

United States General Accounting Office

GAO

Program Evaluation and Methodology
Division

June 1992

**Quantitative Data
Analysis: An
Introduction**

GAO/PEMD-10.1.11



Preface

GAO assists congressional decisionmakers in their deliberative process by furnishing analytical information on issues and options under consideration. Many diverse methodologies are needed to develop sound and timely answers to the questions that are posed by the Congress. To provide GAO evaluators with basic information about the more commonly used methodologies, GAO's policy guidance includes documents such as methodology transfer papers and technical guidelines.

This methodology transfer paper on quantitative data analysis deals with information expressed as numbers, as opposed to words, and is about statistical analysis in particular because most numerical analyses by GAO are of that form. The intended reader is the GAO generalist, not statisticians and other experts on evaluation design and methodology. The paper aims to bridge the communications gap between generalist and specialist, helping the generalist evaluator be a wiser consumer of technical advice and helping report reviewers be more sensitive to the potential for methodological errors. The intent is thus to provide a brief tour of the statistical terrain by introducing concepts and issues important to GAO's work, illustrating the use of a variety of statistical methods, discussing factors that influence the choice of methods, and offering some advice on how to avoid pitfalls in the analysis of quantitative data. Concepts are presented in a nontechnical way by avoiding computational procedures, except for a few illustrations, and by avoiding a rigorous discussion of assumptions that underlie statistical methods.

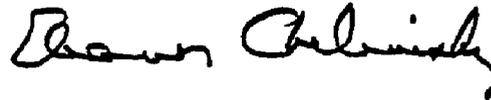
Quantitative Data Analysis is one of a series of papers issued by the Program Evaluation and Methodology Division (PEMD). The purpose of the series is to provide GAO evaluators with guides to various aspects of audit and evaluation methodology, to illustrate applications, and to indicate where more detailed information is available.

Preface

We look forward to receiving comments from the readers of this paper. They should be addressed to Eleanor Chelmsky at 202-275-1854.



Werner Grosshans
Assistant Comptroller General
Office of Policy



Eleanor Chelmsky
Assistant Comptroller General
for Program Evaluation and Methodology

Contents

Preface		1
Chapter 1		8
Introduction	Guiding Principles	8
	Quantitative Questions Addressed in the Chapters of This Paper	11
	Attributes, Variables, and Cases	13
	Level of Measurement	16
	Unit of Analysis	18
	Distribution of a Variable	19
	Populations, Probability Samples, and Batches	24
	Completeness of the Data	26
	Statistics	27
Chapter 2		29
Determining the Central Tendency of a Distribution	Measures of the Central Tendency of a Distribution	31
	Analyzing and Reporting Central Tendency	33
Chapter 3		37
Determining the Spread of a Distribution	Measures of the Spread of a Distribution	38
	Analyzing and Reporting Spread	46
Chapter 4		48
Determining Association Among Variables	What Is an Association Among Variables?	48
	Measures of Association Between Two Variables	52
	The Comparison of Groups	64
	Analyzing and Reporting the Association Between Variables	67

Contents

Chapter 5		71
Estimating Population Parameters	Histograms and Probability Distributions	73
	Sampling Distributions	77
	Population Parameters	80
	Point Estimates of Population Parameters	81
	Interval Estimates of Population Parameters	84
<hr/>		
Chapter 6		88
Determining Causation	What Do We Mean by Causal Association?	88
	Evidence for Causation	91
	Limitations of Causal Analysis	100
<hr/>		
Chapter 7		102
Avoiding Pitfalls	In the Early Planning Stages	102
	When Plans Are Being Made for Data Collection	105
	As the Data Analysis Begins	106
	As the Results Are Produced and Interpreted	109
<hr/>		
Bibliography		112
<hr/>		
Glossary		117
<hr/>		
Contributors		127
<hr/>		
Papers in This Series		128
<hr/>		
Tables	Table 1.1: Generic Types of Quantitative Questions	11
	Table 1.2: Data Sheet for a Study of College Student Loan Balances	15
	Table 1.3: Tabular Display of a Distribution	24
	Table 2.1: Distribution of Staff Turnover Rates in Long-Term Care Facilities	30
	Table 2.2: Three Common Measures of Central Tendency	31

Contents

Table 2.3: Illustrative Measures of Central Tendency	34
Table 3.1: Measures of Spread	39
Table 4.1: Data Sheet With Two Variables	49
Table 4.2: Cross-Tabulation of Two Ordinal Variables	50
Table 4.3: Percentaged Cross-Tabulation of Two Ordinal Variables	51
Table 4.4: Cross-Tabulation of Two Nominal Variables	54
Table 4.5: Two Ordinal Variables Showing No Association	67
Table 5.1: Data Sheet for 100 Samples of College Students	78
Table 5.2: Point and Interval Estimates for a Set of Samples	85

Figures

Figure 1.1: Histogram of Loan Balances	20
Figure 1.2: Two Distributions	21
Figure 1.3: Histogram for a Nominal Variable	23
Figure 3.1: Histogram of Hospital Mortality Rates	37
Figure 3.2: Spread of a Distribution	41
Figure 3.3: Spread in a Normal Distribution	45
Figure 4.1: Scatter Plots for Spending Level and Test Scores	56
Figure 4.2: Regression of Test Score on Spending Level	60
Figure 4.3: Regression of Spending Level on Test Score	62
Figure 4.4: Linear and Nonlinear Associations	69
Figure 5.1: Frequency Distribution of Loan Balances	73
Figure 5.2: Probability Distribution of Loan Balances	75
Figure 5.3: Sampling Distribution for Mean Student Loan Balances	79
Figure 6.1: Causal Network	93

Contents

Abbreviations

AIDS	Acquired immune deficiency syndrome
GAO	U.S. General Accounting Office
PEMD	Program Evaluation and Methodology Division
PRE	Proportionate reduction in error
WIC	Special Supplemental Food Program for Women, Infants, and Children

Introduction

Guiding Principles

Data analysis is more than number crunching. It is an activity that permeates all stages of a study. Concern with analysis should (1) begin during the design of a study, (2) continue as detailed plans are made to collect data in different forms, (3) become the focus of attention after data are collected, and (4) be completed only during the report writing and reviewing stages.¹

The basic thesis of this paper is that successful data analysis, whether quantitative or qualitative, requires (1) understanding a variety of data analysis methods; (2) planning data analysis early in a project and making revisions in the plan as the work develops; (3) understanding which methods will best answer the study questions posed, given the data that have been collected; and (4) once the analysis is finished, recognizing how weaknesses in the data or the analysis affect the conclusions that can properly be drawn. The study questions govern the overall analysis, of course. But the form and quality of the data determine what analyses can be performed and what can be inferred from them. This implies that the evaluator should think about data analysis at four junctures:

- when the study is in the design phase,
- when detailed plans are being made for data collection,
- after the data are collected, and
- as the report is being written and reviewed.

¹Relative to GAO job phases, the first two checkpoints occur during the job design phase, the third occurs during data collection and analysis, and the fourth during product preparation. For detail on job phases see the General Policy Manual, chapter 6, and the Project Manual, chapters 6.2, 6.3, and 6.4.

Chapter 1
Introduction

Designing the Study

As policy-relevant questions are being formulated, evaluators should decide what data will be needed to answer the questions and how they will analyze the data. In other words, they need to develop a data analysis plan. Determining the type and scope of data analysis is an integral part of an overall design for the study. (See the transfer paper entitled Designing Evaluations, listed in "Papers in This Series.") Moreover, confronting data collection and analysis issues at this stage may lead to a reformulation of the questions to ones that can be answered within the time and resources available.

Data Collection

When evaluators have advanced to the point of planning the details of data collection, analysis must be considered again. Observations can be made and, if they are qualitative (that is, text data), converted to numbers in a variety of ways that affect the kinds of analyses that can be performed and the interpretations that can be made of the results. Therefore, decisions about how to collect data should be influenced by the analysis options in mind.

Data Analysis

After the data are collected, evaluators need to see whether their expectations regarding data characteristics and quality have been met. Choice among possible analyses should be based partly on the nature of the data—for example, whether many observed values are small and a few are large and whether the data are complete. If the data do not fit the assumptions of the methods they had planned to

use, the evaluators have to regroup and decide what to do with the data they have.² A different form of data analysis may be advisable, but if some observations are untrustworthy or missing altogether, additional data collection may be necessary.

