q_4, but that expression is seldom if ever used.

Essence of the Concept

The *interquartile range*, like any measure of variability, is an interval. In this case the interval extends from Q_1 to Q_3. The calculation of the interquartile range is extremely easy: You simply find the difference between Q_1 and Q_3. That's it.[1]

The IQR, like the median, is insensitive to the precise values of most of the scores in a distribution, so when you use it instead of the standard deviation you discard information. But if the distribution is skewed, you can return some important information (namely, the direction of the skew) by reporting not only the *difference* between Q_1 and Q_3, but their exact *locations* as well, along with that of Q_2. (*If* $Q_2 - Q_1$ is longer than $Q_3 - Q_2$, the skew is negative, as in Figure 4-4; if $Q_3 - Q_2$ is the longer, the skew is positive.)

Appropriate Applications

The interquartile range is to measures of variability what the median is to measures of central tendency. Although insensitive to the exact values of many of the scores in a distribution, it is preferred to the standard deviation in the same kinds of situations in which the median is preferred to the mean—namely, when distributions are radically asymmetrical. It is the perfect companion to the median wherever the latter is properly applied.

A report of the incomes of a university faculty, for example, might be skewed by the high salaries and consulting fees of the college of business. If faculty

* What other familiar statistic has essentially the same definition as the second quartile? If you are not sure, turn back to page 28.

Box 4-1 Calculation of Standard Deviation (see Figure 4-2, verbal test)

(1) Class interval	(2) Midpoint X_m	(3) Frequency f	(4) $X_m - \overline{X}$ x	(5) $(X_m - \overline{X})^2$ x^2	(6) $f(X_m - \overline{X})^2$ fx^2
33–37	35	2	15	225	450
28–32	30	13	10	100	1300
23–27	25	32	5	25	800
18–22	20	106	0	0	0
13–17	15	32	− 5	25	800
8–12	10	13	− 10	100	1300
3–7	5	2	− 15	225	450
		$\Sigma = 200$			$\Sigma = 5100$

Column 1: Class intervals. See pages 12 and 16–21.

Column 2: Midpoints of class intervals. Remember when data are grouped into class intervals, all individuals (X) within each interval are treated as though they were at the midpoint of that interval (X_m).

Column 3: Frequency (f) of scores in each class interval.

Column 4: Deviation scores (x), which equal the differences between the midpoints (X_m) and the mean of the distribution ($\overline{X}$). In this case, $\overline{X} = 20$.

Column 5: Squares (x^2) of the deviation scores listed in column 4.

Column 6: Product of the squared deviation scores and the frequency of scores in the interval (fx^2).

$$S = \sqrt{\frac{\Sigma x^2}{n}}$$

The sum of the frequencies listed in column 3 is n, and the sum of column 6 is Σx^2.

$$S = \sqrt{\frac{5100}{200}} = 5.01$$

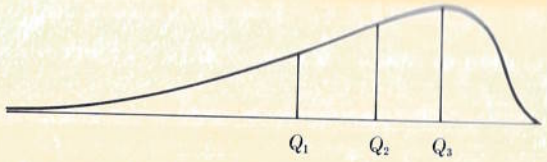


FIGURE 4-4 Negatively skewed distribution with area divided into quarters.

incomes without the college of business are distributed symmetrically, the high incomes there give the total distribution a marked positive skew, making the mean and standard deviation difficult to interpret. The median and interquartile range would be better in those circumstances than the mean and standard deviation.

THE RANGE

There is another measure of variability—one that probably gets more attention than it deserves, both in makeshift analyses and in this book. It is given attention in makeshift analyses because it is so easy to compute, and much of the space assigned to it in this book is occupied by an exposure of its fundamental weakness and an attempt to prepare you for possible differences in its interpretation.

Essence of the Concept

Sometimes we are interested specifically in the most extreme cases in a sample; on these occasions we may report its *total range*—or simply *range*—which is the distance from the lowest score to the highest. As you can see, it is extremely easy to calculate.[2] That is about its only virtue, however.

In fact, the most important feature of the range is a weakness—its extreme instability. Note that the range in Figure 3-1 is 8. Now turn to Figure 3-2. It is the same as Figure 3-1 with the exception that two scores have been moved away from the main group. Note the effect on the range. (Instead of 8, it is now 35.5 − 1.5, or 34!) Note, too, that not even *two* extremely high scores were necessary to have that effect; one would have done exactly the same thing. The fact is that the range is determined by two, and only two, scores in any distribution: the lowest and the highest. (Compare that to the mean, which is sensitive to *all* scores.) That is why the range is so easy to compute. That is also why it is so unreliable.

Appropriate Applications

We have seen that the standard deviation is analogous to the mean and is its proper companion, and we have noted a similar relationship of the interquartile range to the median. The relationship of the range to the mode is not quite as neat, but there are similarities that should help you to remember both.

One similarity is that each is the *quickest* estimate of its kind. Another closely related similarity is that each is *less stable* than alternative indices (although the range usually is much the worse in that respect because of its total dependence on only two scores). Finally, both the mode and the range are used, more often than are other statistics, to answer questions related directly and very simply to their definitions. For the mode, the question is "Which is the typical (most frequent) case?" For the range, it is "How much of the scale must be used to represent the distribution?"

SUMMARY

Two important ways of describing a sample are by its central tendency or average, which indicates the general level of the scores, and by its variability, which tells the extent to which individual scores deviate from that average. This chapter has been about the latter type of description—measures of *variability*.

You have been invited to use your previously developed understanding of measures of central tendency as an anchor for the new concepts; the latter were presented as analogs of the former. Specifically, the standard deviation is roughly analogous to the mean and is its companion statistic, the interquartile range goes with the median, and the range is similar to the mode.

The *standard deviation* is a kind of average of individual deviations from the mean of a distribution. The mean of the squared deviations is called the *variance*, and the standard deviation is the square root of the variance. In a normal distribution, the standard deviation carries the most information of all measures of variability. It also lends itself to the computation of many more advanced measures.

The *interquartile range* is the distance from Q_1 to Q_3. It embodies less information than the standard deviation but is preferred to it whenever the distribution is markedly asymmetrical, and it is the perfect companion to the median.

The *range* is the distance from the lowest score to the highest, is completely determined by those two scores, and thus carries little information. Like the mode, it

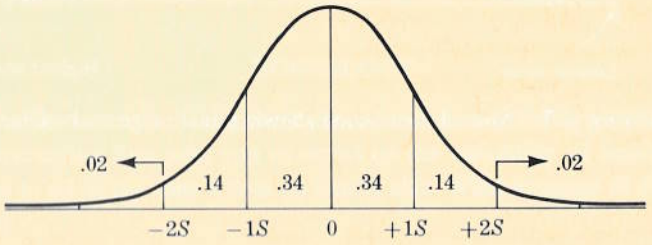


FIGURE 4-5 Normal distribution with baseline divided into standard deviation units.

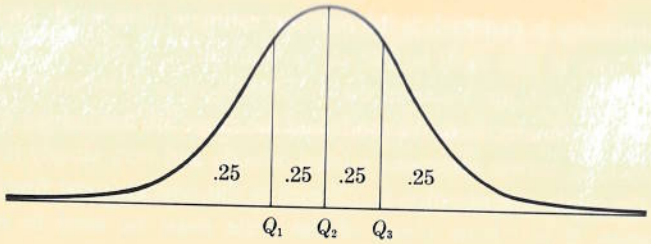


FIGURE 4-6 Normal distribution with area divided into quarters.

is both easier to compute and less stable than any other measure of its kind. It is, of course, the one statistic that gives information specifically about the distance between the highest and lowest scores in the sample.

We have seen that in a normal distribution, the mean, the median, and the mode are all at the same place. When we make a similar graphical comparison of measures of variability, the situation is not quite as simple.

In Figure 4-5, the baseline is divided into equal units—namely, standard deviations—and the areas subtended by them (.02, .14, .34, etc.) are *un*equal. The number above the line indicates the proportion of the total area that is subtended by each segment. The proportions have been rounded off because they are easier to remember in that form, and they are precise enough for our present purposes in any case (more exact proportions are given in the figure in note 3 of Chapter 5).

In Figure 4-6, the baseline is divided into unequal segments, and it is the parts of the area that are equal. And, of course, the range subtends the entire distribution, as shown in Figure 4-7.

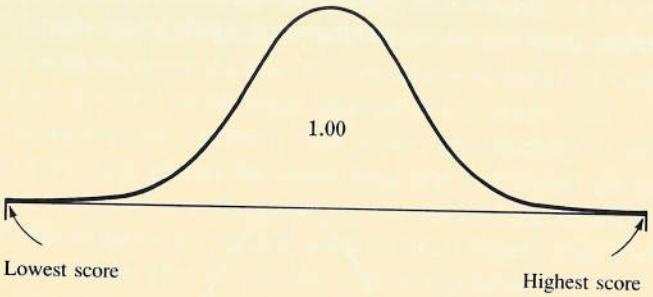


FIGURE 4-7 Normal distribution showing total range on baseline.

Sample Applications

EDUCATION

You are appointed to a committee of elementary school teachers and administrators. The committee decides to make the teaching of reading a top priority for the next two-year period. You get permission to use the annual in-service training fund for purchasing a reading program in which all elementary teachers will participate. After an extensive search, you narrow the field to two programs that have been used and evaluated on usefulness and practicality by a large number of elementary school teachers. The means of their ratings of the two programs are approximately the same. Is there any other statistic that might help you make your decision?

POLITICAL SCIENCE

You are interested in the incidence of military coups in Latin America. Specifically, you want to know whether most Latin American countries have experienced a number of coups that is close to the mean number for the region. After you have gathered your data, how do you obtain that information?

PSYCHOLOGY

You are one of a five-person team of observers sent into a home to evaluate the degree of aggressiveness among family members over a one-week period. The observers indicate that, on the average, eight aggressive acts occur per day. But there is also some disagreement. How might you quantify that disagreement?

SOCIAL WORK

You are the new director of a community fund-raising organization. The member agencies have widely varying needs, but you suspect that recently the board has been shirking its duty to investigate those needs. Specifically, you suspect that recent allocations have not been sufficiently differentiated. How might you document your case?

SOCIOLOGY

In order to make some decisions about the construction of family dwellings in your state, a construction firm asks you to report on the variability of family size. You have access to all of the state's data on family size. What measure of dispersion do you report?

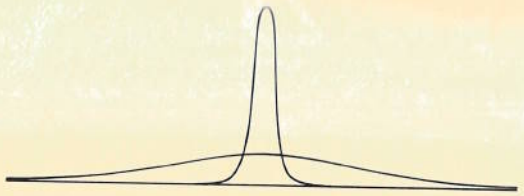


FIGURE 4-1 Two distributions with the same N's but different variabilities.

the single scholastic aptitude test mentioned in Chapter 2, those students took a *pair* of tests—one of verbal aptitude, the other of numerical. The distributions of scores on the two tests might look something like Figure 4-2.* In that diagram, the means of the two distributions have been aligned so that we may concentrate on their

* The indicated central tendencies and variabilities of the two distributions make somewhat plausible the proposition that if combined, they would form the distribution shown in Figure 2-1. However, that was not the primary consideration in their selection. Rather it was *simplicity*. We are deemphasizing computation; therefore, computations that *are* required have been made as easy as possible. It is easier to think about an interval that extends from 20 to 25, for example, than one that extends from 23 to 27, or even from 23 to 28. You will find that you can do most—possibly all—of this book's computations entirely in your head.

But do not attempt any computations unless the requisite data are readily available. For example, in Figure 4-2 do not attempt to confirm the standard deviations of 15 and 5 in the two distributions. To do so, you would need a list of the individual scores, and those have been withheld deliberately to encourage you to focus on the big picture—the configuration of the entire sample for each test and the comparison of two configurations that are very different from each other.

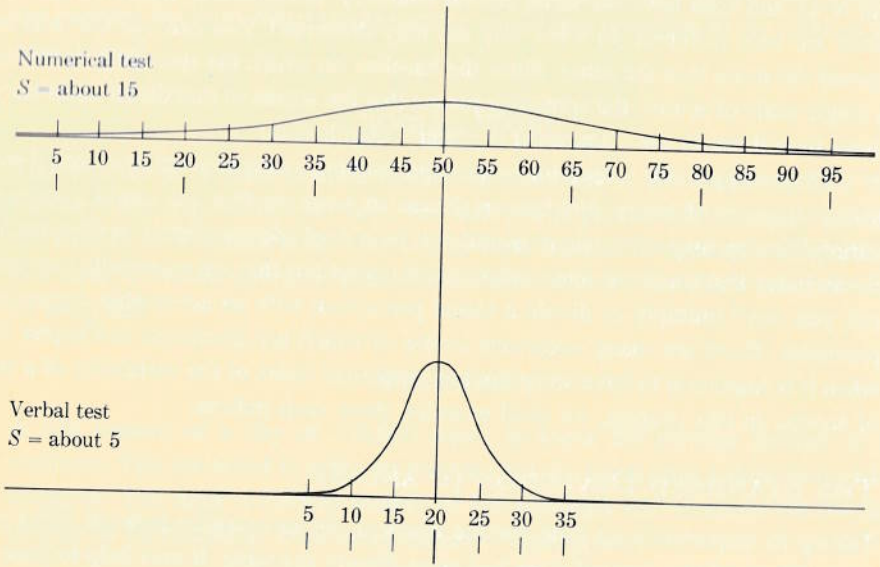


FIGURE 4-2 Two distributions with the same N's but different variabilities.

respective variabilities. As we look at the figure, we are immediately impressed by the striking difference between the dispersions of the two distributions across their baselines. But it is not enough to be impressed; we need a numerical index of dispersion (variability).