As the evaluators proceed with data analysis, intermediate results should be monitored to avoid pitfalls that may invalidate the conclusions. This is not just verifying the completeness of the data and the accuracy of the calculations but maintaining the logic of the analysis. Yet it is more, because the avoidance of pitfalls is both a science and an art. Balancing the analytic alternatives calls for the exercise of considerable judgment. For example, when observations take on an unusual range of values, what methods should be used to describe the results? What if there are a few very large or small values in a set of data? Should we drop data at the extreme high and low ends of the scale? On what grounds?

**Writing and
Reviewing**

Finally, as the evaluators interpret the results and write the report, they have to close the loop by making judgments about how well they have answered the questions, determining whether different or supplementary analyses are warranted, and deciding the form of any recommendations that may be suitable. They have to ask themselves questions about their data collection and analysis: How much of the variation in the data has been accounted for? Is the method of analysis sensitive enough to detect the

²An example would be a study in which the data analysis method evaluators planned to use required the assumption that observations be from a probability sample, as discussed in chapter 5. If the evaluators did not obtain observations for a portion of the intended sample, the assumption might not be warranted and their application of the method could be questioned.

effects of a program? Are the data “strong” enough to warrant a far-reaching recommendation? These questions and many others may occur to the evaluators and reviewers and good answers will come only if the analyst is “close” to the data but always with an eye on the overall study questions.

Quantitative Questions Addressed in the Chapters of This Paper

Most GAO statistical analyses address one or more of the four generic questions presented in table 1.1. Each generic question is illustrated with several specific questions and examples of the kinds of statistics that might be computed to answer the questions. The specific questions are loosely based on past GAO studies of state bottle bills (U.S. General Accounting Office, 1977 and 1980).

Table 1.1: Generic Types of Quantitative Questions

Generic question	Specific question	Useful statistics
What is a typical value of the variable?	At the state level, how many pounds of soft drink bottles (per unit of population) were typically returned annually?	Measures of central tendency (ch.2)
How much spread is there among the cases?	How similar are the individual states' return rates?	Measures of spread (ch. 3)
To what extent are two or more variables associated?	What factors are most associated with high return rates: existence of state bottle bills? state economic conditions? state levels of environmental awareness?	Measures of association (ch. 4)
To what extent are there causal relationships among two or more variables?	What factors cause high return rates: existence of state bottle bills? state economic conditions? state level of environmental awareness?	Measures of association (ch.4): Note that association is but one of three conditions necessary to establish causation (ch.6)

Bottle bills have been adopted by about nine states and are intended to reduce solid waste disposal problems by recycling. Other benefits can also be sought, such as the reduction of environmental litter and savings of energy and natural resources. One of GAO's studies was a prospective analysis, intended to inform discussion of a proposed national bottle bill. The quantitative analyses were not the only relevant factor. For example, the evaluators had to consider the interaction of the merchant-based bottle bill strategy with emerging state incentives for curbside pickups or with other recycling initiatives sponsored by local communities. The quantitative results were, however, relevant to the overall conclusions regarding the likely benefits of the proposed national bottle bill.

The first three generic questions in table 1.1 are standard fare for statistical analysis. GAO reports using quantitative analysis usually include answers in the form of descriptive statistics such as the mean, a measure of central tendency, and the standard deviation, a measure of spread. In chapters 2, 3, and 4 of this paper, we focus on descriptive statistics for answering the questions.

To answer many questions, it is desirable to use probability samples to draw conclusions about populations. In chapter 5, we address the first three questions from the perspective of inferential statistics. The treatment there is necessarily brief, focused on point and interval estimation methods.

The fourth generic question, about causality, is more difficult to answer than the others. Providing a good answer to a causal question depends heavily upon the study design and somewhat advanced statistical methods; we treat the topic only lightly in chapter 6. Chapter 7 discusses some broad strategies for avoiding pitfalls in the analysis of quantitative data.

Before describing these concepts, it is important to establish a common understanding about some ideas that are basic to data analysis, especially those applicable to the quantitative analysis we describe in this paper. Each of GAO's assignments requires considerable analysis of data. Over the years, many workable tools and methods have been developed and perfected. Trained evaluators use these tools as appropriate in addressing an assignment's objectives. This paper tries to reinforce the uses of these tools and put consistent labels on them.³ It also gives helpful hints and illustrates the use of each tool. In the next section, we discuss the basic terminology that is used in later chapters.

**Attributes,
Variables, and
Cases**

Observations about persons, things, and events are central to answering questions about government programs and policies. Groups of observations are called data, which may be qualitative or quantitative. Statistical analysis is the manipulation, summarization, and interpretation of quantitative data.

We observe characteristics of the entities we are studying. For example, we observe that a person is female and we refer to that characteristic as an attribute of the person. A logical collection of attributes is called a variable; in this instance, the

³Inconsistencies in the use of statistical terms can cause problems. We have tried to deal with the difficulty in three ways: (1) by using the language of current writers in the field, (2) by noting instances where there are common alternatives to key terms, and (3) by including a glossary of the terms used in this paper.

variable would be gender and would be composed of the attributes female and male.⁴ Age might be another variable composed of the integer values from 0 to 115.

It is convenient to refer to the variables we are especially interested in as response variables. For example, in a study of the effects of a government retraining program for displaced workers, employment rate might be the response variable. In trying to determine the need for an acquired immune deficiency syndrome (AIDS) education program in different segments of the U.S. population, evaluators might use the incidence of AIDS as the response variable. We usually also collect information on other variables with which we hope to better understand the response variables. We occasionally refer to these other variables as supplementary variables.

The data that we want to analyze can be displayed in a rectangular or matrix form, often called a data sheet (see table 1.2). To simplify matters, the individual persons, things, or events that we get information about are referred to generically as cases. (The intensive study of one or a few cases, typically combining quantitative and qualitative data, is referred to as case study research. See the GAO transfer paper entitled Case Study Evaluations.) Traditionally, the rows in a data sheet correspond to the cases and the columns correspond to the variables

⁴Instead of referring to the attributes of a variable, some prefer to say that the variable takes on a number of "values." For example, the variable gender can have two values, male and female. Also, some statisticians use the expression "attribute sampling" in reference to probability sampling procedures for estimating proportions. Although attribute sampling is related to attribute as used in data analysis, the terminology is not perfectly parallel. See the discussion of attribute sampling in the transfer paper entitled Using Statistical Sampling, listed in "Papers in This Series."

of interest. The numbers or words in the cells then correspond to the attributes of the cases.

Table 1.2: Data Sheet for a Study of College Student Loan Balances

Case	Age	Class	Type of institution	Loan balance
1	23	Sophomore	Private	\$3,254
2	19	Freshman	Public	1,501
3	21	Junior	Public	2,361
4	30	Graduate	Private	8,100
5	21	Freshman	Private	1,970
6	22	Sophomore	Public	3,866
7	21	Sophomore	Public	2,890
8	20	Freshman	Public	6,300
9	22	Junior	Private	2,639
10	21	Sophomore	Public	1,718
11	19	Freshman	Private	2,690
12	20	Sophomore	Public	3,812
13	20	Sophomore	Public	2,210
14	23	Senior	Private	3,780
15	24	Senior	Private	5,082

Table 1.2 shows 15 cases, college students, from a hypothetical study of student loan balances at higher education institutions. The first column shows an identification number for each case, and the rest of the columns indicate four variables: age of student, class, type of institution, and loan balance. Two of the variables, class and type of institution, are presently in text form. As will be seen shortly, they can be converted to numbers for purposes of quantitative analysis. Loan balance is the response variable and the others are supplementary.

The choice of a data analysis method is affected by several considerations, especially the level of measurement for the variables to be studied; the unit of analysis; the shape of the distribution of a variable, including the presence of outliers (extreme values); the study design used to produce the data from populations, probability samples, or batches; and the completeness of the data. Each factor is considered briefly.

**Level of
Measurement**

Quantitative variables take several forms, frequently called levels of measurement, which affect the type of data analysis that is appropriate. Although the terminology used by different analysts is not uniform, one common way to classify a quantitative variable is according to whether it is nominal, ordinal, interval, or ratio.

The attributes of a nominal variable have no inherent order. For example, gender is a nominal variable in that being male is neither better nor worse than being female. Persons, things, and events characterized by a nominal variable are not ranked or ordered by the variable. For purposes of data analysis, we can assign numbers to the attributes of a nominal variable but must remember that the numbers are just labels and must not be interpreted as conveying the order of the attributes. In the study of student loans, the type of institution is a nominal variable with two attributes—private and public—to which we might assign the numbers 0 and 1 or, if we wish, 12 and 17. For most purposes, 0 and 1 would be more useful.⁵

With an ordinal variable, the attributes are ordered. For example, observations about attitudes are often

⁵A variable for which the attributes are assigned arbitrary numerical values is usually called a "dummy variable." Dummy variables occur frequently in evaluation studies.

arrayed into five classifications, such as greatly dislike, moderately dislike, indifferent to, moderately like, greatly like. Participants in a government program might be asked to categorize their views of the program offerings in this way. Although the ordinal level of measurement yields a ranking of attributes, no assumptions are made about the "distance" between the classifications. In this example, we do not assume that the difference between persons who greatly like a program offering and ones who moderately like it is the same as the difference between persons who moderately like the offering and ones who are indifferent to it. For data analysis, numbers are assigned to the attributes (for example, greatly dislike = -2, moderately dislike = -1, indifferent to = 0, moderately like = +1, and greatly like = +2), but the numbers are understood to indicate rank order and the "distance" between the numbers has no meaning. Any other assignment of numbers that preserves the rank order of the attributes would serve as well. In the student loan study, class is an ordinal variable.

The attributes of an interval variable are assumed to be equally spaced. For example, temperature on the Fahrenheit scale is an interval variable. The difference between a temperature of 45 degrees and 46 degrees is taken to be the same as the difference between 90 degrees and 91 degrees. However, it is not assumed that a 90-degree object has twice the temperature of a 45-degree object (meaning that the ratio of temperatures is not necessarily 2 to 1). The condition that makes the ratio of two observations uninterpretable is the absence of a true zero for the variable. In general, with variables measured at the interval level, it makes no sense to try to interpret the ratio of two observations.

The attributes of a ratio variable are assumed to have equal intervals and a true zero point. For example, age is a ratio variable because the negative age of a

person or object is not meaningful and, thus, the birth of the person or the creation of the object is a true zero point. With ratio variables, it makes sense to form ratios of observations and it is thus meaningful, for example, to say that a person of 90 years is twice as old as one of 45. In the study of student loans, age and loan balance are both ratio variables (the attributes are equally spaced and the variables have true zero points). For analysis purposes, it is seldom necessary to distinguish between interval and ratio variables so we usually lump them together and call them interval-ratio variables.

Unit of Analysis

Units of analysis are the persons, things, or events under study—the entities that we want to say something about. Frequently, the appropriate units of analysis are easy to select. They follow from the purpose of the study. For example, if we want to know how people feel about the offerings of a government program, individual people would be the logical unit of analysis. In the statistical analysis, the set of data to be manipulated would be variables defined at the level of the individual.

However, in some studies, variables can potentially be analyzed at two or more levels of aggregation. Suppose, for example, that evaluators wished to evaluate a compensatory reading program and had acquired reading test scores on a large number of children, some who participated in the program and some who did not. One way to analyze the data would be to treat each child as a case.

But another possibility would be to aggregate the scores of the individual children to the classroom level. For example, they could compute the average scores for the children in each classroom that participated in their study. They could then treat each classroom as a unit, and an average reading test score would be an attribute of a classroom. Other variables,

such as teacher's years of experience, number of students, and hours of instruction could be defined at the classroom level. The data analysis would proceed by using classrooms as the unit of analysis. For some issues, treating each child as a unit might seem more appropriate, while in others each classroom might seem a better choice. And we can imagine rationales for aggregating to the school, school district, and even state level.

Summarizing, the unit of analysis is the level at which analysis is conducted. We have, in this example, five possible units of analysis: child, classroom, school, school district, and state. We can move up the ladder of aggregation by computing average reading scores across lower-level units. In effect, the definition of the variable changes as we change the unit of analysis. The lowest-level variable might be called child-reading-score, the next could be classroom-average-reading-score, and so on.

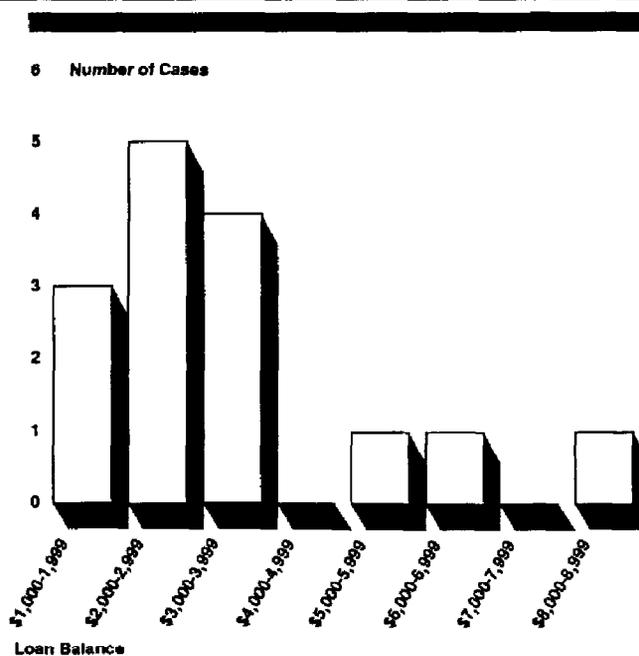
In general, the results from an analysis will vary, depending upon the unit of analysis. Thus, for studies in which aggregation is a possibility, evaluators must answer the question: What is the appropriate unit of analysis? Several situation-specific factors may need consideration, and there may not be a clear-cut answer. Sometimes analyses are carried out with several units of analysis. (GAO evaluators should seek advice from technical assistance groups.)

Distribution of a Variable

The cases we observe vary in the characteristics of interest to us. For example, students vary by class and by loan balance. Such variation across cases, which is called the distribution of a variable, is the focus of attention in a statistical analysis. Among the several ways to picture or describe a distribution, the histogram is probably the simplest. To illustrate, suppose we want to display the distribution of the loan balance variable for the 15 cases in table 1.2. A

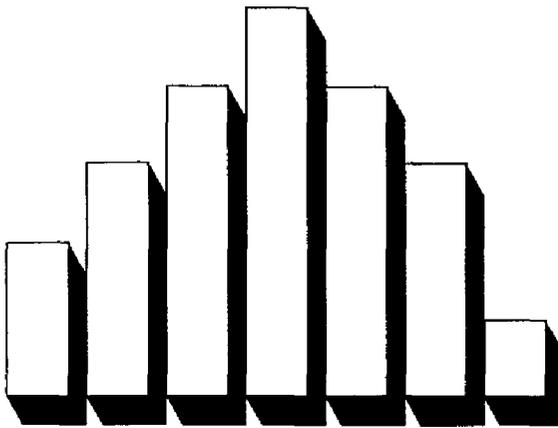
histogram for the data is shown in figure 1.1. The length of the lefthand bar corresponds to the number of observations between \$1,000 and \$1,999. There are three: \$1,500, \$1,970, and \$1,718. The lengths of the other bars are determined in a similar fashion, and the overall histogram gives a picture of the distribution. In this example, the distribution is rather "piled up" on one end and spread out at the other; two intervals have no observations.