Essence of the Concept

How might such an index be devised? One way would be to compare every score in the sample to every other and to average the differences obtained thereby. But that would be entirely too cumbersome, especially with large distributions; we can get the same effect by selecting a point of reference in the middle of the distribution and measuring the distance of each individual score from that point, as in Figure 4-3. The average (mean) of those differences (without regard to sign) can also serve as an index of variability. If, for example, we were to choose the mean of our target population as our reference point, our index would be an average of the individual distances from the mean and could be obtained via the following formula:

$$\text{AD} = \frac{\Sigma|x|}{N} \tag{4-1}$$

where AD is the *average deviation*, $\Sigma|x|$ is the sum of *individual deviations* from the mean of the population, and N is the size of the population. The distance of any

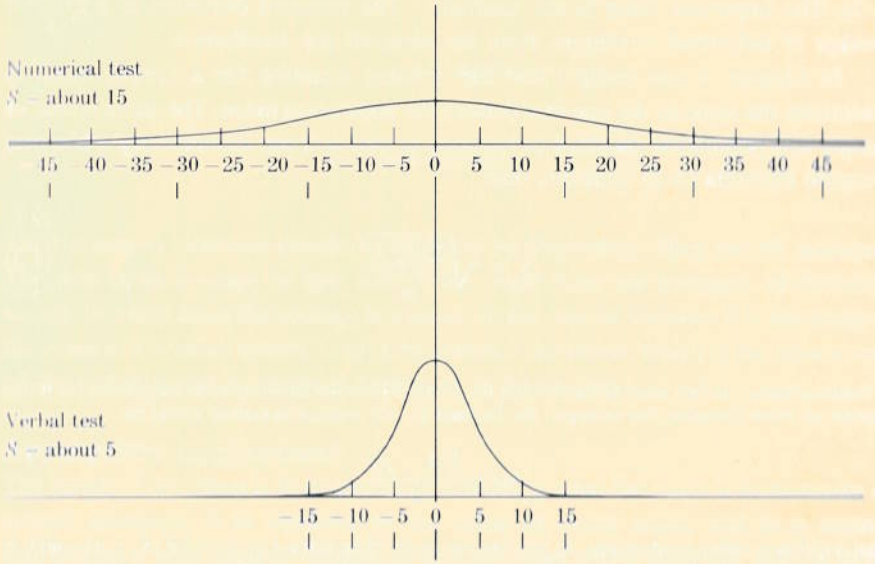


FIGURE 4-3 The distributions in Figure 4-2 with raw scores converted to deviation scores.

both of which will be discussed later. In scientific work, at least, the standard deviation clearly is the prevalent measure of dispersion.

As in Chapter 3 a calculation box (Box 4-1) provides you with an opportunity to try your hand at actually crunching some numbers. But here again, and for the same reason as before, the operations in the box are specified by the defining equation, not by a specialized calculating formula.

THE INTERQUARTILE RANGE (IQR)

A much easier statistic to comprehend (and to compute) is the *interquartile range* (IQR). Look at Figure 4-4. What you see is a negatively skewed distribution cut into four equal parts. At the upper end of each of those quarters is a point on the baseline called a *quartile* (Q)—the first quartile (Q_1) above the lowest quarter, the second (Q_2) above the lowest two quarters,* and the third (Q_3) above the lowest three quarters. The point just above the highest score in the distribution would logically be Q_4, but that expression is seldom if ever used.

Essence of the Concept

The *interquartile range*, like any measure of variability, is an interval. In this case the interval extends from Q_1 to Q_3. The calculation of the interquartile range is extremely easy: You simply find the difference between Q_1 and Q_3. That's it.[1]

The IQR, like the median, is insensitive to the precise values of most of the scores in a distribution, so when you use it instead of the standard deviation you discard information. But if the distribution is skewed, you can return some important information (namely, the direction of the skew) by reporting not only the *difference* between Q_1 and Q_3, but their exact *locations* as well, along with that of Q_2. (*If* $Q_2 - Q_1$ is longer than $Q_3 - Q_2$, the skew is negative, as in Figure 4-4; if $Q_3 - Q_2$ is the longer, the skew is positive.)

Appropriate Applications

The interquartile range is to measures of variability what the median is to measures of central tendency. Although insensitive to the exact values of many of the scores in a distribution, it is preferred to the standard deviation in the same kinds of situations in which the median is preferred to the mean—namely, when distributions are radically asymmetrical. It is the perfect companion to the median wherever the latter is properly applied.

A report of the incomes of a university faculty, for example, might be skewed by the high salaries and consulting fees of the college of business. If faculty

* What other familiar statistic has essentially the same definition as the second quartile? If you are not sure, turn back to page 28.

Box 4-1 Calculation of Standard Deviation (see Figure 4-2, verbal test)

(1) Class interval	(2) Midpoint X_m	(3) Frequency f	(4) $X_m - \bar{X}$ x	(5) $(X_m - \bar{X})^2$ x^2	(6) $f(X_m - \bar{X})^2$ fx^2
33–37	35	2	15	225	450
28–32	30	13	10	100	1300
23–27	25	32	5	25	800
18–22	20	106	0	0	0
13–17	15	32	− 5	25	800
8–12	10	13	− 10	100	1300
3–7	5	2	− 15	225	450
		$\Sigma = 200$			$\Sigma = 5100$

Column 1: Class intervals. See pages 12 and 16–21.

Column 2: Midpoints of class intervals. Remember when data are grouped into class intervals, all individuals (X) within each interval are treated as though they were at the midpoint of that interval (X_m).

Column 3: Frequency (f) of scores in each class interval.

Column 4: Deviation scores (x), which equal the differences between the midpoints (X_m) and the mean of the distribution ($\bar{X}$). In this case, $\bar{X} = 20$.

Column 5: Squares (x^2) of the deviation scores listed in column 4.

Column 6: Product of the squared deviation scores and the frequency of scores in the interval (fx^2).

$$S = \sqrt{\frac{\Sigma x^2}{n}}$$

The sum of the frequencies listed in column 3 is n, and the sum of column 6 is Σx^2.

$$S = \sqrt{\frac{5100}{200}} = 5.01$$

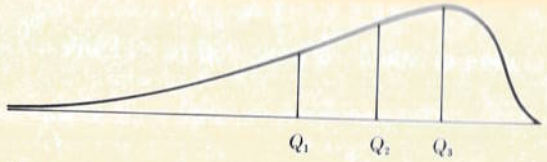


FIGURE 4-4 Negatively skewed distribution with area divided into quarters.

incomes without the college of business are distributed symmetrically, the high incomes there give the total distribution a marked positive skew, making the mean and standard deviation difficult to interpret. The median and interquartile range would be better in those circumstances than the mean and standard deviation.

THE RANGE

There is another measure of variability—one that probably gets more attention than it deserves, both in makeshift analyses and in this book. It is given attention in makeshift analyses because it is so easy to compute, and much of the space assigned to it in this book is occupied by an exposure of its fundamental weakness and an attempt to prepare you for possible differences in its interpretation.

Essence of the Concept

Sometimes we are interested specifically in the most extreme cases in a sample; on these occasions we may report its *total range*—or simply *range*—which is the distance from the lowest score to the highest. As you can see, it is extremely easy to calculate.[2] That is about its only virtue, however.

In fact, the most important feature of the range is a weakness—its extreme instability. Note that the range in Figure 3-1 is 8. Now turn to Figure 3-2. It is the same as Figure 3-1 with the exception that two scores have been moved away from the main group. Note the effect on the range. (Instead of 8, it is now 35.5 − 1.5, or 34!) Note, too, that not even *two* extremely high scores were necessary to have that effect; one would have done exactly the same thing. The fact is that the range is determined by two, and only two, scores in any distribution: the lowest and the highest. (Compare that to the mean, which is sensitive to *all* scores.) That is why the range is so easy to compute. That is also why it is so unreliable.

Appropriate Applications

We have seen that the standard deviation is analogous to the mean and is its proper companion, and we have noted a similar relationship of the interquartile range to the median. The relationship of the range to the mode is not quite as neat, but there are similarities that should help you to remember both.

One similarity is that each is the *quickest* estimate of its kind. Another closely related similarity is that each is *less stable* than alternative indices (although the range usually is much the worse in that respect because of its total dependence on only two scores). Finally, both the mode and the range are used, more often than are other statistics, to answer questions related directly and very simply to their definitions. For the mode, the question is "Which is the typical (most frequent) case?" For the range, it is "How much of the scale must be used to represent the distribution?"

SUMMARY

Two important ways of describing a sample are by its central tendency or average, which indicates the general level of the scores, and by its variability, which tells the extent to which individual scores deviate from that average. This chapter has been about the latter type of description—measures of *variability*.

You have been invited to use your previously developed understanding of measures of central tendency as an anchor for the new concepts; the latter were presented as analogs of the former. Specifically, the standard deviation is roughly analogous to the mean and is its companion statistic, the interquartile range goes with the median, and the range is similar to the mode.

The *standard deviation* is a kind of average of individual deviations from the mean of a distribution. The mean of the squared deviations is called the *variance*, and the standard deviation is the square root of the variance. In a normal distribution, the standard deviation carries the most information of all measures of variability. It also lends itself to the computation of many more advanced measures.

The *interquartile range* is the distance from Q_1 to Q_3. It embodies less information than the standard deviation but is preferred to it whenever the distribution is markedly asymmetrical, and it is the perfect companion to the median.

The *range* is the distance from the lowest score to the highest, is completely determined by those two scores, and thus carries little information. Like the mode, it

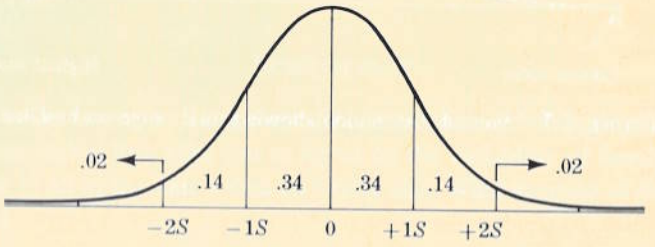


FIGURE 4-5 Normal distribution with baseline divided into standard deviation units.

5

Interpreting Individual Measures

Measurement has been defined as "rules for assigning numbers to objects to represent quantities of attributes."[1] The importance of *rules* lies primarily in the fact that the resulting assignment of numbers nearly always functions as a *communication*. Unless the receiver of a communication knows the rules by which the sender has made that assignment, however, the receiver will be unsure of its meaning. Many such rules are intuitively obvious. If you tell me that you have measured the distance from point A to point B, I shall assume that the rule you followed was to lay a ruler across the two points and note how many units (for example, inches) lie between them.

But most rules are not so obvious. What rule do you follow if you want to tell me the size of a particular circle? Do you lay your ruler across the center and the circumference and report the number of units between the two? If so, you had better label the resulting number as a "radius," so that I shall know exactly what you did to get the measurement. There are, of course, two other labels—"diameter" and "circumference"—that you could use and that would specify two rather different operations.

Those labels actually indicate different operations for measuring a circle, but because the operations have become standardized over the years, descriptions are no longer necessary to communicate precisely—the mere labels suffice. In some educational and psychological measurements, a similar standardization has occurred—though it is not as complete because the operations are much more complex. "Level of intelligence," for example, can be represented by a number, but the number is rendered more meaningful by a specification of the particular standardized test from which the number came, and even then the specificity of the operations performed falls far short of that indicated by the labels used in the measurement of circles.

Other measurements are not standardized at all. If we want to do a rat study in which one of the variables is "drive level," we have to figure out a way to measure that level; then, after doing so and running our experiment, we are obliged to describe in some detail the operations that resulted in whatever scores (measures) we have to report. It is in situations like this one that the need for specifying rules is most apparent; nevertheless, the rules are there, even when, as in the measurement of a circle, they are not explicitly stated. Were it not so, there could be no communication.

Our definition says that the rules we have been discussing are "for assigning numbers to objects." A number does not represent an object as such, but rather the quantity of some *attribute* of the object. How smart is this boy? How tall is that toy soldier? How long is the segment $A-B$ on the surface of this burial site? The numbers that we assign to the objects represent the magnitudes of their attributes.

The definition of measurement quoted above is adequate as far as it goes; it is a good general definition. But for our particular purpose, one further point needs to be made: In many measurement problems, scores can be legitimately interpreted only as *individual differences*. Indeed, individual differences may have an important effect on measurement even when the person doing the measurement is unaware of it. For example, Miss Jones has been teaching mathematics for 20 years. She tells all her students that they will be graded "on absolute standards," meaning standards inherent in the subject matter and independent of student performance in that subject matter; Miss Jones is openly contemptuous of "grading on the curve."* Indeed, we must acknowledge that if there is any discipline in which absolute standards should be used, mathematics surely must be it. But even in mathematics, when Miss Jones says that she is grading her students on absolute standards, there is reason to suspect that her statement is not entirely correct. She may require that first-year work be mastered before she passes a student to the second year, but she must have some idea, from either her own experience or that of her mentors and colleagues, what can reasonably be asked of first-year students. Her absolute standards turn out to be less absolute than she thought.

This example is from education. I am not saying that it is never possible to make an educational measurement of any one student without taking into account many other students on the same dimension. What I am saying is that in most

* If Miss Jones had completed her teacher training 10 years later, she might be talking about "criterion-referenced" versus "norm-referenced" scoring instead of "absolute standards" versus "grading on the curve," but the different terms refer to very similar concepts.

applications, the more those other measurements are utilized in the interpretation of an individual score, the more sophisticated the interpretation will be.

Consider the case of Joseph O. Cawledge. JOC's high school grades were pretty bad, but then he was a full-time athlete, and the rest of the school's program never interested him very much. How well might he perform scholastically if he were motivated to do so? To find out, we give him Central University's scholastic aptitude test.

You will recall that the CU test is divided into two parts—a numerical test and a verbal test. Joe Cawledge takes them both because they are both required of all entering freshmen. He gets a score of 60 on the numerical test and a score of 30 on the verbal. Now, what do we know about Joe?

It *looks* as though he is twice as good with numbers as he is with words. But mightn't that appearance be deceiving? Turn to page 34 and place each of Joe's two scores within the appropriate distribution in Figure 4-2. (Insert a bookmark there, for we shall be using Figure 4-2 repeatedly for the next few moments.) In these as in all interval measurements, two questions must be answered about each scale: (1) What is the scale's *point of reference*, and (2) what is its *unit of measurement*? Answers to these questions will make it possible for us to answer two parallel questions about any individual score: (1) Is it *above or below* the reference point, and (2) *how far* is it from that point?

It is customary to use the mean of some well-defined "standardization" sample as the standard reference point. You can see immediately that each of Joe's scores is above that point on the appropriate scale. But how far above? "Ten points each. They are both the same distance above their respective means." Is that what you say? Look again. One score (the numerical) is in the populous middle part of its distribution, but the other (the verbal) is way out in the tail of its distribution. On second thought, doesn't it seem to you that the latter must be a much higher score than the former?