Figure 1.1: Histogram of Loan Balances

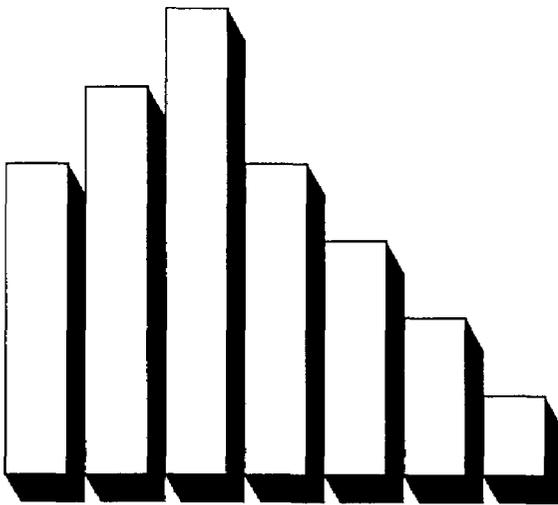


Histograms show the shape of a distribution, a factor that helps determine the type of data analysis that will be appropriate. For example, some techniques are suitable only when the distribution is approximately symmetrical (as in figure 1.2a), while others can be

Figure 1.2: Two
Distributions



a. Approximately Symmetrical Distribution



b. Asymmetrical Distribution

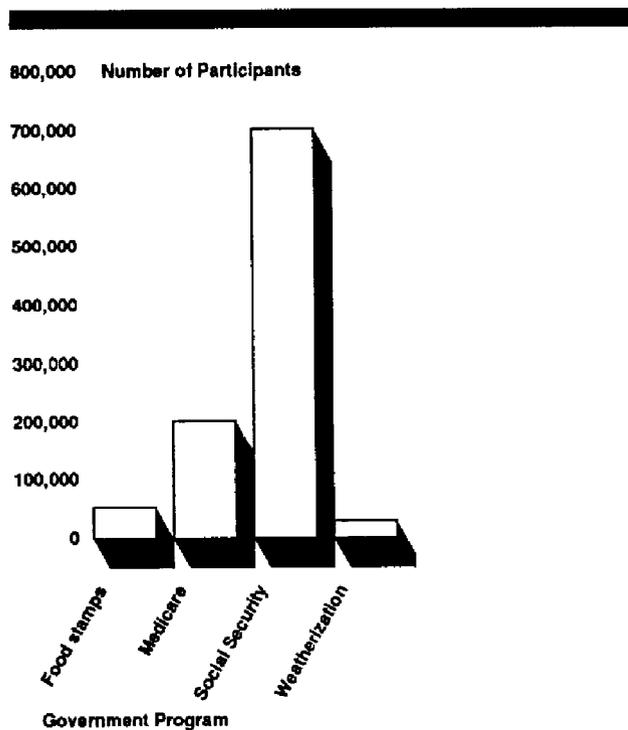
used when the observations are asymmetrical (figure 1.2b). Once data are collected for a study, we need to inspect the distributions of the variables to see what initial steps are appropriate for the data analysis. Sometimes it is advisable to transform a variable (that is, systematically change the values of the observations) that is distributed asymmetrically to one that is symmetric. For example, taking the square root of each observation is a transformation that will sometimes work. Velleman and Hoaglin (1981, ch. 2) provide a good introduction to transformation strategies (they refer to them as “re-expression”) and Hoaglin, Mosteller, and Tukey (1983, ch. 4) give a more complete treatment. With proper care, transformations do not alter the conclusions that can be drawn from data.

Another aspect of a distribution is the possible presence of outliers, a few observations that have extremely large or small values so that they lie on the outer reaches of the distribution. For the student loan observations, case number 4, which has a value of \$8,100, is far from the center of the distribution. Outliers can be important because they may lead to new understanding of the variable in question. However, outliers attributable to measurement error may produce misleading results with some statistical analyses, so an early decision must be made about how to handle outliers—a decision not easy to make. The usual way is to employ analytical methods that are relatively insensitive to outliers—for example, by using the median instead of the mean. Sometimes outliers are dropped but only if there is good reason to believe that the observations are in error.

Considerations about the shape of a distribution and about outliers apply to ordinal, interval, and ratio variables. Because the attributes of a nominal variable have no inherent order, these spatial relationships have no meaning. However, we can still display the results from observations on a nominal variable as a

histogram, as long as we remember that the order of the attributes is arbitrary. Figure 1.3 shows hypothetical data on the number of participants in four government programs. There is no inherent order for displaying the programs.

Figure 1.3: Histogram for a Nominal Variable



Another way of showing the distribution of a variable is to use a simple table. Suppose evaluators have data on 341 homeowners' attitudes toward energy conservation with three categories of response: indifferent, somewhat positive, and positive. Table 1.3

shows the data in summary form. This kind of display is not often used when only one variable is involved, but with two it is common (see chapter 4).

Table 1.3: Tabular Display of a Distribution

Attitude toward energy conservation	Number of homeowners
Indifferent	120
Somewhat positive	115
Positive	106
Total	341

Populations, Probability Samples, and Batches

Statistical analysis is applied to a group of cases. The process by which the group was chosen (that is, the study design) affects the type of data analysis that is appropriate and the interpretations that may be drawn from the analysis. Three types of group are of interest: populations, probability samples, and batches.

A population is the full set of cases that the evaluators have a question about. For example, suppose they want to know the age of Medicaid participants and the amount of benefits these participants received last year. The population would be all persons who received such benefits, and the evaluators might obtain data tapes containing the attributes for all such persons. They could perform statistical analyses to describe the distributions of certain variables such as age and amount of benefits received. The results of such an analysis are called descriptive statistics.

A second way to draw conclusions about the Medicaid participants is to use a probability sample from the population of beneficiaries. A probability sample is a group of cases selected so that each member of the population has a known, nonzero probability of being

selected. (For detailed information on probability sampling, see the transfer paper entitled Using Statistical Sampling.) Studies based on probability samples are usually less expensive than those that use data from the entire population and, under some conditions, are less error-prone.⁶ The study of probability samples can use descriptive statistics but the study of the population, upon which the probability sample is based, uses inferential statistics (discussed in chapter 5).

A group of cases can also be treated as a batch, a group produced by a process about which we make no probabilistic assumptions. For example, the evaluators might use their judgment, not probability, to select a number of interesting Medicaid cases for study. Being neither a population nor a probability sample, the set of cases is treated as a batch. As such, the techniques of descriptive statistics can be applied but not those of inferential statistics. Thus, conclusions about the population of which the batch is a part cannot be based on statistical rules of inference.

When do we regard a group of cases as a batch? Evaluators who have purposely chosen a nonprobability sample, or who have doubts about whether cases in hand fit the definition of a probability sample—for example, because they are using someone else's data and the selection procedures were not well described—should treat the cases as a batch. Actually, any group of cases can be

⁶Error in using probability samples to answer questions about populations stems from the net effects of both measurement error and sampling error. Conclusions based upon data from the entire population are subject only to measurement error. The total error associated with data from a probability sample may be less than the total error (measurement only) of data from a population.

regarded as a batch. The term is applied whenever we do not wish to assume the grouping is a population or a probability sample.

Completeness of the Data

When we design a study, we plan to obtain data for a specific number of cases. Despite our best plans, we usually cannot obtain data on all variables for all cases. For example, in a sample survey, some persons may decline to respond at all and others may not answer certain questionnaire items. Or responses to some interview questions may be inadvertently "lost" during data editing and processing. In another study, we may not be allowed to observe certain events. Almost inevitably, the data will be incomplete in several respects, and data analysis must contend with that eventuality.

Incompleteness in the data can affect analysis in a variety of ways. The classic example is when we draw a probability sample with the aim of using inferential statistics to answer questions about a population. To illustrate, suppose evaluators send a questionnaire to a sample of Medicaid beneficiaries but only 45 percent provide data. Without increasing the response rate or satisfying themselves that nonrespondents would have answered in ways similar to respondents (or that the differences would have been inconsequential), the evaluators would not be entitled to draw inferential conclusions about the population of Medicaid beneficiaries. If they knew the views of the nonrespondents, their overall description of the population might be quite different. They would be limited, therefore, to descriptive statistics about the 45 percent who responded, and that information might not be useful for answering a policy-relevant question.

The problem of incomplete data entails several considerations and a variety of analytic approaches. (See, for example, Groves, 1989; Madow, Olkin, and

Rubin, 1983; and Little and Rubin, 1987.) One important strategy is to minimize the problems by using good data collection techniques. (See the transfer papers entitled Using Structured Interviewing Techniques and Developing and Using Questionnaires.)

Statistics

In GAO work, we may be interested in analyzing data from a population, a probability sample, or a batch. Regardless of how the group of cases is selected, we make observations on the cases and can produce a data sheet like that of table 1.2. A main purpose of statistical analysis is to draw conclusions about the real world by computing useful statistics.⁷ A statistic is a number computed from a set of data. For example, the midpoint loan balance for the 15 students, \$2,890, is a statistic—the median loan balance for the batch in statistical terminology.