Yes, but how much higher? Once again we find that drawings are excellent aids to comprehending basic principles, but for precision we need a numerical index. We need a single number that will tell us how far away from the mean any given score is and in what direction.

Standard Scores: The *z* Scale

The answer to the question of how much higher Joe's verbal score is than his numerical must be in terms that make it possible to *compare* an individual's scores on two different scales. To put it another way, we must find a way to put both scores onto a single scale—a *standard* scale, if you will. That probably doesn't mean much to you right now, but in a moment it will.

We have already found a common reference point for both distributions, have we not? Every distribution has a mean; therefore the mean can be used as the reference point for the standard scale that we are about to develop. Every score that

falls precisely on the mean of its own distribution will be given a score of 0, because all counting of units will start from there.

Now that we have a common reference point, all that remains is to find a common *unit* for our new scale. As Figure 4-2 demonstrates, it must be a unit that can be used to *compensate for the variability* of a distribution; in order to do that, it must *be* a measure of variability. If we were to divide Joe's numerical deviation score by a large number and his verbal score by a small one (look back at Figure 4-2), wouldn't we have a more realistic index of his position in each distribution for purposes of comparison?

Because a measure of variability *would* be large for the numerical and small for the verbal distribution, that is how we shall accomplish our purpose of converting both scores to a common scale. If you had a measurement stated originally in inches and you wanted to convert it to feet, how would you proceed? You would divide by the number of inches in a foot, would you not? Well, you do the same thing here: You divide by the number of raw-score units in a *standard deviation* (the standard deviation of that particular distribution). The result gives you the number of standard deviations in the interval being measured. The formula looks like this:

$$z = \frac{x}{S} \tag{5-1}$$

where z is a standard score, x is a deviation score, and S is the standard deviation of the distribution in which x occurs.

Let's see how it works in Joe's case. If the standard deviation of the numerical distribution is 15, and Joe's score of 60 is 10 raw-score points above the mean, his deviation score is 10 and his standard score is $\frac{10}{15}$, or 0.67. If the standard deviation of the verbal distribution is 5, and Joe's score of 30 is again 10 points above the mean, his deviation score is 10, but his standard score is $\frac{10}{5}$, or 2.00. Quite a difference, is it not? Figure 5-1 shows where Joe's two scores would be in a distribution of standard scores from both tests.

We have accomplished our purpose: We have moved scores from two different

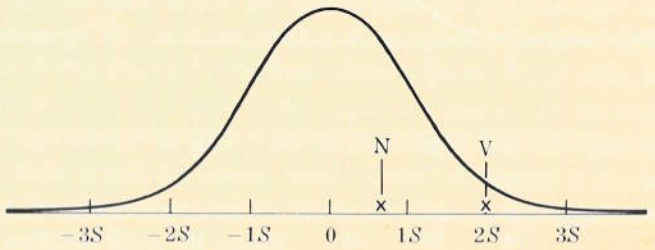


FIGURE 5-1 Distribution of standard scores from two different tests. N = numerical; V = verbal.

scales onto a single standard scale. When each was expressed in terms of its own scale, the two could not be compared; now they can be.

Other Standard Scores

Effective though it is, our standard scale still has some shortcomings in actual practice. For one thing, the negative numbers that can result are a nuisance. (If instead of 60 and 30 on the numerical and verbal tests, Joe had made a raw score of only 20 on each, his standard scores would have been -2 and 0, respectively.) Another shortcoming is that the scale unit can be awkwardly large, thus necessitating the use of (decimal) fractions.

Fortunately, however, neither of those shortcomings is difficult to overcome. To get rid of the negative numbers, we have but to add a constant to every z score; if that constant were 5, the mean score would be $0 + 5 = 5$, 1 standard deviation below the mean would be 4, 2 above it would be 7, and so on (see line B in Figure 5-2). To reduce the size of the scale unit requires nothing more than dividing it into smaller parts, which of course makes the *number* of parts larger. If we want our new units to be $\frac{1}{10}$ the size of the original, there will be 10 of the new units to every standard deviation. Combining those maneuvers (first adding 5 to every z score and then multiplying by 10) produces the scale of T scores that you see in line C of Figure 5-2.[2] The result is often referred to as a *derived* standard scale and its scores as derived scores—or sometimes "scaled scores."

Other standard scales are constructed in much the same way. In the scale for the College Boards ("CEEB scores" in the figure accompanying note 3, page 158), 5 again is added, but this time 100 is the multiplier; the resulting distribution has a mean of 500 and a standard deviation of 100. The Army General Classification Test

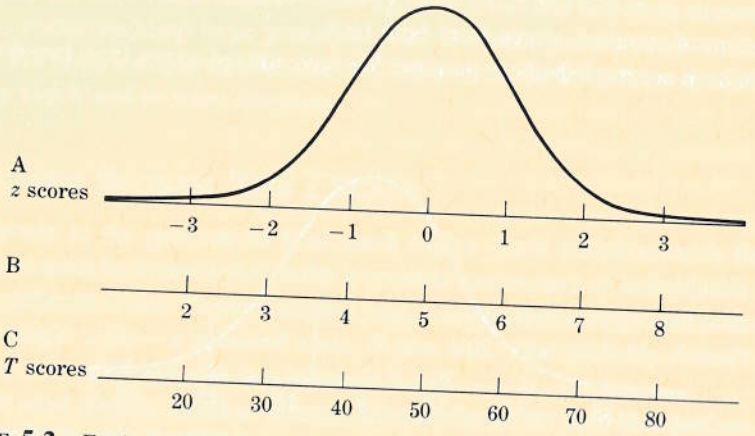


Figure 5-2 Evolution of a derived standard scale with a mean of 50 and a standard deviation of 10.

mean ("AGCT scores" in the same figure) is arbitrarily set at 100 and its standard deviation at 20. In every case, what we have is but a modification of what is still basically a standard deviation scale.

Centile (or Percentile) Scores

Standard, or scaled, scores can be used to compare an individual's scores on two different tests or to trace with one test his progress over time. Whatever else it may do, every such score compares the individual's performance to that of a standardization group. But the size of that group is commonly monstrous and its composition heterogeneous. Often what is most needed is a comparison with a group that is less cumbersome and more precisely defined.

Take Joe Cawledge's numerical performance, for example (pages 46–47). Imagine now that the test, far from being homemade by Central University as I have asked you to think of it until now, is only being borrowed by the university as a part of a national testing program. The test has already been *standardized*—that is, it has been administered to a very large number of college students, measures of central tendency and variability have been computed, scaled scores (say, T scores) have been derived, and *norm tables* have been prepared that make it possible to convert quickly any individual raw score to other kinds of scores. (You will recall that a raw score by itself is meaningless.) Imagine further that Joe's score on the numerical test is 1 standard deviation above the national mean (it was only $\frac{1}{4}$ of a standard deviation above the local mean).

That is important information in itself, and it also makes possible a comparison of Joe's present performance with his performance on the same test months or years later. What this information does *not* do, however, is to compare him with the competition at CU; more specifically, it does not compare his performance with the performances of other entering freshmen at Central University. However, each year the CU people gather their own data on the performance of their entering freshmen. Those data then appear as local norms in the form of *centiles*—or, as they are more often called, *percentiles*. Actually, of course, the data themselves are raw scores; they must be transformed somehow in order for the information to be used in this new way.

What is this new way? It is in fact very simple. Do you remember the concept of the quartile (page 38)? It is the number of quarters of the distribution that lie below the score being reported. If the point above a quarter is a quartile, how might we refer to the point above a *tenth* of the scores in a sample? *Decile* would be the parallel term. And what if we were to divide the sample into 100 equal parts? The parallel term there would be *centile*, would it not? A centile is the number of *hundredths* of the distribution that lie below the score being reported. The term *percentile* is technical slang for *centile*.

Now back to Joe. His numerical raw score was 60, which happens by coincidence to be precisely 1 standard deviation above the national mean and converts to a

TABLE 5-1 Norms for numerical test

Standard score	Percentile				
	Entering freshman	Freshman	Sophomore	Junior	Senior
88					
86					
84					
82					
80					99
78				99	98
76			99	98	97
74		99	98	97	95
72	99	98	97	95	92
70	98	96	95	92	88
68	96	94	92	88	83
66	95	92	88	84	78
64	92	88	84	78	71
62	88	84	78	71	63
60	84	79	71	63	55
58	79	72	63	55	47
56	73	64	56	48	39
54	66	57	48	39	31
52	58	49	40	31	24
50	50	41	32	24	18
48	42	33	25	18	13
46	34	26	19	13	9
44	27	20	13	9	6
42	21	14	10	6	4
40	16	10	7	4	2
38	12	7	4	2	1
36	8	5	3	1	
34	5	3	2		
32	4	2	1		
30	2	1			
28	1				
26					
24					
22					
20					

T score that is also 60. That conversion has been done by a computer at national headquarters, and neither Joe nor his counselor ever sees the raw scores. What they do see is a standard score of 60 and a norms table that allows them to convert that score into a percentile in any of several populations. (Samples from those populations are called *norm groups*, and the process of gathering data from them is often referred to as the *standardization* of the test.) Table 5-1 shows how the norms might look for the numerical test. The table is represented graphically in Figure 5-3.

The important thing to notice is that Joe has not one score but many. If it is true that his performance is meaningless until it has been compared with other performances, it is also true that its meaning is enhanced by a precise definition of the group with which he is being compared at any given time. Once that definition has been made, individuals who fit it can be tested and the distribution of their scores divided into hundredths. The result is a set of *norms* like that in Table 5-1.

Now our Joe can see how his performance compares not only with the performances of the entering American college freshmen on whom the numerical test was originally standardized, but with those of any other group that has been

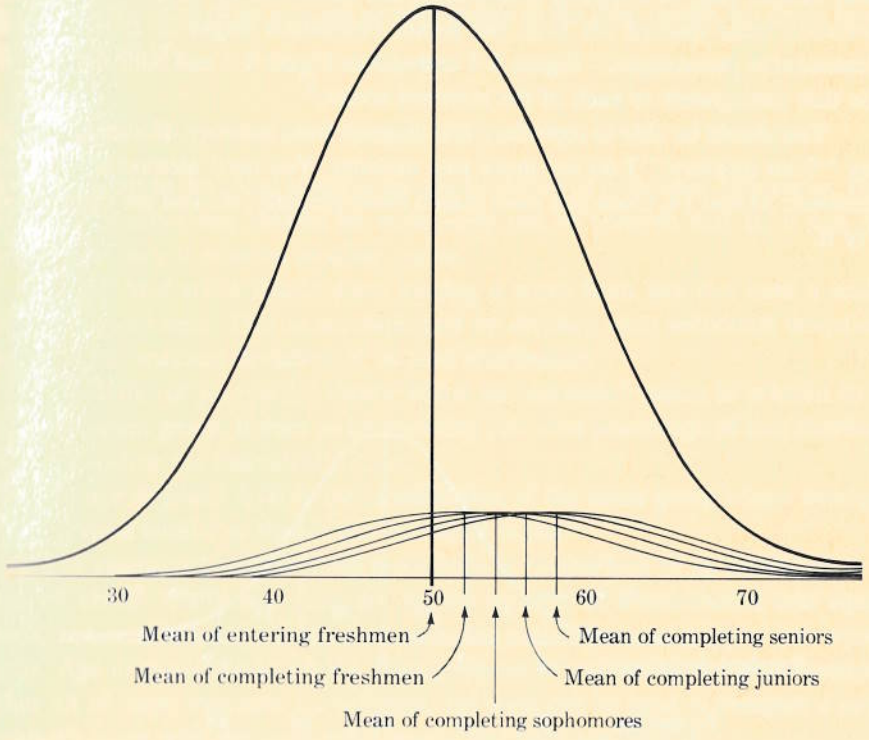


FIGURE 5-3 Distribution of four survivor groups artificially equated in size and superimposed upon original freshmen distribution.

tested. His score is better than 84 percent of entering freshmen, 79 percent of students completing their freshman year, 71 percent of completing sophomores, 63 percent of completing juniors, and 55 percent of graduating seniors. Those norm groups are students from all kinds of programs, including art, music, literature, and so forth, in which little numerical work is done. If Joe were thinking of entering an engineering curriculum, it would be appropriate to use norms derived from the testing of engineering students. Meanwhile, we should suspect that his score, compared with scores of engineering freshmen, might be well below the median and that among graduating engineering seniors it might even be near the bottom.

The small curves in Figure 5-3 have been artificially equated in size and shape. You might expect the "completing freshmen" group to be closer to the size of the "entering freshmen" group than the drawing might lead you to believe, and you would be right. You might also expect that the shapes of the distributions would change systematically from entering freshman to graduating senior; after all, it is the less able whose elimination accounts for the rising averages. But there you would be wrong. In practice, it turns out that distributions of more select groups tend to be very nearly normal.

One thing more. In Chapter 4 (Figure 4-5) we identified the proportions of the total sample that, in a normal distribution, are subtended by successive standard deviation units marked off from the mean. Percentiles might be thought of as "cumulative percentages." Figure 5-4 reproduces Figure 4-5 and adds the percentile that corresponds to each of the standard scores.[3]

You should be able to reproduce this diagram from memory. However, when I say "from memory," I do *not* mean rote memory. You need memorize only two numbers: 34 and 14. Once you have placed those properly, the rest are determined. Try it.

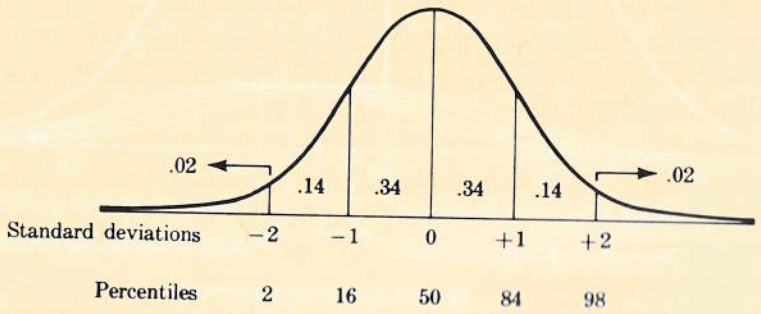


FIGURE 5-4 Normal distribution with baseline divided into standard deviation units, showing proportion of total area subtended by each standard deviation and listing cumulative percentages (percentiles).

AGE AND GRADE NORMS

A percentile tells us where an individual is in relation to a specified norm group. Another way to interpret a scaled score (or sometimes even a raw score) is to find that group which the individual's score most nearly resembles. For example, each comparison group can be taken entirely from a particular chronological age group or from a particular grade in school.

By far the best known of age norms is the *mental age*, which used to be a prerequisite to the derivation of an intelligence quotient (IQ). An individual child's score is compared with the average (usually median) scores of many chronological age groups; the age of the group whose average is closest to his determines his mental age. For example, if six-year-old Cathy does as well as an *average* six-year-old on an intelligence test, her mental age is 6; but if she does as well as the average eight-year-old, her mental age is 8, not 6.[4]

The comparison group usually used in interpreting scholastic achievement tests is the *school grade* rather than the age group. If a child's score is nearest to that made by the average entering sixth grader, his *grade level* is said to be 6.1, meaning "the first month of the sixth grade." If a child's score is equal to the average pupil in the middle of the sixth year of school, his grade level is 6.5; if his performance is most similar to that of an average seventh grader in the third month, his grade level is 7.3; and so on.

SUMMARY

Measurements in the social sciences are nearly always of *individual differences*. The report of a score can seldom be understood out of context; that is, it can be understood only in relation to other scores.

A *standard score* results from placing a score from any test onto a scale common to all tests. That is accomplished by dividing each individual deviation score by the standard deviation of its own distribution.

A *centile* (or *percentile*) shows where an individual stands in relation to a specified norm group. It does so by reporting just what percentage of that distribution falls below the individual's score.

An *age* or *grade level* is the identification of the norm group whose average score is most similar to that of the individual tested.

The essence of each of those procedures is *comparison* of an individual to a defined group. A standard scale divides the baseline of a distribution into equal parts, and percentiles divide the area under the curve (the total sample) into equal parts. Age or grade norms are the average scores of various age or grade groups. Thus all of these measures are different, but each can be used to compare a new subject's score with the many other scores that have preceded it.

Sample Applications

EDUCATION

A fourth-grade student in your school has been having considerable difficulty in several subjects. His teacher has tried various approaches, but although the student can read words at grade level, he does not seem to remember what he reads, and although he can complete simple addition and subtraction problems, he cannot complete problems that involve several steps. The teacher refers the student to you, the school psychologist, and you give the student a battery of tests to determine his specific strengths and weaknesses. After the tests have been scored, what should you do before making your report?

POLITICAL SCIENCE

Many of the concepts in political theories are so complex that they cannot be expressed in terms of simple indices. As a result, researchers often create complex indices by summing several separately derived scores, each of which represents a different dimension of the concept. Imagine, for example, that you want to measure civil strife. A simple index (such as the frequency of riots) would not suffice for measuring such a complex concept, but adding the scores on labor-hours lost per strike, assassinations per 1000 population, and other such indices derived through different measuring procedures would not work either because each is measured in a unit different from that used in measuring each of the others. What can you do?

PSYCHOLOGY

June is a year-old infant, the product of a difficult pregnancy and premature delivery. At birth, the attending physician feared that June's development would be delayed or fragmented (high in some areas, low in others).

You are asked to evaluate the quality of June's overall development. You are concerned about her growth in three different areas: intellectual, social, and psychomotor. You administer separate tests corresponding to those areas and get three measures (scores). However, each of the three distributions has a different mean and standard deviation from the other two. How can you compare June's development in the three areas?

SOCIAL WORK

The state personnel office conducts an examination to establish a hiring register for social workers. It consists of a battery of tests concerning such functions as client assessment, client treatment, and community resources. Each test has its own mean and standard deviation, but all have been standardized on the same population. How can the personnel office inform you about your relative performance on the various parts of the examination?

SOCIOLOGY

A study of authoritarianism in the various departments of your university has just been completed, and a friend of yours has access to the results. You are about to enroll in a required English course, but because you yourself harbor authoritarian attitudes you want to postpone enrollment unless the course is offered this semester by a professor with attitudes similar to yours. What information do you request from your insider friend?

6

Correlation

There are many times, both in basic science and in professional practice, when we want to know the *relationship* between one thing and another. Indeed, all of science is concerned with such relationships, and without knowledge of them, professional practice could never check up on itself.

Take our scholastic aptitude test, for example (pages 12ff). If we merely assume that it is measuring what we want it to, then much money and even more time and effort may be spent in vain. If, on the other hand, we define the test's effectiveness in terms of its prediction of success in college, then we have a way of checking up on it. Once we discover how closely the scores are related to some criterion of success, we can judge the test's effectiveness.

What we need, then, is an index of relationship—a number that when low indicates a low degree, and when high a high degree of relationship between two variables. We need a *coefficient of correlation.*

Imagine now that the two variables in which we are interested are the height and the weight of a population of toy soldiers. Imagine further that all the soldiers are exactly the same *shape* and that they differ in size and therefore in weight. (This is, of course, an unnatural situation; I am using it because it enables us to focus our

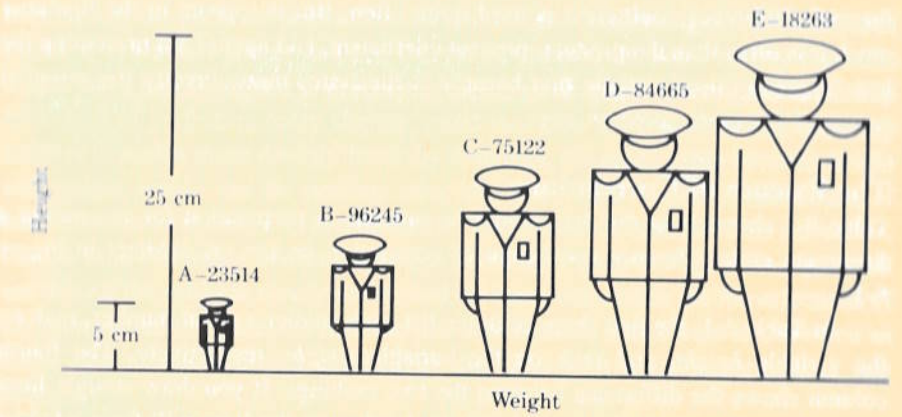


FIGURE 6-1 Sample of a population in which height and weight have a perfect positive correlation. (The label above each toy soldier is a serial number for identification.)

attention exclusively on the two variables of our present concern.) Figure 6-1 pictures five soldiers. If we were actually doing the computation, we would need a much larger set than five, but for learning the concept, the smaller number is better.

Just by looking at the five soldiers, what would you say the relationship is between height and weight? You notice immediately that the short soldiers are light in weight and that the tall ones are heavy. How would you describe that relationship? Strong? Directly proportional? Perfect? If you were asked to place the strength of the relationship on a scale from .00 to 1.00, you would have to place it at 1.00, because it *is* perfect.[1]

A coefficient of correlation provides just such a scale—that is, a scale with limits of .00 and 1.00—except that it carries information not only about the *strength* of the relationship but also about its *direction*; some correlations are positive and some are negative.

There are several kinds of coefficient. The easiest to comprehend is the rank-difference coefficient; the most useful (and most used) is the product-moment coefficient. We shall examine both.

THE RANK-DIFFERENCE COEFFICIENT (ρ)

Near the beginning of Chapter 4, "Measures of Variability," I introduced you to a statistic that you will probably never see again. I did so because that statistic, the average deviation, was the best device available for building an understanding of the concept that lies behind the most used of all measures of variability, the standard deviation. What I am about to do with correlation is not quite as drastic as that, for

the rank-difference coefficient is used quite often. But it appears in the literature much less often than the product-moment coefficient, and again I am presenting the less frequently used statistic first because it illustrates more directly the essential nature of the more common one.

The Essence of Correlation

Table 6-1 shows how the data would be ordered in preparation for computing a Spearman rank-difference coefficient of correlation on the toy soldiers in Figure 6-1.

In Table 6-1, the first three columns list each soldier's serial number, rank on the variable *height*, and rank on the variable *weight*, respectively. The fourth column shows the difference between the two rankings. If you draw straight lines between equal ranks in the second and third columns, the lines will form a ladder pattern. That has been done in Table 6-1. (In a moment, we'll look at a similar table that forms a different pattern.) Notice, too, that *all the rank differences are zero*. Keep that in mind while looking at the formula for the *rank-difference coefficient*, labeled ρ (rho):

$$\rho = 1 - \frac{6\Sigma D^2}{n(n^2 - 1)} \qquad (6\text{-}1)$$

where ρ is Spearman's rank-difference coefficient, D is the difference between the two ranks of a single subject, and n is the number of subjects—that is, the number of *pairs* of measurements. We are not interested in the computation as such[3]; the formula is included here only to demonstrate that the rank differences are in the numerator of a fraction that is subtracted from 1.00. That means that the larger the rank differences, the smaller the ρ, down to a limit of .00 (after that, still larger differences contribute to rising *negative* coefficients, up to a limit of −1.00).

Any correlation coefficient carries information about two aspects of a relationship: its *strength*—measured on a scale from zero to unity—and its *direction*—

TABLE 6-1 Ordering of data for computing rank-difference correlation of heights and weights of toy soldiers in Figure 6-1

Identification of subject	Rank on variable X (height)	Rank on variable Y (weight)	Difference between ranks, D	Square of difference, D^2
E-18263	1	1	0	0
D-84665	2	2	0	0
C-75122	3	3	0	0
B-96245	4	4	0	0
A-23514	5	5	0	0
				$\Sigma = 0$

indicated by the presence or absence of a minus sign. Consider again our example in Figure 6-1. I asked you to imagine that all of the subjects were exactly the same shape, and I had attempted to draw them all the same shape. But my draftsmanship falls short of perfection (only slightly, mind you!), so if we were to cast those toy soldiers in molds made directly from my drawings, the relationship between height and weight would *not* be perfect, after all. You might expect a coefficient of correlation to be sensitive to that difference between perfection and near perfection, but the rank-difference coefficient is not. Examine Figure 6-1 again, and you will see that all the subjects clearly are ranked the same even when the imperfections of my drawing are recognized. I mention that here because later I want to show you an index (the product-moment coefficient) that *does* take those imperfections into account.

As for the negative sign that accompanies some coefficients, there is an important point to be made (or rather, emphasized, for it has already been made). *The size* (strength) *of a coefficient is entirely independent of its direction* (positive or negative). Although the formula for the rank-difference coefficient does not encourage it, we should think not of a single continuum from negative 1.00 through .00 to positive 1.00 but rather of two separate dimensions—one negative, one positive—each of which begins at .00 and ends at 1.00. A correlation of +1.00 is no bigger (stronger) than a correlation of −1.00.

Negative Correlation

What does it mean to say that two variables are *negatively* correlated? Let us look again at the variables of height and weight, but this time in a population quite different from the toy soldiers that we examined in Figure 6-1 and Table 6-1. Consider now a human population in which the shortest individuals are the heaviest and the tallest are the lightest. (Such an arrangement is beyond your experience, but not your imagination.) Figure 6-2 is a rendering of a sample of five people drawn

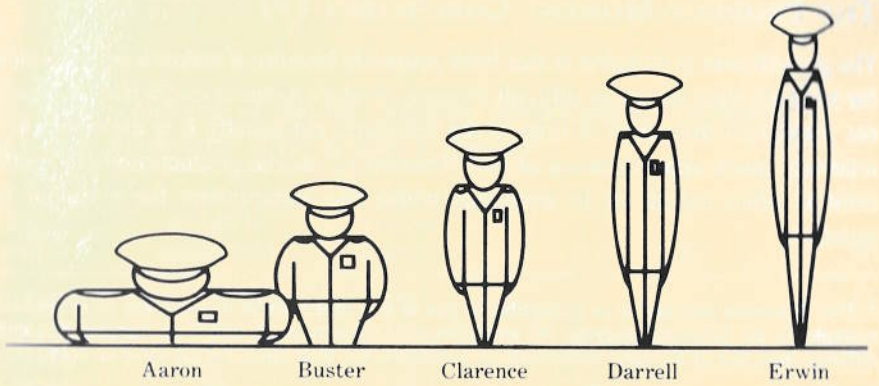


FIGURE 6-2 Sample of population in which height and weight have a perfect negative correlation.

TABLE 6-2 Ordering of data for computing rank-difference correlation of heights and weights of men in Figure 6-2

Identification of subject	Rank on variable X (height)	Rank on variable Y (weight)	Difference between ranks	Square of difference
Erwin	1	5	4	16
Darrell	2	4	−2	4
Clarence	3	3	0	0
Buster	4	2	2	4
Aaron	5	1	4	16
				$\Sigma = 40$

from such a population. Table 6-2 reveals the different pattern I promised a moment ago. If you look at the lines drawn between equivalent ranks, you will find that a star pattern emerges instead of the ladder that you see in Table 6-1.

If you sum the squared rank differences, you will find a marked contrast with the zero that comes from Table 6-1. (Remember that D^2 is in the numerator of a fraction that is *subtracted* from 1.00 to obtain the correlation coefficient.) But my drawing is a considerably less adequate representation of a perfect relationship than was Figure 6-1. Does the ρ coefficient detect my inadequacy this time? No. Nothing will affect it until the *ranks* change. In Table 6-2, the correlation of height to weight is a full -1.00, even though the weights of Darrell and Erwin, for example, are perceptibly farther apart than their heights.*

In other words, information is lost when we convert interval measurements into rank orders.[3] That difficulty can be overcome, at considerable cost in computational labor, by using a different coefficient of correlation. The next section will aim at an understanding, without computational labor, of that coefficient.

THE PRODUCT-MOMENT COEFFICIENT (r)

The ρ coefficient is included in this book primarily because it makes a better vehicle for teaching than the more difficult Pearson *product-moment coefficient*. You will see ρ reported from time to time in the literature, but usually it is used only as a relatively quick approximation of the "Pearson r," as the product-moment coefficient is often called, or in situations where the assumptions for r cannot be satisfied.[4]

* That difference may not be as perceptible to you as it is to me, so let me give you an exaggerated example. In the following diagram, the arrangement of subjects A, B, and C on variable X is quite different from that of the same individuals on variable Y, but their ranks are the same on both:

X: A B C

B C

The Meaning of "Product Moment"

To begin our discussion of this more sophisticated index, let us look again at those toy soldiers (Figure 6-1). Once again we construct a table, and indeed it is of the same general form as Table 6-1, but this time instead of *ranks*, we enter ordinary interval scores that also tell how far apart the subjects are on each variable (Table 6-3). (See the footnote on the facing page.)