Many statistics are possible but only a relative few are useful in the sense of helping us understand the data and answer policy-relevant questions. Another possibly useful statistic from the batch of 15 is the range—the difference between the maximum loan balance and the minimum. The range, in this example, is $8,100 - 1,500 = 6,600$. In this instance, the “computation” of the statistic is merely a sorting through the attributes for the loan balance variable to find the largest and smallest values and then computing the difference between them. Many statistics can be imagined but most would not be useful in describing the batch. For example, the square root of the difference between the maximum

⁷Another purpose, though one that has received less attention in the statistical literature, is to devise useful ways to graphically depict the data. See, for example, Du Toit, Steyn, and Stumpf, 1986; and Tufte, 1983.

Chapter 1
Introduction

loan balance and the mean loan balance is a statistic but not a useful one.

The methods of statistical analysis provide us with ways to compute and interpret useful statistics. Those that are useful for describing a population or a batch are called descriptive statistics. They are used to describe a set of cases upon which observations were made. Methods that are useful for drawing inferences about a population from a probability sample are called inferential statistics. They are used to describe a population using merely information from observations on a probability sample of cases from the population. Thus, the same statistic can be descriptive or inferential or both, depending on its use.

Determining the Central Tendency of a Distribution

Descriptive analyses are the workhorses of GAO, carrying much of the message in many of our reports. There are three main forms of descriptive analysis: determining the central tendency in the distribution of a variable (discussed in this chapter), determining the spread of a distribution (chapter 3), and determining the association among variables (chapter 4).

The determination of central tendency answers the first of GAO's four basic questions, What is a typical value of the variable? All readers are familiar with the basic ideas. Sample questions might be

- How satisfied are Social Security beneficiaries with the agency's responsiveness?
- How much time is required to fill requests for fighter plane repair parts?
- What was the dollar value in agricultural subsidies received by wealthy farmers?
- What was the turnover rate among personnel in long-term care facilities?

The common theme of these questions is the need to express what is typical of a group of cases. For example, in the last question, the response variable is the turnover rate. Suppose evaluators have collected information on the turnover rates for 800 long-term care facilities. Assuming there is variation among the facilities, they would have a distribution for the turnover rate variable. There are two approaches for describing the central tendency of a distribution: (1) presenting the data on turnover rates in tables or figures and (2) finding a single number, a descriptive statistic, that best summarizes the distribution of turnover rates.

The first approach, shown in table 2.1, allows us to "see" the distribution. The trouble is that it may be hard to grasp what the typical value is. However, evaluators should always take a graphic or tabular approach as a first step to help in deciding how to

Chapter 2
Determining the Central Tendency of
a Distribution

proceed on the second approach, choosing a single statistic to represent the batch. How a display of the distribution can help will be seen shortly.

Table 2.1: Distribution of Staff Turnover Rates in Long-Term Care Facilities

Turnover rates (percent new staff per year)	Frequency count (number of long-term care facilities)
0-0.9	155
1.0-1.9	100
2.0-2.9	125
3.0-3.9	150
4.0-4.9	100
5.0-5.9	75
6.0-6.9	50
7.0-7.9	25
8.0-8.9	15
9.0-9.9	5

The second approach, describing the typical value of a variable with a single number, offers several possibilities. But before considering them, a little discussion of terminology is necessary. A descriptive statistic is a number, computed from observations of a batch, that in some way describes the group of cases. The definition of a particular descriptive statistic is specific, sometimes given as a recipe for calculation. Measures of central tendency form a class of descriptive statistics each member of which characterizes, in some sense, the typical value of a variable—the central location of a distribution.¹ The definition of central tendency is necessarily somewhat vague because it embraces a variety of computational

¹Measures of central tendency also go by other, equivalent names such as “center indicators” and “location indicators.”

procedures that frequently produce different numerical values. Nonetheless, the purpose of each measure would be to compress information about a whole distribution of cases into a single number.

Measures of the Central Tendency of a Distribution

Three familiar and commonly used measures of central tendency are summarized in table 2.2. The mean, or arithmetic average, is calculated by summing the observations and dividing the sum by the number of observations. It is ordinarily used as a measure of central tendency only with interval-ratio level data. However, the mean may not be a good choice if several cases are outliers or if the distribution is notably asymmetric. The reason is that the mean is strongly influenced by the presence of a few extreme values, which may give a distorted view of central tendency. Despite such limitations, the mean has definite advantages in inferential statistics (see chapter 5).

Table 2.2: Three Common Measures of Central Tendency

Measurement level	Use of measure ^a		
	Mode	Median	Mean
Nominal	Yes	No	No
Ordinal	Yes	Yes	No ^b
Interval-ratio	Yes	Yes	Yes ^c

^a“Yes” means the indicator is suitable for the measurement level shown.

^bMay be OK in some circumstances. See chapter 7.

^cMay be misleading when the distribution is asymmetric or has a few outliers.

The median—calculated by determining the midpoint of rank-ordered cases—can be used with ordinal,

Chapter 2
Determining the Central Tendency of
a Distribution

interval, or ratio measurements and no assumptions need be made about the shape of the distribution.² The median has another attractive feature: it is a resistant measure. That means it is not much affected by changes in a few cases. Intuitively, this suggests that significant errors of observation in several cases will not greatly distort the results. Because it is a resistant measure, outliers have less influence on the median than on the mean. For example, notice that the observations 1,4,4,5,7,7,8,8,9 have the same median (7) as the observations 1,4,4,5,7,7,8,8,542. The means (5.89 and 65.44, respectively), however, are quite different because of the outlier, 542, in the second set of observations.

The mode is determined by finding the attribute that is most often observed.³ That is, we simply count the number of times each attribute occurs in the data, and the mode is the most frequently occurring attribute. It can be used as a measure of central tendency with data at any level of measurement. However, the mode is most commonly employed with nominal variables

²With an odd number of cases, the midpoint is the median. With an even number of cases, the median is the mean of the middle pair of cases.

³This definition is suitable when the mode is used with nominal and ordinal variables—the most common situation. A slightly different definition is required for interval-ratio variables.

and is generally less used for other levels. A distribution can have more than one mode (when two or more attributes tie for the highest frequency). When it does, that fact alone gives important information about the shape of the distribution.

Measures of central tendency are used frequently in GAO reports. In a study of tuition guarantee programs (U.S. General Accounting Office, 1990c), for example, the mean was often used to characterize the programs in the sample, but when outliers were evident, the median was reported. In another GAO study (U.S. General Accounting Office, 1988), the distinctions between properties of the mode, median, and mean figured prominently in an analysis of procedures used by the Employment and Training Administration to determine prevailing wage rates of farmworkers.

Analyzing and Reporting Central Tendency

To illustrate some considerations involved in determining the central tendency of a distribution, we can recall the earlier study question about the views of Social Security beneficiaries regarding program services. Assume that a questionnaire has been sent to a batch of 800 Social Security recipients asking how satisfied they are with program services.⁴ Further, imagine four hypothetical distributions of the responses. By assigning a numerical value of 1 to the item response "very satisfied" and 5 to "very dissatisfied," and so on, we can create an ordinal

⁴To keep the discussion general, we make no assumptions about how the group of recipients was chosen. However, in GAO, a probability sample would usually form the basis for data collection by a mailout questionnaire.

Chapter 2
Determining the Central Tendency of
a Distribution

variable. The three measures of central tendency can then be computed to produce the results in table 2.3.⁵ Although the data are ordinal, we have included the mean for comparison purposes.

Table 2.3: Illustrative Measures of Central Tendency

Attribute	Code	Distribution			
		A	B	C	D
Very satisfied	1	250	250	100	159
Satisfied	2	200	150	150	159
Neither satisfied nor dissatisfied	3	125	0	300	164
Dissatisfied	4	125	150	150	159
Very dissatisfied	5	100	250	100	159
Total responses		800	800	800	800
Mean		2.5	3	3	3
Median		2	3	3	3
Mode		1 and 5		3	3

In distribution A, the data are distributed asymmetrically. More persons report being very satisfied than any other condition, and mode 1 reflects this. However, 225 beneficiaries expressed some degree of dissatisfaction (codes 4 and 5), and these observations pull the mean to a value of 2.5 (that is,

⁵Although computer programs automatically compute a variety of indicators and although we display three of them here, we are not suggesting that this is a good practice. In general, the choice of an indicator should be based upon the measurement level of a variable and the shape of the distribution.