Don't be intimidated by that table. One reason it is so large is that information must be collected for the computation of two standard deviations. (The x^2 and y^2 columns are used exclusively for that.) The reason for including that computation should be clear, based on the discussion of standard scores in Chapter 5. The formula for r is:

$$r = \frac{\Sigma xy}{nS_xS_y} \tag{6-2}$$

where r is the Pearson coefficient, Σxy is the sum of the cross products of deviations from the means of X and Y distributions, n is the number of such products (or of subjects or of pairs of observations), and S_x and S_y are the standard deviations of the two distributions.

The main idea behind this formula is that when X and Y scores are arranged in parallel, so to speak, the product of the corresponding deviation scores (xy) is maximally large. This is so because in such an arrangement the largest deviation scores—both positive and negative—occur together in the table and thus are multiplied together. Compare the "ordered" and "random" parts of Table 6-4 and you will see what I mean.

Close examination of the two parts will reveal that the scores are the same in both but that they are arranged differently. On the left, the arrangement is perfectly ordered, as the drawing of toy soldiers in Figure 6-1 would suggest; on the right, the arrangement of the Y scores is random, which means that the relationship between X and Y is also random.

TABLE 6-3 Ordering of data for computing product-moment correlation of heights and weights of toy soldiers in Figure 6-1

Serial number	X (height in centimeters)	Y (weight in grams)	x	y	x^2	y^2	xy
E-18263	25	250					
D-84665	20	160					
C-75122	15	90					
B-96245	10	40					
A-23514	5	10					

TABLE 6-4 Effect of ordering on sum of cross products (Σxy)

	Ordered					Random				
	X	Y	x	y	xy	X	Y	x	y	xy
	25	250	10	140	1400	25	160	10	50	500
	20	160	5	50	250	20	40	5	−70	−350
	15	90	0	−20	0	15	10	0	−100	0
	10	40	−5	−70	350	10	250	−5	140	−700
	5	10	−10	−100	1000	5	90	−10	−20	200
Σ	75	550			3000	75	550			−350
$\bar{X}$	15	110				15	110			

Now see what that arrangement does to the correlation coefficient. On the left, every large positive deviation score gets multiplied by another that is similarly large and also positive; whereas on the right, a large positive deviation may be neutralized by a low multiplier, and *any* positive deviation may be multiplied by one of an opposite sign, thereby yielding a negative product. The result is that with the random arrangement, the sum of the *cross products* (Σxy)—and hence the correlation coefficient—is always small. (In fact, it varies from zero only by chance!) In terms of the weightless beam concept introduced on page 25, you could say that the deviation scores on the two variables are often either close to the fulcrum or on opposite sides of it, whereas in a more substantial correlation, x and y scores for a given individual are consistently together—frequently far from the fulcrum—so that they can apply an extremely large torque, or *moment*, to the beam. Hence the term *product moment*—a reference to the product of the moments, which is largest when all of the scores are perfectly ordered. Our defining equation for the Pearson product-moment coefficient is given in Formula (6-2). Box 6-1 shows how an r might be calculated using that formula.

The Scatterplot

There is another way of looking at correlation that gets directly at the fundamental idea. I refer to the *scatterplot*. A scatterplot is an approximation of a three-dimensional frequency distribution. It is a surface on which both the X and Y scores of each individual can be represented by a single point. If, for example, we were correlating scholastic aptitude test scores with first-year grade point averages and if

BOX 6-1 Calculation of Correlation Coefficient, r (data not discussed in text)

(1) Sugar intake X	(2) $X - \bar{X}$ x	(3) Time on task Y	(4) $Y - \bar{Y}$ y	(5) xy
45	15	22	−28	−420
43	13	18	−32	−416
42	12	50	0	0
40	10	64	14	140
37	7	56	6	42
36	6	44	−6	−36
35	5	41	−9	−45
30	0	59	9	0
25	−5	36	−14	70
24	−6	73	23	−138
23	−7	31	−19	133
20	−10	78	28	−280
18	−12	69	19	−228
17	−13	27	−23	299
15	−15	82	32	−480
				$\Sigma = -1359$

Column 1: Sugar intake as percentage of total carbohydrates ingested (X).

Column 2: Deviation scores (x) of sugar intake, which equal the differences between the individual scores (X) and the mean of the X scores ($\bar{X}$; see sample calculation on pages 26–27). In this case, $\bar{X} = 30$.

Column 3: Time on task as percentage of time available (Y), used as an index of hyperactivity. (A *low* score indicates hyperactivity.)

Column 4: Deviation scores (y) of time on task, which equal the differences between the individual scores (Y) and the mean of the Y scores ($\bar{Y}$). In this case, $\bar{Y} = 50$.

Column 5: Product of deviation scores, x and y. Notice that a negative product occurs only where a negative x is paired with a positive y or vice versa.

$$r = \frac{\Sigma xy}{nS_xS_y}$$

The number of xy products is n. The sum of column 5 is Σxy. $S_x = 10$ and $S_y = 20$. (See page 39 for a sample calculation of the standard deviation.)

$$r = \frac{-1359}{(15)(10)(20)} = -.45$$

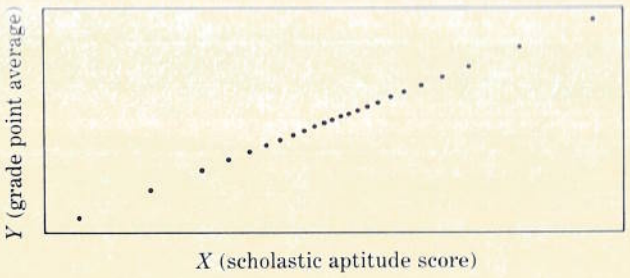


FIGURE 6-3 Scatterplot of perfect positive correlation.

all students' test scores were precisely proportional to their subsequent scholastic performance, the scatterplot might look like Figure 6-3. Each student is represented by a single point (here a black dot) that locates the student on two dimensions, X and Y.

A third dimension is formed whenever two or more individuals occupy the same point on that two-dimensional surface. Because the third dimension is extremely difficult to represent on the flat pages of a book, I have, in places where individuals tend to pile up, separated them enough so that you can see all of them.

In the case represented by Figure 6-3, the students' scores on X and Y form a straight line (called a *regression line** because it embodies the mathematical regression of Y on X) with no scatter. If the scores were all in standard units (discussed in Chapter 5), the regression line in a perfect positive correlation (e.g., the zero scatter of Figure 6-3) would necessarily have a slope of 1.00, and any smaller correlation (e.g., the moderate scatter of Figure 6-4) would produce a slope smaller than 1.00. When scores are in standard units, the smaller the slope, the

* *Regression* in this context is a functional relation of correlated variables; in this application, the correlated variables are labeled X and Y. The *regression line* yields an approximation of the mean value of Y for any specified value of X: a line of best fit. For example, in Figure 6-11B, if X is 1.0, the mean value of Y is about 0.75; and if X is 2.0, the mean of Y is about 1.5. The slope of the line is 0.75. But in Figure 6-11C, if X is 1.0, the mean of Y is 0; if X is 2.0, the mean of Y is still 0. The slope of that line is 0.00.

The slope of a line in a coordinate plot like Figure 6-3 is the ratio of the amount of change on the ordinate (Y) to the amount of change on the abscissa (X). If Y increases $\frac{1}{2}$ unit for every unit increase in X, the slope is 0.5. If while following a regression line with your pencil you find that you have to move 4 units on the X-axis for every 1 you cover on Y, the slope is 0.25—providing that the changes on both variables are positive. The same *degree* of slope could occur in a *negative* direction if Y were *dropping* $\frac{1}{4}$ unit with every 1-unit rise in X.

The relation between scatter and slope is explained later (pages 70–73). For an interesting historical account of the development of regression and correlation concepts, see J. P. Guilford and B. Fruchter, *Fundamental Statistics in Psychology and Education*, 6th ed. (New York: McGraw-Hill, 1977).

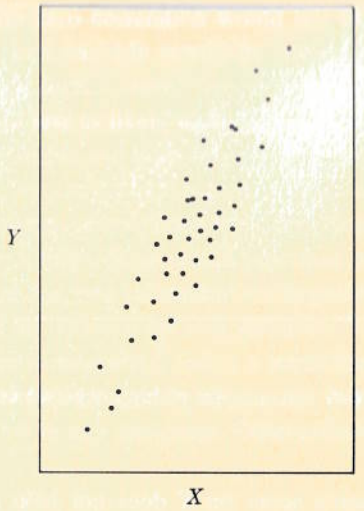


FIGURE 6-4 Scatterplot of high positive correlation.

greater the scatter, and the smaller the scatter, the greater the slope. When scores are *not* in standard units, the slope is determined mostly by the whim of the graph maker. We shall return to this issue on pages 70–73, after we have finished our discussion of scatterplots.

If, as in real life, a correlation is *not* perfect, there is always some scatter. But if the X-Y relationship is extremely strong, most of the deviations from the regression line are small (see Figure 6-4); we can make very accurate predictions of Y scores from a knowledge of X scores. At the other extreme is the random X-Y relationship depicted in Figure 6-5, where there is no tendency for a low Y to be associated with a low X or a high Y with a high X, and as you can see, there is a near maximum of

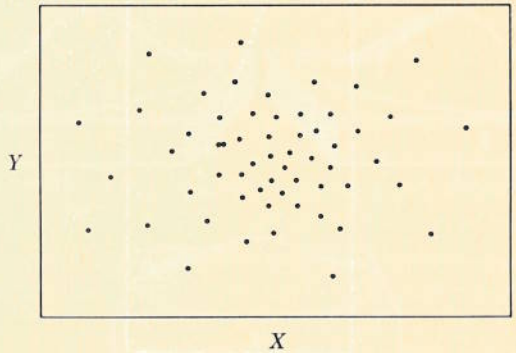


FIGURE 6-5 Scatterplot of zero correlation.

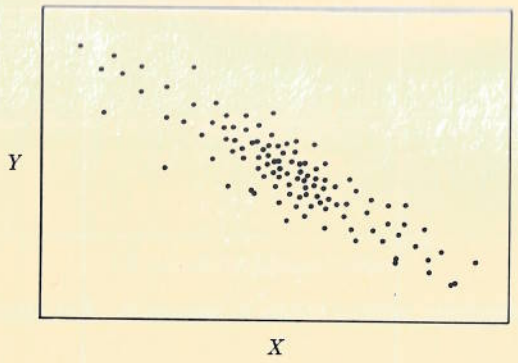


FIGURE 6-6 Scatterplot of high negative correlation.

scatter. Knowing a person's score on X does not help at all in estimating that person's score on Y.

Negative correlations behave in exactly the same way as the positive ones except that the regression line slopes *down* from left to right instead of up. Figures 6-6 and 6-7 are examples.

Figures 6-3 through 6-7 have been contrived to represent perfect positive, high positive, zero, high negative, and perfect negative correlations, respectively. Either of the perfect correlations would enable us to predict Y precisely from a knowledge

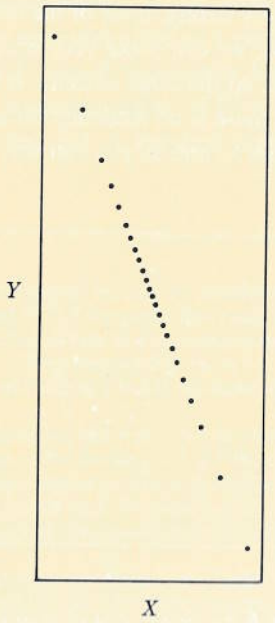


FIGURE 6-7 Scatterplot of perfect negative correlation.

of X, and vice versa. The zero correlation would tell us that a knowledge of an individual's position on one variable would be useless in predicting where that individual is on the other. Such a correlation would say that high scorers on the scholastic aptitude test are just as likely to flunk out of school as low scorers. The other correlations (high positive and high negative) would not ensure that we could make perfect predictions, but they would significantly reduce our errors, should we wish to try predicting college performance from test scores.

Figure 6-8 attempts a three-dimensional rendition of five scatterplots. It includes two perfect correlations, two substantial but imperfect ones, and one plot that reveals no relationship at all between the two variables. You can see that when the correlation is perfect there is *no* scatter, and as a result the scores are piled high. In stark contrast, a zero correlation spreads scores out over most of the horizontal surface, with relatively few in any one place. Other correlations approximate those

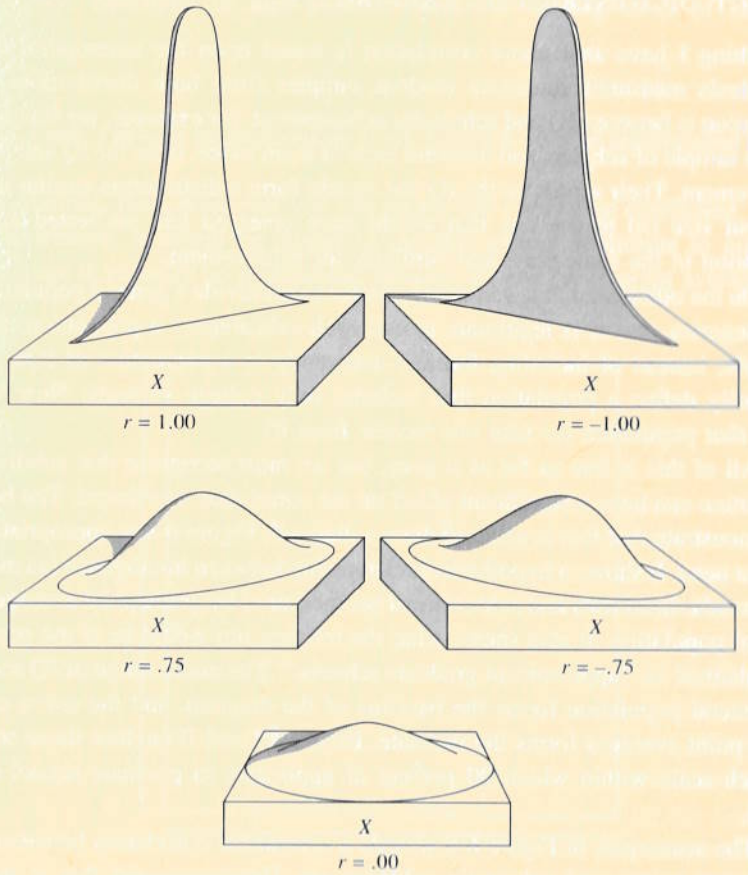


FIGURE 6-8 Three-dimensional renderings of five scatterplots.

two in varying degrees. The relationship represented by $r = .75$, for example, is far from perfect; but it is a long way from zero. The relationship that it reveals may be interesting for itself, and it also may have practical consequences. One illustration of the latter is the fact that with a correlation that high and a knowledge of X you could substantially reduce your errors in predicting Y.

A distribution need not be normal to support calculating an r. It need only be unimodal and fairly symmetrical. And the regression line must be approximately straight. Always construct a scattergram, whether or not you calculate a coefficient of correlation. The Pearson r, based as it is on standard scores, guards against distortions of scale, but the scatterplot may detect configurational anomalies that would be missed without it—asymmetry, for example, a curved regression line, or even more than one line. The coefficient will not catch any of these.

EFFECT OF RESTRICTED VARIABILITY

Everything I have said about correlation is based upon the assumption that the individuals measured constitute random samples from both distributions. If the correlation is between IQ and scholastic achievement, for example, we should select a large sample of subjects and measure each of them twice, once for IQ and once for achievement. Their scores on the IQ test should form a distribution similar in every way but size (n) to the one that would have emerged had we tested an entire population in the same way, and similarly for achievement.

On the other hand, it is sometimes useful to subdivide a general population into subcategories, and it is legitimate to call each subcategory a population, too. For example, instead of including literally everybody in a population of IQs, we can arbitrarily define a population that includes only college students. Now we will study *that* population (or take our sample from it).

All of this is fine as far as it goes, but we must recognize that subdividing a population can have a significant effect on the correlation coefficient. The best way to demonstrate that fact is again to draw a diagram; Figure 6-9 is appropriate to our present need. It shows a hypothetical relationship between intelligence (as measured by a standardized test) and achievement (as measured by grade point average) in the general population. It also shows what the relationship would be if the population were defined as "applicants to graduate schools." The entire range of IQ scores for the general population forms the baseline of the diagram, and the entire range of grade point averages forms the ordinate. Brackets I and II enclose those segments on each scale within which 90 percent of applicants to graduate school may be found.

The scatterplot in Figure 6-9 reveals a substantial correlation between intelligence test scores and scholastic achievement. However—and this is the point toward which this discussion has been directed—a plot of the area marked off by

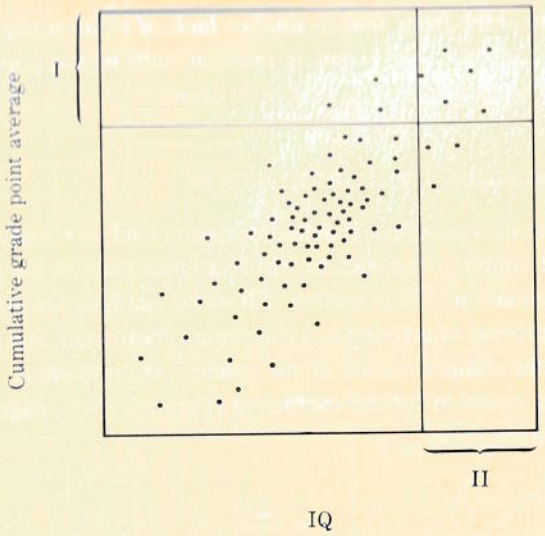


FIGURE 6-9 Scatterplot of high positive correlation showing effect of restricted variability.

segments I and II reveals no relationship at all. Figure 6-10 is an enlargement of that area.

Be cautious, then, whenever you interpret a low correlation coefficient. In the above illustration, it would be proper to say that among applicants to graduate schools, there is virtually no relationship between IQ and grade point average, but

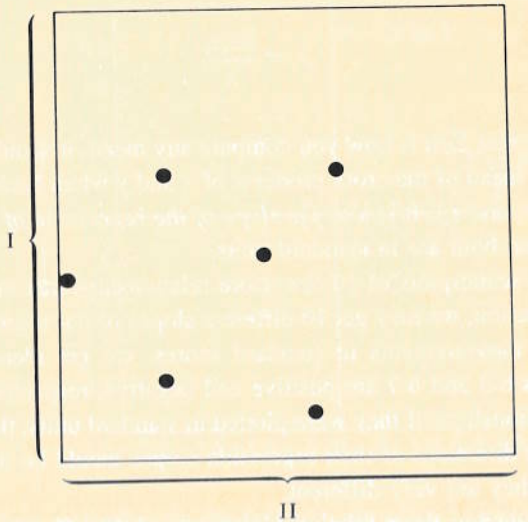


FIGURE 6-10 Restricted area (enlarged) of scatterplot in Figure 6-9.

be careful *not* to infer from that *a similar lack of relationship in the general population*. A correlation coefficient is relevant only to the population that you study directly or from which your sample is drawn.

STANDARD SCORES IN CORRELATION

While introducing the concept of standard deviation back in Chapter 4, I mentioned that "among the common statistics to which it lends itself is the product-moment coefficient of correlation." This section will show you how the standard deviation functions in the process of arriving at a correlation coefficient, and at the same time it will deepen your comprehension of the *concept* of correlation.

The defining equation for the coefficient [Formula (6-2), page 61] is repeated here for your convenience:

$$r = \frac{\Sigma xy}{nS_xS_y}$$

Now according to Formula (5-1), a standard score (z) is a deviation score divided by the standard deviation of the distribution in which it is found; in other words, it is a deviation score in standard deviation units. In Formula (6-2), we have two such scores:

$$\frac{x}{S_x} \text{ and } \frac{y}{S_y} = z_x \text{ and } z_y$$

so we can rewrite the formula in terms of standard (z) scores:

$$r = \frac{\Sigma z_x z_y}{n} \qquad (6\text{-}3)$$

If you remember that Σ/n is how you compute any mean, it should now be clear to you that r is the mean of the cross products of x and y when *both are expressed in standard units* (z_x and z_y). It is also *the slope of the regression of the Y variable on the X*, again when both are in standard units.

If we make scatterplots of 10 raw-score relationships that are all of the same strength and direction, we may get 10 different slopes of our regression lines; if we first convert all measurements to standard scores, we get identical slopes. The slopes in Figures 6-3 and 6-7 are positive and negative, respectively, but they are both perfect relationships; if they were plotted in standard units, the degree (though of course not the direction) of their regression slopes would be the same. Without the conversion, they are very different.

Figure 6-11 depicts those relationships in a slightly different way. Of the four lines that cross diagrams A, B, and C, one line (the perpendicular) is at the mean of

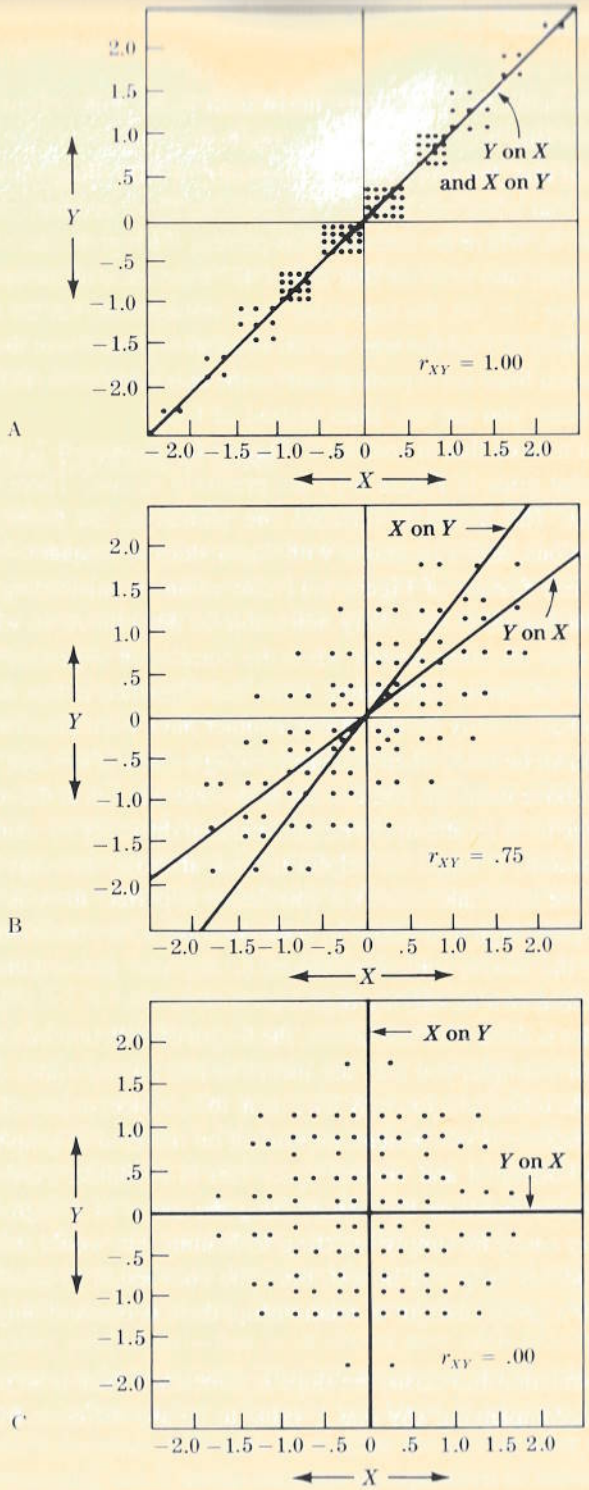


FIGURE 6-11 Three scatterplots on standard grids (all scores in standard units): (A) one regression line, (B) two regression lines, and (C) two regression lines. See also Figure 6-8.

all the X scores and another line (the horizontal) is the mean of the Y's. The third and fourth lines (the bold diagonals) are regression lines.

The reason that you don't see four lines at $r = 1.00$ (diagram A) and $r = .00$ (diagram C) is that when one line corresponds exactly to another, you can see only one. In diagram A, where the correlation is perfect, the regressions of Y on X and of X on Y are merged into one line that is just 45° from each of the two baselines. In diagram C, where $r = .00$, the regression lines are separate, but the regression of Y on X is at the mean of Y all the way across, and that of X on Y is at the mean of X; so the two regression lines are superimposed on the two mean lines. (They *are* the two mean lines.) Thus, you see two lines instead of four.*

Have you noticed that in contrast to Figures 6-3 through 6-7, pages 64–66, all of the scatterplot grids in Figure 6-11 are precisely square? There is a very good reason for that: The X and Y variables are plotted in the same units, namely, standard deviations, and a rectangle with equal sides is a square.

There is one feature of Figure 6-11 that could be misleading if it were not explained. That feature is especially noticeable in diagram A, in which it appears that there is, after all, some scatter when the correlation is perfect ($r = +1.00$).

That is an erroneous impression. What has happened is that dots that should have been stacked directly on top of one another have been dispersed horizontally so that they can all be seen; when you are viewing a stack of identical objects from a point directly above them, all you can see is the one that sits at the top of the stack. Figure 6-8, page 67 is an attempt to circumvent that difficulty by rendering the plots in three dimensions. It might be helpful to look at those drawings again.

You can see there that when the correlation is perfect there is *no* scatter, and the scores are piled high as a result. In stark contrast, a zero correlation spreads out over most of the board, with few scores in any one place. Other correlations approximate those two in varying degrees.

If your plot is done with raw scores, the length of each unit on either dimension is arbitrary. For example, you may use inches or feet on one dimension and ounces or pounds on the other, and you may let 1 inch, foot, ounce, or pound be represented by whatever distance is convenient, depending on the kind of graph paper you are using and the horizontal and vertical space that is available. Furthermore, if you wanted to deceive someone by manipulating the slope of a regression line, you could do it very easily by simply selecting scale units that would suit your purpose. But standard scores keep you honest; once the conversion to standard scores has been made, for a given strength of relationship, there is one and only one degree of slope.

When no relationship exists, the slope is zero; since there is no reason to expect the Y scores accompanying any one X value to be any different from those associated with any other, the best estimated central tendency of the Y's at any given value of X is simply the mean of *all* the Y's in the entire distribution. So as our regression line moves from left to right on the X-axis, it moves not at all on the Y; which is to say, the slope of the regression line is zero, as in Figure 6-11C. For a perfect relationship, it is 1.00, as in Figure 6-11A. All other relationships are somewhere between these two extremes. (Remember that a correlation of −1.00 is as perfect as a correlation of 1.00.)

In short, one meaningful interpretation of r is as *the slope of the regression of Y on X when both are laid out in standard units.*

* If you will turn Figure 6-11 over on its side and examine the regression of X on Y, you will discover that in every diagram it has exactly the same slope as that of Y on X. (The direction of the slope is reversed because the sequence of Y scores is reversed when the diagram is viewed from the side.)

A MATRIX OF CORRELATIONS

Investigators seldom publish studies that report but a single coefficient of correlation. Many published studies cite tens and some even hundreds of such coefficients. There needs to be some systematic way of recording all that information so that any part of it can be seen and compared to other parts. The system commonly used is called a *correlation matrix*.

A matrix is a table consisting of rows and columns of numbers. A matrix of correlation coefficients differs from most others in that it is perfectly symmetrical across its diagonal; that is, the numbers in the upper right half form a mirror image of those in the lower left. A single example should suffice to clarify the concept.

The ultimate objective of many psychological studies is to reveal the *structure* of human behavior and experience. One approach to that objective is correlational. Investigators take measurements and correlate the results, thus discovering what is related to what. Frequently the number of variables in the resulting table (the matrix) is enormous, but the *form* of the table can be illustrated just as well with only a few. Table 6-5 is a complete matrix of the coefficients that might result from the correlation of five variables, each one with the other four.

The coefficient of correlation between any two variables appears at the intersection of a row and a column, the row representing one of the two variables, the column representing the other. A brief scrutiny of the matrix will reveal how that comes about and how the table provides a systematic way of recording the results of a correlational study. You may also notice the special feature of matrices that I mentioned earlier: the perfect symmetry across the diagonal.