Chapter 2
Determining the Central Tendency of
a Distribution

toward the dissatisfied end of the scale). The median is 2, between the mode and the mean. Although the mean might be acceptable for some ordinal variables, in this example it can be misleading and shows the danger of using a single measure with an asymmetrical distribution. The mode seems unsatisfactory also because, although it draws attention to the fact that more respondents reported satisfaction with the services than any other category, it obscures the point that 225 reported that they were dissatisfied or very dissatisfied. The median seems the better choice for this distribution if we can display only one number, but showing the whole distribution is probably wise.

In distribution B, the mean and the median both equal 3 (a central tendency of “neither satisfied nor dissatisfied”). Some would say this is nonsense in terms of the actual distribution, since no one actually chose the middle category. Modes 1 and 5 seem the better choices to represent the clearly bimodal distribution, although again a display of the full distribution is probably the best option.

In distribution C, the mean, median, and mode are identical; the distribution is symmetrical. Any one of the three would be appropriate. One easy check on the symmetry of a distribution, as this shows, is to compare the values of the mean, median, and mode. If they differ substantially, as with distribution A, the distribution is probably such that the median should be used.

As distribution D illustrates, however, this rule-of-thumb is not infallible. Although the mean, median, and mode agree, the distribution is almost flat. In this case, a single measure of central tendency could be misleading, since the values 1, 2, 3, 4, and 5 are all about equally likely to occur. Thus, the full distribution should be displayed.

Chapter 2
Determining the Central Tendency of
a Distribution

The lesson of this example? First, before representing the central tendency by any single number, evaluators need to look at the distribution and decide whether the indicator would be misleading. Second, there will be occasions when displaying the results graphically or in tabular form will be desirable instead of, or in addition to, reporting statistics.

The interpretation of a measure of central tendency comes from the context of the associated policy question. The number itself does not carry along a message saying whether policymakers should be complacent or concerned about the central tendency. For example, the observed mean agricultural subsidy for farmers can be interpreted only in the context of economic and social policy. Comparison of the mean to other numbers such as the wealth or income level of farmers or to the trend over time for mean subsidies might be helpful in this regard. And, of course, limits on mean values are sometimes written into law. An example is the fleet-average mileage standard for automobiles. Information that can be used to interpret the observed measures of central tendency is a necessary part of the overall answer to a policy question.

Determining the Spread of a Distribution

Spread refers to the extent of variation among cases—sometimes cases cluster closely together and sometimes they are widely spread out. When we determine appropriate policy action, the spread of a distribution may be as much a factor, or more, than the central tendency.

The point is illustrated by the issue of variation in hospital mortality rates. Consider two questions. How much do hospital mortality rates vary? If there is substantial variation, what accounts for it? We consider questions of the first type in this chapter and questions of the second type in chapter 6.

Figure 3.1: Histogram of Hospital Mortality Rates

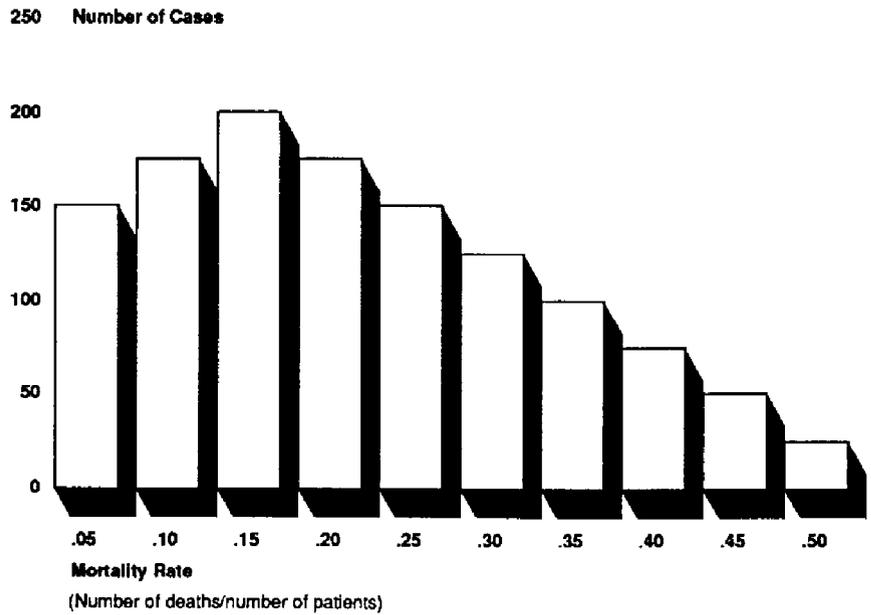


Figure 3.1 shows the distribution of hypothetical data on mortality rates in 1,225 hospitals. While the depiction is useful both in gaining an initial understanding of the spread in mortality rates and in communicating findings, it is also usually desirable to produce a number that characterizes the variation in the distribution.

Other questions in which spread is the issue are

- What is the variability in timber production among national forests?
- What is the variation among the states in food stamp participation rates?
- What is the spread in asset value among failed savings and loan institutions?

In each of these examples, we are addressing the generic question, How much spread (or variation in the response variable) is there among the cases? (See table 1.1.)

Even when spread is not the center of attention, it is an important concept in data analysis and should be reported when a set of data are described. Whenever evaluators give information about the central tendency of a distribution, they should also describe the spread.

Measures of the Spread of a Distribution

There is a variety of statistics for gauging the spread of a distribution. Some measures should be used only with interval-ratio measurement while others are appropriate for nominal or ordinal data. Table 3.1 summarizes the use of four particular measures.

**Chapter 3
Determining the Spread of a
Distribution**

Table 3.1: Measures of Spread

Measurement level	Use of measure			
	Index of dispersion	Range	Interquartile range	Standard deviation
Nominal	Yes	No	No	No
Ordinal	Sometimes	Sometimes	Yes	No
Interval-ratio	No	Yes	Yes	Yes

The index of dispersion is a measure of spread for nominal or ordinal variables. With such variables, each case falls into one of a number of categories. The index shows the extent to which cases are bunched up in one or a few categories rather than being well spread out among the available categories.

The calculation of the index is based upon the concept of unique pairs of cases. Suppose, for example, we want to know the spread for gender, a nominal variable. Assume a batch of 8 cases, 3 females and 5 males. Each of the 3 females could be paired with each of the 5 males to yield 15 unique pairs (3 x 5).

The index is a ratio in which the numerator is the number of unique pairs (15 in the example) that can be created given the observed number of cases ($n = 8$ in the example). The denominator of the ratio is the maximum number of unique pairs of cases that can be created with n cases. The maximum occurs when the cases are evenly divided among the available categories.

The maximum number of unique pairs (for $n = 8$) would occur if the batch included 4 females and 4 males (the 8 cases evenly divided among the two categories). Under this condition, 16 unique pairs (4 x 4) could be formed. The index of dispersion for the example would thus be $15/16 = .94$. Although this example illustrates the concept of the index, the

Chapter 3
Determining the Spread of a
Distribution

calculation of the index becomes more tedious as the number of cases and the number of categories increase. Loether and McTavish (1988) give a computational formula and a computer program for the index of dispersion.

As the cases become more spread out among the available categories, the index of dispersion increases in value. The index of dispersion can be as large as 1, when the categories have equal numbers of cases, and as small as 0, when all cases are in one category.