The fact is, however, that you don't need the symmetry, and you don't need the diagonal. The latter is derived not from data but from theory: each *1.00* is an idealized reliability coefficient—that is, it shows how each test would correlate with itself if it were perfectly reliable. That information is useful only as a reference point—or rather, a reference line—because you know in advance that a perfectly reliable test would correlate 1.00 with itself. (See the section after this one.) The symmetry is interesting, but it results from the incorporation of redundant information: If you know what is on one side of the diagonal, you know what is on the other.

TABLE 6-5 Intercorrelations of five ability measures: the complete matrix

		Shop work quality	Mechanical assembly	Mechanical information	IQ	School grades
		1	2	3	4	5
Shop work quality	1	1.00	.60	.40	.20	.40
Mechanical assembly	2	.60	1.00	.40	.00	.10
Mechanical information	3	.40	.40	1.00	.70	.50
IQ	4	.20	.00	.70	1.00	.60
School grades	5	.40	.10	.50	.60	1.00

Adapted from P. E. Vernon. *The Structure of Human Abilities*. London: Methuen & Company, Ltd, 1950, p. 102.

Table 6-6 omits everything (1) that you know in advance and (2) that you can infer with certainty by examining the entries on the other side of the diagonal. That is the form in which reports are commonly made (although the diagonal series of perfect correlations is frequently not deleted). The essential information is all there, namely, the intercorrelations among five measures taken of a large number of workers in automobile repair shops. Scores on *shop work quality* are derived from supervisors' ratings; *mechanical assembly* is a test that confronts examinees with a dismembered machine and requires them to reassemble it. The data for the *mechanical information* entries are also test scores, as are those for *IQ*. *School grades* are overall averages (means) from high school.

TABLE 6-6 Intercorrelations of five ability measures: an abbreviated report

		1	2	3	4	5
Shop work quality	1					
Mechanical assembly	2	.60				
Mechanical information	3	.40	.40			
IQ	4	.20	.00	.70		
School grades	5	.40	.10	.50	.60	

Source: Adapted from P. E. Vernon. *The Structure of Human Abilities*. London: Methuen & Company, Ltd, 1950, p. 102.

Now look again at Table 6-6, this time for content. Think about it. Think about the relative magnitudes of the relations represented by these coefficients. (All the relations are positive, which is nearly always the case with abilities; so you needn't be concerned with *direction* when making these comparisons.) Are your preconceptions generally confirmed? Are there any surprises? Do any possible explanations occur to you concerning counterintuitive results? These are the kinds of thoughts that form in the mind of a good researcher in the presence of a correlation matrix.

EXPECTANCY TABLES AND PREDICTIVE VALIDITY

There are many occasions in contemporary medical and social science research on which a relation between *two* variables is the focus of interest. On such occasions, various correlation techniques may be applied. When a correlation coefficient is used to predict scores on one variable from scores on another, the accuracy of the prediction is known as *predictive validity*.

The Expectancy Table as a Scatterplot

The correlation coefficient can be a very useful statistic when the relation between two variables is of interest. However, if you need to communicate relational information to a person untrained in statistics (for instance, a high school student or the student's parent) some other way must be found—some display that is concrete enough to be comprehended without benefit of previous study. A scatterplot is one such display. You will recall that in seeking the Pearson r we were really attempting to define the slope of the regression line with both X and Y scores in standard units. The formula for r converts the scores into standard units and makes possible a very concise communication (r) to *anyone who is familiar with the process*. But by using a lot more space, we can communicate essentially the same information without converting to standard scores. A scatterplot displays that information, and most scatterplots use raw scores.

Figure 6-12 shows the regression of grade point averages on test scores. The line across the drawing is the regression line—that is, the *line of best fit*. (In this case, the slope is .37, as it would have been if the two scales had been laid out in standard units.) It would be easy to predict grade point averages from test scores by simply referring to that line. The only trouble with such predictions is that they leave out some important information: They ignore *variability*. If X were an aptitude test score and Y were grade point average at Central University, the regression line would tell a student with a score of 57, for example, that in college his grade point average would be 2.2. The entire scatterplot, on the other hand, would say that although a 2.2 grade point average might be the central tendency of students who scored 57 on the test, it would not be the only possibility.

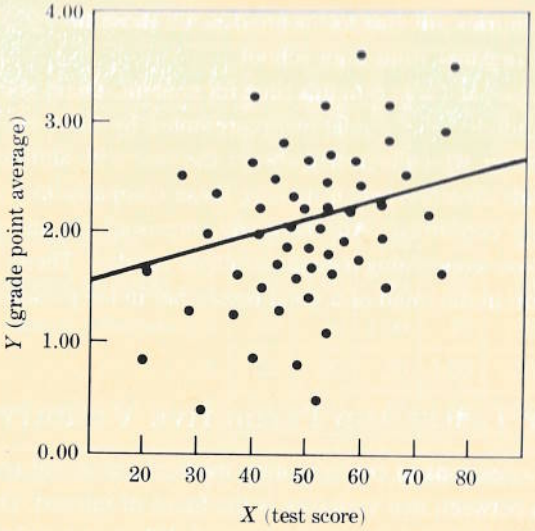


FIGURE 6-12 Scatterplot with superimposed regression line when $r = .37$.

An *expectancy table* functions in the same way as a scatterplot but is easier to use. The space in the table is divided on two dimensions into rows and columns, thus forming an array of cells. If the entry in each cell is a frequency, then the scatterplot is almost exactly duplicated, as in Table 6-7. Usually, however, the frequencies in each column are transformed into percentages of the column sum, as in Table 6-8. In either case, you enter the table with the raw score from a student's test and emerge with more than one possible grade point average.

Using an Expectancy Table

Now you can tell your student not only his most probable grade point average in college, but also his chances of doing better (or worse) than the most probable performance. Let's say that Joe Cawledge's score on test X is 57; Table 6-8 says that

TABLE 6-7 Expectancy of frequencies for Figure 6-13 ($r = .37$)

Y	10–19	20–29	30–39	40–49	50–59	60–69	70–79	80–89
3.00–3.99				1	2	1	1	
2.00–2.99		1	2	7	7	3	2	
1.00–1.99		2	3	7	7	2	1	
0.00–0.99		1	1	2	1			

X

TABLE 6-8 Expectancy of percentages for Figure 6-12 ($r = .37$)

Y	10–19	20–29	30–39	40–49	50–59	60–69	70–79	80–89
3.00–3.99				6	12	17	25	
2.00–2.99		25	33	41	41	50	50	
1.00–1.99		50	50	41	41	33	25	
0.00–0.99		25	17	12	6	0	0	

X

of all students who previously scored in the 50s and subsequently attended Central University, 6 percent earned a grade point average below 1.00, 41 percent a GPA in the interval 1.00 through 1.99, another 41 percent between 2.00 and 2.99, and 12 percent a 3.00 or better. That kind of information would be comprehensible to any interested person. Indeed, it may be helpful even to the experienced counselor, for although the counselor knows immediately that any r below a magnitude of 1.00 implies some scatter, it is often difficult to visualize the *amount* of scatter. The expectancy table gets right down to cases (as does, in a slightly different way, the scatterplot).

A refinement often made in the interpretation of such a table is the derivation of *cumulative* percentages. Table 6-9, for example, implies not that Joe Cawledge has a 41 percent chance of achieving a grade point average between 2.00 and 2.99, but that the probability of his doing *2.00 or better* is .41 + .12 = .53. That sum is often more useful than either probability by itself—so often, in fact, that many expectancy tables are made up entirely of cumulative percentages. Table 6-9 is based upon the same data as Table 6-8; the only difference is that all the percentages are cumulative. Although that structure makes Table 6-9 less flexible than Table 6-8, the many persons who are interested primarily in cumulative percentages will find it easier to use.

TABLE 6-9 Expectancy of cumulative percentages for Figure 6-12 ($r = .37$)

Y	10–19	20–29	30–39	40–49	50–59	60–69	70–79	80–89
3.00				6	12	17	25	
2.00		25	33	47	53	67	75	
1.00		75	83	88	94	100	100	
0.00		100	100	100	100	100	100	

X

RELIABILITY AND VALIDITY

At the beginning of this chapter, I suggested that an index of correlation would allow us to check on the accuracy of predictions made by a scholastic aptitude test. Now, by substituting the test score for X and college grade point average for Y in each of the diagrams in Figure 6-11, you can see just how accurate the predictions would be if the correlation were .00, .75, or 1.00.

A correlation coefficient that is used to predict a criterion score (Y) from a test score (X), as in the case of the scholastic aptitude test, is called a *validity coefficient*.* (The *criterion* is the entity that our test is supposed to be measuring—in this case, scholastic aptitude, or the capacity to profit from schooling. (Grade point average is generally accepted, though sometimes reluctantly, as an index of scholastic achievement—that is, the extent to which an individual has profited from schooling.) If we were to administer our scholastic aptitude test to all entering freshmen as indicated earlier, and if we were to test them again a week later, we could compute an r that would tell us the relationship between the first and second testings. Instead of an r_{xy}, we would have an r_{xx}, and it would be called a *reliability* coefficient because it would tell us the extent to which we can depend on the test to give us the same results from one administration to the next. (You can't measure anything reliably with a rubber yardstick!)

The rubber yardstick metaphor is a good one to help you remember the concept of reliability, but it bears no direct relation to *validity*. Here is an analogy that illustrates both reliability and validity: Imagine a woman aiming a rifle in the general direction of a target that is several hundred feet away from her. I say "in the general direction" because there is a huge sheet of white gauze that hides not only the target but a considerable area around it. Our marksman takes aim at a spot that she hopes is the bull's eye. She fires five shots—and scatters them all over that cloth screen. Figure 6-13 shows the pattern that the shots make on the screen and on the hidden target.

Another marksman takes aim at the same spot as the first one did. He groups his shots as shown in Figure 6-14, thereby missing the target completely.

A third marksman knows where the target is, takes aim in that direction, and pumps off five shots that are as close together as those of the second marksman. The result is shown in Figure 6-15.

Now review the concepts of reliability and validity, and see if you can tell which of them is illustrated by which marksman. Reliability? That would have to be the closely grouped patterns, wouldn't it? Shot after shot lands in nearly the same place; you can count on it, just as you can count on a reliable test to yield nearly the same scores when the same people are retested. Proximity to the target, on the other

* The examples given are of predictive validity and test–retest reliability. There are other kinds of validity and of reliability, but these examples illustrate how correlation coefficients can represent the relationships involved.

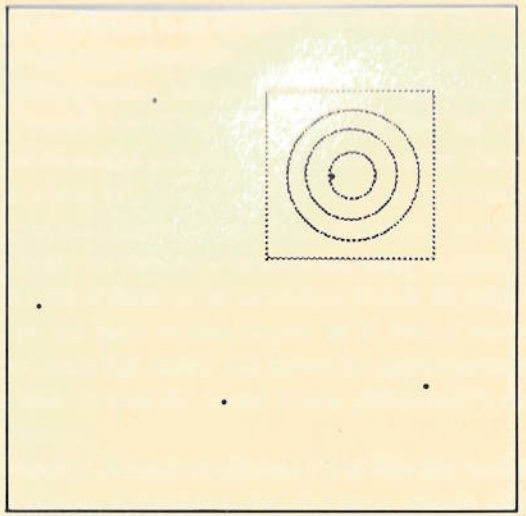


FIGURE 6-13 Widely spaced shots fired at invisible target concealed by sheet of white gauze.

hand, is the analog of validity. Close grouping of shots doesn't help a bit unless it is on target. Reliable testing is useless unless we know what we are testing.

So high reliability does not ensure high validity. But it is also important to note that *low* reliability absolutely precludes it. Unless your shots are all very close

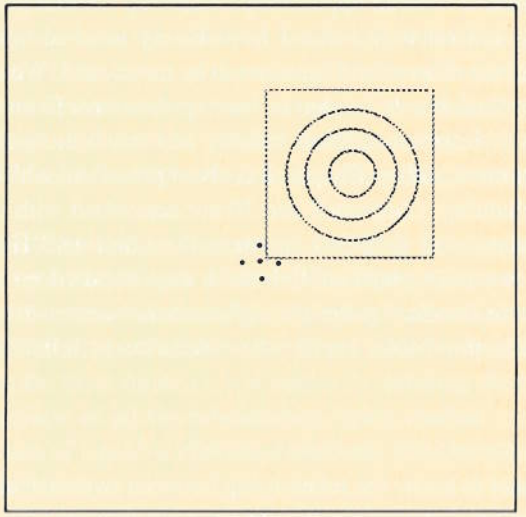


FIGURE 6-14 Closely grouped shots fired off target with target concealed by sheet of white gauze.

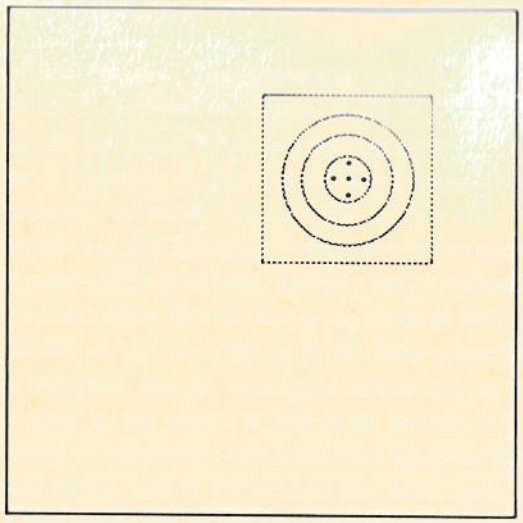


FIGURE 6-15 Closely grouped shots fired on target with target concealed by sheet of white gauze.

together, it is impossible to score well. If a test has low reliability, it can have *some* validity (see Figure 6-13) but not much. To put it another way, a test can be highly reliable without being at all valid, but it cannot be very valid without being highly reliable.

If when you have finished this book you continue your study of statistics and measurement, you will have an opportunity to develop a better understanding of the kinds of validity and reliability that I have barely touched upon; you will also discover that there are others that I have not even mentioned. We will not go further into those matters here, but do be alert to later applications. Even in this book, there are opportunities to learn more about validity and much more about reliability.

With respect to reliability, Chapter 8 is about precision, which is really another way of saying reliability. Chapters 9 and 10 are concerned with the significance of obtained differences, and that, too, answers the question "How reliable is the information that we have obtained?" Indeed, any standard error (as in *standard error of the mean* or *standard error of a difference between means*, to name the two specifically cited in this book) implies the estimation of reliability.