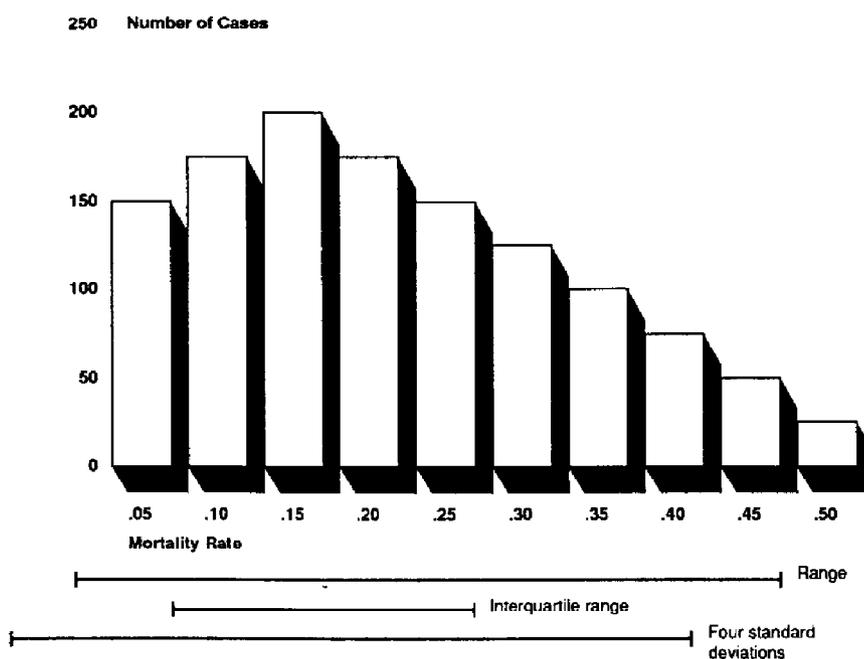
The range is a commonly used measure of spread when a variable is measured at least at the ordinal level. The range is the difference between the largest and smallest observations in the distribution. Because the range is based solely on the extreme values, it is a crude measure that is very sensitive to sample size and to outliers. The effect of an outlier is shown by the two distributions we considered in chapter 2: (1) 1,4,4,5,7,7,8,8,9, and (2) 1,4,4,5,7,7,8,8,542. The range for the first distribution is 8, and for the second it is 541. The huge difference is attributable to the presence of an outlier in the second distribution.

A range of 0 means there is no variation in the cases, but unlike the index of dispersion, the range has no upper limit. The range is not used with nominal

Chapter 3
Determining the Spread of a
Distribution

variables because the measure makes sense only when cases are ordered. To illustrate the measure, the distribution of hospital mortality rates, is reproduced in figure 3.2. Inspection of the data showed that the minimum rate was .025 and the maximum was .475, so the range is .45.

Figure 3.2: Spread of a Distribution



Another measure of spread, the interquartile range, is the difference between the two points in a distribution that bracket the middle 50 percent of the cases. These two points are called the 1st and 3rd quartiles and, in effect, they cut the upper and lower 25 percent of the cases from the range. The more closely the cases are bunched together, the smaller will be the value of the

Chapter 3
Determining the Spread of a
Distribution

interquartile range. Like the range, the interquartile range requires at least an ordinal level of measurement, but by discounting extreme cases, it is not subject to criticism for being inappropriately sensitive to outliers. In the hospital mortality example, the 1st quartile is .075 and the 3rd quartile is .275 so the interquartile range is the difference, .2.

A fourth measure of spread, one often used with interval-ratio data, is the standard deviation. It is the square root of the average of the squares of the deviations of each case from the mean. As with the preceding measures, the standard deviation is 0 when there is no variation among the cases. It has no upper limit, however. For the distribution of hospital mortality rate, the standard deviation is .12 but note, from figure 3.2, that the distribution is somewhat asymmetric, so this measure of spread is apt to be misleading. The four-standard-deviation band shown in figure 3.2 is .48 units wide and centered on the sample mean of .19.¹

One way of interpreting or explaining the spread of a distribution (for ordinal or higher variables) is to look at the proportion of cases "covered" by a measure of dispersion. To do this, we think of a spread measure as a band having a lower value and an upper value and then imagine that band superimposed on the distribution of cases. A certain proportion of the cases have observations larger than the lower value of the band and less than the upper value; those cases are thus covered by the spread measure. For the range, the lower value is the smallest observation among all cases and the upper value is the largest observation

¹ Expressing the spread as a band of four standard deviations is a common but not unique practice. Any multiple of standard deviations would be acceptable but two, four, and six are commonplace.

Chapter 3
Determining the Spread of a
Distribution

(see figure 3.2, based upon 1,225 cases). Then 100 percent of the cases are covered by the range.

Likewise, we know that when the interquartile range is used, 50 percent of the cases are always covered. The situation with the standard deviation is more complex but ultimately, in terms of inferential statistics, more useful.

When we use the standard deviation as the measure of spread, we can define the width of the band in an infinite number of ways but only two or three are commonly used. One possibility is to define the lower value of the band as the mean minus one standard deviation and the upper value as the mean plus one standard deviation. In other words, this band is two standard deviations wide (and centered on the mean). We could then simply count the cases in the batch that are covered by the band. However, it is important to realize that the number of cases can vary from study to study. For example, 53 percent of the cases might be covered in one study, to pick an arbitrary figure, and 66 percent in another. Just how many depends upon the shape of the distribution. So, unlike the situation with the range or the interquartile range, the measure by itself does not imply that a specified proportion of cases will be covered by a band that is two standard deviations wide. Thus if we know only the width of the band, we may have difficulty interpreting the meaning of the measure. Other bands could be defined as four standard deviations wide or

Chapter 3
Determining the Spread of a
Distribution

any other multiple of the basic measure, a standard deviation.²

We can obtain some idea of the effect of distribution shape on the interpretation of the standard deviation by considering three situations. We may believe that the distribution is (1) close to a theoretical curve called the normal distribution (the familiar bell-shaped curve), (2) has a single mode and is approximately symmetric (but not necessarily normal in shape), or (3) of unknown or “irregular” shape.³ For this example, we define the band to be four standard deviations wide (that is, two standard deviations on either side of the mean).

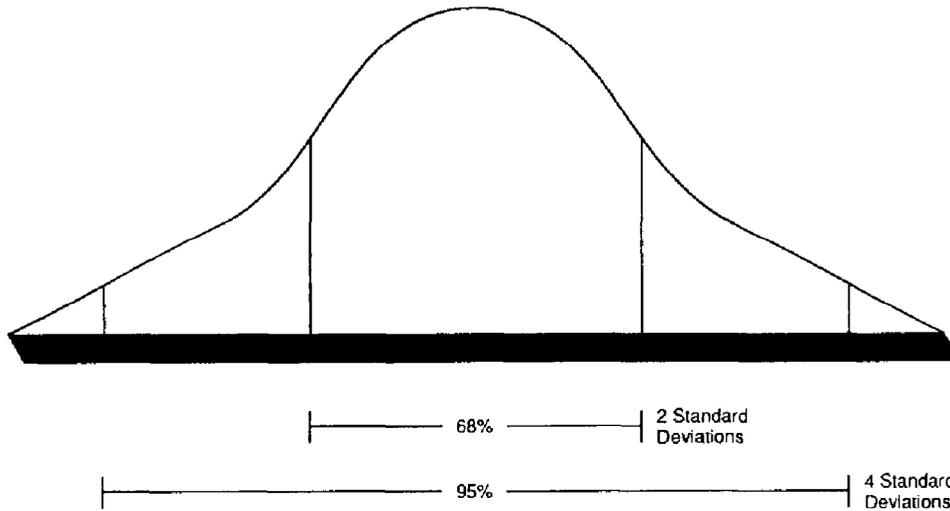
When the distribution of a batch is close to a normal distribution, statistical theory permits us to say that approximately 95 percent of the cases will be covered by the four-standard-deviation band. (See figure 3.3.) However, if we know only that the distribution is unimodal and symmetric, theory lets us say that, at minimum, 89 percent of the cases will be covered. If the distribution is multimodal or asymmetric or if we simply do not know its shape, we can make a weaker statement that applies to any distribution: that, at minimum, 75 percent of the cases will be covered by the four-standard-deviation band.

²The term “standard deviation” is sometimes misunderstood to be implying some substantive meaning to the amount of variation—that the variation is a large amount or a small amount. The measure by itself does not convey such information, and after we have computed a standard deviation, we still have to decide, on the basis of nonstatistical information, whether the variation is “large” or not.

³The name for the set of theoretical distributions called “normal” is unfortunate in that it seems to imply that distributions that have this form are “to be expected.” While many real-world distributions are indeed close to a normal (or Gaussian) distribution in shape, many others are not.

Chapter 3
Determining the Spread of a
Distribution

Figure 3.3: Spread in a Normal Distribution



From this example, it should be evident that care must be taken when using the standard deviation to describe the spread of a batch of cases. The common interpretation that a four-standard-deviation band covers about 95 percent of the cases is true only if the distribution is approximately normal.

One GAO example of describing the spread of a distribution comes from a report on Bureau of the Census methods for estimating the value of noncash benefits to poor families (U.S. General Accounting Office, 1987b). Variation in the amount of noncash benefits was described in terms of both the range and the standard deviation. In a study of homeless children and youths (U.S. General Accounting Office, 1989a), GAO evaluators asked shelter providers, advocates for the homeless, and government officials to estimate the proportion of the homeless persons, in a county, that seek shelter in a variety of settings (for

example, churches, formal shelters, and public places). The responses were summarized by reporting medians and by the first and third quartiles (from which the interquartile range can be computed).