SUMMARY

It is often important to know the relationship between two variables. *Coefficients of correlation* are indices of relationship. Many such indices have been devised for various applications, but we have discussed just two: the rank-difference and the product-moment coefficients.

Any correlation coefficient is an index of the extent to which measurements of the same individuals are to be found on corresponding segments (e.g., low, middle, and high) of two different scales. When such a relationship holds precisely and with no exceptions, the correlation is said to be perfect, and the coefficient is 1.00. Less-than-perfect relationships produce coefficients smaller than 1.00. A *negative* correlation is just as strong as a *positive* correlation if its coefficient is equally large; the coefficient indexes magnitude and direction separately in a single number that varies from .00 to 1.00 and that does or does not carry a minus sign.

The Spearman *rank-difference* (ρ) procedure places all subjects in order from smallest to largest on the basis of their scores on X; then it ascertains how nearly their Y scores approximate that order. The better the approximation, the smaller are the differences in rank (hence the name "rank difference") and the larger the coefficient of correlation.

The Pearson *product-moment coefficient* (r) is like the rank-difference coefficient except that it retains information that the latter throws away—namely, the distances that separate subjects of adjacent ranks on either of the variables being correlated.

If the position of each subject (a pair of observations) is plotted on both X and Y coordinates, a graph called a *scatterplot* emerges. In that graph it is possible to perceive directly the degree and direction of the relationship between two variables. (We might still want to compute an r, however; the most *direct* index of an entity is often not the most *precise*.) The same general pattern, expressed somewhat more crudely in tabular form, has been named an *expectancy table* after its intended use in the prediction of outcomes.

The formula for r features in its numerator the deviation scores of both X and Y variables, but the presence of their standard deviations in the denominator is important, too, for they have the effect of converting deviation scores to standard scores. The Pearson r is the mean of the cross products of deviation scores on both variables when those deviation scores are expressed in standard units. The conversion to standard units also permits another definition of r: It is the slope of the regression of the Y variable on the X when both have been converted to standard scores.

Correlational studies seldom stop at two variables. Some relate tens and even hundreds of them, each one to each of the others. Whenever more than one pair of variables is involved, a convenient way of displaying the coefficients is the *correlation matrix*, in which all the variables are listed both horizontally, labeling the columns of cells in the table (matrix), and vertically, labeling the rows. It is a good way to get an overview of all the relationships being studied.

Reliability is one of a pair of important attributes that characterize all measurements; the other is *validity*.

In the kinds of situations cited here, the *coefficient of reliability* tells us whether a test does *consistently* whatever it does; the *coefficient of validity* tells us whether it is doing what we have *believed* it to be doing.

A coefficient of correlation quantifies the empirical relation between two variables. But does an *empirical* relation imply a *causal* relation? Chapter 11 will consider that question.

Sample Applications

Education

You are an elementary school consultant helping teachers plan activities to enhance students' emotional and social development. You have observed informally that students who have a negative view of themselves (low self-concept) seem not to become involved in helping class groups achieve common goals (low social responsibility). You wonder whether those observations can be confirmed objectively. To investigate this relationship you administer a self-concept and a social-responsibility scale to 200 fourth-, fifth-, and sixth-grade students. After the data have been collected, how do you organize them to answer your question about the relationship?

Political Science

Researchers have frequently asked whether there is any relationship between the amount of domestic conflict within a given country (X) and the amount of foreign conflict which that country initiates (Y). Assume that you have constructed a conflict scale and collected data for 50 countries on the values of both X and Y. What statistic will answer your question?

Psychology

You are a child psychologist interested in the possible positive relationship between the amount of sugar in the breakfast diet of young children and hyperactivity (i.e., inattentiveness, excessive muscular activity, etc.). You ask 100 mothers of elementary school children to keep records on what their children eat and drink each morning for breakfast. A nutritionist then analyzes the parental reports and notes the average weekly amount of sugar ingested. Data on hyperactivity are gathered by behavior analysts who visit classrooms on a daily basis and rate each child on a 10-point continuum. What statistic will indicate the strength and direction of the relationship between sugar intake and activity level?

Social Work

You are a social worker studying the records of women who receive AFDC payments. You are interested in the possible association between a woman's level of self-esteem and the number of months she receives benefits. What single index can summarize information pertinent to this problem?

Sociology

Your research class is locked in a disagreement over the possible relationship between the conservatism/liberalism of academic disciplines and authoritarianism. (A study of authoritarianism has already been done; see sociology application in Chapter 5.) The question is whether liberal arts disciplines such as English, history, philosophy, and psychology either attract or produce professors who are significantly more liberal and less authoritarian than professors of disciplines such as math, biology, chemistry, business, and physics, who, some students claim, are more conservative and authoritarian.

Having chosen this issue as your research topic, you begin your project by acquiring the data from the authoritarianism study. Then, to identify the conservative/liberal orientation of academic disciplines, you pick a panel of 10 judges to place the disciplines (not the professors) on an interval scale from 0 (most conservative) to 100 (most liberal).

How do you quantify the relation between these two variables?

7

Description to Inference: A Transition

In Chapter 1, I pointed out that statistics has two main functions: (1) *description* of populations, or of samples taken from those populations and (2) *inference* from the properties of samples (*statistics*) to those of populations (*parameters*). If you are doing research for your employers and for their eyes only, description may be all you will need. But if you intend to generalize the outcome beyond your own situation, you will need something more, for in that event the subjects of your study are but a sample of a larger population. It is then that you will need inferential, sometimes called "sampling," statistics.

Chapter 1 identified description and inference as the two main functions of statistics. Since then you have seen references to both samples and populations in the chapter on *central tendency* and *variability*. In this chapter we shall address those topics once again, this time at a higher level of discourse.

Describing Observed Distributions via $\overline{X}$ and μ, and Estimating μ from $\overline{X}$

We begin with a discussion of description and of inference (estimation) with respect to a measure of central tendency—specifically, the mean.

Describing Observed Distributions via X and μ

This section will be short, mainly because Chapter 3 showed you that the formula for the population mean (μ) is identical to that of a sample mean ($\overline{X}$). To put it another way,

$$\overline{X} = \frac{\Sigma X}{n}$$

where ΣX is the sum of the (raw) scores in a sample and n is the number of such scores. Similarly,

$$\mu = \frac{\Sigma X}{N}$$

where ΣX is the sum of scores in the *population* and N is the number of such scores. If after calculating a sample mean you were suddenly informed that your "sample" is really the *population* of interest, no recalculation would be necessary; you would simply report your mean as μ instead of $\overline{X}$.

Estimating μ from $\overline{X}$

There is a much more important implication of learning that your "sample" is really the whole population: You need not be concerned about the hazards of making inferences about a population you have not been privileged to observe.

If, on the other hand, your "sample" really *is* but a sample of the target population, you *do* have to be concerned: You need (1) to select the best available estimates of important parameters and (2) to ascertain the *reliability* of such estimates.

With respect to the parameter known as the mean, your first concern—the selection of the best estimate—is easily abated: *The best estimate of the population mean is the mean of a sample* taken from that population.

With respect to ascertaining the *reliability* of your estimate, the answer is not so simple. It involves a new concept: *the standard error of the mean*, which is a measure of the variability of a distribution of sample means—a notion that I shall develop for you in Chapter 8. The variability of sample means is significantly dependent on that of individual scores in the population, however; so we must deal with that first.

Describing Observed Distributions via S and σ, and Estimating σ From S

In the preceding section I asserted that the best estimate of the population mean is the mean of the sample. The state of affairs with respect to the standard deviation, however, is not nearly so simple.

Describing Observed Distributions: S and σ

You may recall that as a kind of recapitulation of our discussion of the standard deviation in Chapter 4, I listed four formulas that are related in one way or another to the concept of the standard deviation. I have reproduced them here for your convenience:

$$\text{The mean of the raw scores} = \frac{\Sigma X}{n}$$

$$\text{The average deviation} = \frac{\Sigma |x|}{n}$$

$$\text{The variance} = \frac{\Sigma x^2}{n}$$

$$\text{The standard deviation} = \sqrt{\frac{\Sigma x^2}{n}}$$

The numbers that enter these formulas are taken directly from your sample, and the numbers that emerge from them refer directly to that sample. The very word "sample" implies, however, that your interest is not primarily in the sample. Rather, it is in the *population* from which the sample was drawn: It is in *parameters* rather than *statistics* per se. Indeed, when the population is small enough to be measured in its entirety, you won't even *draw* a sample.

In Chapter 1 we compared description to inference, and I suggested several examples of each. Some of those examples appear in Table 7-1, along with comments concerning the description or inference involved. Study that table.

One implication of this discussion is that the standard deviation of a sample (S) is of only theoretical importance, because whenever you draw a sample from a population, your primary interest is always in the population, not the sample. Except for its theoretical function, a sample is useful only insofar as it helps you to learn something about the population. Unfortunately, a sample's standard deviation (S) is a biased estimate of the standard deviation of its parent population (σ).* So now that you thoroughly comprehend S, I am telling you that it should never be used!

"But," you should be asking about now, "what about situations in which I can observe the *entire population*, as in examples 1, 3, and 5 in Table 7-1? What formula will give me the population standard deviation in those situations?"

The answer is filed somewhere in your memory. If you cannot recall it immediately, turn back to Formulas (4-2) and (4-3), page 36. A quick review will remind you that *the two formulas are identical* (excepting for the symbols that

* If we were to calculate an S for each of an infinite number of samples, the mean of those S's would *not* be equal to σ.

TABLE 7-1 Some examples of description and inference[a]

	Example	Comment
1	A college professor tests a class for knowledge of subject matter.	Description of *entire* population; namely, that particular class. Inference not necessary.
2	A political polling agency interviews 1000 potential voters in a national election.	Inference (of preferences) to *entire* electorate from those of sample.
3	A national election is held.	Description of population, namely, the *entire* electorate.
4	A pharmacologist supervises clinical trials of a new drug on 500 patients.	Inference (of a specified effect of the drug) to *all* potential patients. Description of population impossible.
5	The current batting average of every player is made available to the broadcasters of a baseball game.	Description of *all* at-bats. Description possible here even though there may be a very large number of at-bats because every one has been observed and the result recorded.

[a]Samples are always described, but the properties of a population may be inferred from those of a sample.

indicate references to sample and population, respectively). That makes sense, because in both cases you are *describing* a set of actual observations, as distinguished from *inferring* the nature of a set of observations that have not been made.

Estimating σ FROM S: A NEW STATISTIC, s

When you do want to estimate some characteristic (parameter) of a population that, except for a small sample, you have *not* observed, the situation is different. As I mentioned earlier, the standard deviation of a sample (S) is a biased estimate of the standard deviation of the population (σ). Specifically, S tends to underestimate σ.

The sample standard deviation is the closest approximation of the population standard deviation, however, so we'll start with that and then modify it to correct its tendency to underestimate. Diminishing the denominator (n) of the fraction within the formula

$$S = \sqrt{\frac{\Sigma x^2_{\text{sample}}}{n}}$$

would produce a value larger than S, and that is what we do: The formula for the estimated standard deviation of a population is

$$s = \sqrt{\frac{\Sigma x^2_{\text{sample}}}{n-1}} \qquad (7\text{-}1)$$

where s is the standard deviation of the population as estimated from sample data, $\Sigma x^2_{\text{sample}}$ is the sum of the squared deviations in the sample, and $n - 1$ is *degrees of freedom*.

Degrees of freedom in any calculation is the number of values that are free to vary, given whatever mathematical restrictions are inherent in that calculation.[1] The estimation of a population standard deviation involves one such restriction and the consequent loss of 1 degree of freedom. However, that is not the only possibility; *df* can be n (where there are *no* restrictions), $n - 1$ (as in the present case), $n - 2$, $n - 3$, or even smaller.

When a formula is concerned with straightforward *description*, degrees of freedom is n. When the objective of a calculation is *inference*, some restrictions apply, so that degrees of freedom is smaller than n (e.g., $n - 1, n - 2, n - 3$, etc.). You may expect to see more instances of *df* smaller than n, because from here on there will be much ado about inference.

SUMMARY

Earlier chapters demonstrated sophisticated ways to *describe* a distribution of individual measurements; this one prepares you for the later chapters where such data are used to *infer* properties of a distribution not in evidence. The preparation is done in terms of two kinds of measure—a measure of central tendency (the mean) and a measure of variability (the standard deviation).

The message with respect to the mean is simple: The mean of any sample from the population of interest is an unbiased estimate of the population mean.

The case of the standard deviation is more complicated, but it can be abridged as follows:

1. The standard deviation of a sample

$$S = \sqrt{\frac{\Sigma x^2_{\text{sample}}}{n}}$$

can be calculated directly from the data; but you will have no occasion to do so, because when you take a sample of a population, your interest is in the population.

2. The standard deviation of the population

$$\sigma = \sqrt{\frac{\Sigma x^2_{\text{pop.}}}{N}}$$

is calculated directly from data whenever it is feasible to do so. Its formula is essentially the same as that of S.

TABLE 7-2 Six important concepts with their symbols and formulas[a]

$\bar{X}$ = mean of a sample	$\frac{\Sigma X_{\text{sample}}}{n}$
μ = mean of the population	$\frac{\Sigma X_{\text{pop}}}{N}$
$\bar{X}$ = mean of the population as estimated from a sample	$\frac{\Sigma X_{\text{sample}}}{n}$
S = standard deviation of a sample	$\sqrt{\frac{\Sigma x^2_{\text{sample}}}{n}}$
σ = standard deviation of the population	$\sqrt{\frac{\Sigma x^2_{\text{pop.}}}{N}}$
s = standard deviation of the population as estimated from a sample	$\sqrt{\frac{\Sigma x^2_{\text{sample}}}{n-1}}$

[a]If the word "estimated" does not appear opposite a symbol, the value it represents is calculated directly from data.

3. S underestimates σ.

4. The standard deviation of the population as estimated from that of a sample compensates for that underestimation. In other words, s

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

is the same as S except for the loss of 1 degree of freedom (the "-1" in the denominator).

Table 7-2 lists the key concepts of this chapter and organizes them for easy access, along with their symbols and defining formulas. No calculation boxes are included in this chapter because you have already calculated all of those measures, with the exception of the population standard deviation as estimated from sample data (s), and that one is exactly the same as the (S) that you *have* calculated, except for the substitution of $n - 1$ for n (See Box 4-1, pages 38 and 39.)