Analyzing and Reporting Spread

To analyze the spread of a nominal variable, it is probably best just to develop a table or a histogram that shows the frequency of cases for each category of the variable. The calculation of a single measure, such as the index of dispersion, is not common but can be done.

For describing the spread of an ordinal variable, tables or histograms are useful, but the choice of a single measure is problematic. The index of dispersion is a possibility, but it does not take advantage of the known information about the order of the categories. Range, interquartile range, and standard deviation are all based on interval or ratio measurement. When a single measure is used, the best choice is often the interquartile range.

With an interval-ratio variable, graphic analysis of the spread is always advisable even if only a single measure is ultimately reported. The standard deviation is a commonly used measure but, as noted above, may be difficult to interpret if it cannot be shown that the cases have approximately a normal distribution.⁴ Consequently, the interquartile range may be a good alternative to the standard deviation when the distribution is questionable.

⁴A possible approach with a variable that does not have a normal distribution is to change the scale of the variable so that the shape does approximate the normal. See Velleman and Hoaglin (1981) for some examples; they refer to the process of changing the scale as "re-expression," but "transformation" of the variables is a more common term.

Chapter 3
Determining the Spread of a
Distribution

With respect to reporting data, a general principle applies: whenever central tendency is reported, spread should be reported too. There are two main reasons for this. The first is that a key study question may ask about the variability among cases. In such instances, the mean should be reported but the real issue pertains to the spread.

The second reason for describing the spread of a distribution, which applies even when the study question focuses on central tendency, is that knowledge of variation among individual cases tells us the extent to which an action based on the central tendency is likely to be on the mark. The point is that government action based upon the central tendency may be appropriate if the spread of cases is small, but if the spread is large, several different actions may be warranted to take account of the great variety among the cases. For example, policymakers might conclude that the mean mortality rate among hospitals is satisfactory and, given central tendency alone, might decide that no action is needed. If there is little spread among hospitals with respect to mortality rates, then taking no action may be appropriate. But if the spread is wide, then maybe hospitals with low rates should be studied to see what lessons can be learned from them and perhaps hospitals with extremely high rates should be looked at closely to see if improvements can be made.

Determining Association Among Variables

Many questions GAO addresses deal with associations among variables:

- Do 12th grade students in high-spending school districts learn more than students in low-spending districts?
- Are different procedures for monitoring thrift institutions associated with different rates for correctly predicting institution failure?
- Is there a relationship between geographical area and whether farm crop prices are affected by price supports?
- Are homeowners' attitudes about energy conservation related to their income level?
- Are homeowners' appliance-purchasing decisions associated with government information campaigns aimed at reducing energy consumption?

Recalling table 1.1, these examples illustrate the third generic question, *To what extent are two or more variables associated?* An answer to the first question would reveal, for example, whether high achievement levels tend to be found in higher-spending districts and low achievement levels in lower-spending districts (a positive association), or vice versa (a negative association).

What Is an Association Among Variables?

Just what do we mean by an association among variables?¹ The simplest case is that involving two variables, say homeowners' attitudes about energy conservation and income level. Imagine a data sheet as in table 4.1 representing the results of interviews with 341 homeowners. For these hypothetical data, we have adopted the following coding scheme: attitude toward energy conservation (indifferent 1,

¹The term "relationship" is equivalent to "association."

Chapter 4
Determining Association Among
Variables

somewhat positive 2, positive 3); family income level (low 1, medium 2, high 3).

Table 4.1: Data Sheet With Two Variables

Case	Attitude toward energy conservation	Family income level
1	3	2
2	3	1
3	1	3
...
341	2	2

To say that there is an association between the variables is to say that there is a particular pattern in the observations. Perhaps homeowners who respond that their attitude toward energy conservation is positive tend to report that they have low income and homeowners who respond that they are indifferent toward conservation tend to have high income. The pattern is that the cases vary together on the two variables of interest. Usually the relationship does not hold for every case but there is a tendency for it to occur.

The trouble with a data sheet like this is that it is usually not easy to perceive an association between the two variables. Evaluators need a way to summarize the data. One common way, with nominal or ordinal data, is to use a cross-tabular display as in table 4.2. The numbers in the cells of the table indicate the number of homeowners who responded to each possible combination of attitude and income level.

Chapter 4
Determining Association Among
Variables

Table 4.2:
Cross-Tabulation of Two
Ordinal Variables

Attitude toward energy conservation	Family income level			Total
	Low	Medium	High	
Indifferent	27	37	56	120
Somewhat positive	35	39	41	115
Positive	43	33	30	106
Total	105	109	127	341

Notice that the information in table 4.2 is an elaboration of the distribution of 341 homeowners shown in table 1.3. Reading down the total column in table 4.2 gives the distribution of the homeowners with respect to the attitude variable (the same as in table 1.3). In a two-variable table, this distribution is called a marginal distribution; it presents information on only one variable. The last row in table 4.2 (not including the grand total, 341) also gives a marginal distribution—that for the income variable.

There is much more information in table 4.2. If we look down the numbers in the low-income column only, we are looking at the distribution of attitude toward energy conservation for only low-income households. Or, if we look across the indifferent row, we are looking at the distribution of income levels for indifferent households. The distribution of one variable for a given value of the other variable is called a conditional distribution. Four other conditional distributions (for households with medium income, high income, somewhat positive attitudes, and positive attitudes) are displayed in table 4.2, which in its entirety portrays a bivariate distribution.

The new table compresses the data, from 682 cells in the data sheet of table 4.1 to 16, and again we can look for patterns in the data. In effect, we are trying to

Chapter 4
Determining Association Among
Variables

compare distributions (for example, across low-income, medium-income, and high-income households) and if we find that the distributions are different (across income levels, for example) we will conclude that attitude toward energy conservation is associated with income level. Specifically, households with high income tend to be less positive than low-income households. But the comparisons are difficult because the number of households in the categories (for example, low-income and medium-income) are not equal, as we can observe from the row and column totals.

The next step in trying to understand the data is to convert the numbers in table 4.2 to percentages. That will eliminate the effects of different numbers of households in different categories. There are three ways to make the conversion: (1) make each number in a row a percentage of the row total, (2) make each number in a column a percentage of a column total, or (3) make each number in the table a percentage of the batch total, 341. (Computer programs may readily compute all three variations.) In table 4.3, we have chosen the second way. Now we can see much more clearly how the distributions compare for different income levels.

Table 4.3: Percentaged Cross-Tabulation of Two Ordinal Variables

Attitude toward energy conservation	Family income level			Total
	Low	Medium	High	
Indifferent	26	34	44	35
Somewhat positive	33	36	32	34
Positive	41	30	24	31
Total	100	100	100	100

And we could go on and look at the other ways of computing percentages. But even with all three displays, it still may not be easy to grasp the extent of an association, much less readily communicate its extent to another person. Therefore, we often want to go beyond tabular displays and seek a number, a measure of association, to summarize the association. Such a measure can be used to characterize the extent of the relationship and, often, the direction of the association, except for nominal variables. We may sometimes use more than one measure to observe different facets of an association. Although this example involves two ordinal variables, the notion of an association is similar for other combinations of measurement levels.

Measures of Association Between Two Variables

A measure of association between variables is calculated from a batch of observations, so it is another descriptive statistic. Several measures of association are available to choose from, depending upon the measurement level of the variables and exactly how association is defined. For illustrative purposes, we mention four from the whole class of statistics sometimes used for indicating association: gamma, lambda, the Pearson product-moment correlation coefficient, and the regression coefficient.

Ordinal Variables: Gamma

When we have two ordinal variables, as in the energy conservation example, gamma is a common statistic used to characterize an association. This indicator can range in value from -1 to $+1$, indicating perfect negative association and perfect positive association, respectively. When the value of gamma is near zero, there is little or no evident association between the two variables. Gamma is readily produced by available statistical programs, and it can be computed by hand from a table like table 4.2, but the calculation,

sketched out below, is rather laborious. For our hypothetical data set, gamma is found to be $-.24$.

The computed value of gamma indicates that the association between family income level and attitude toward energy conservation is negative but that the extent of the association is modest. One way to interpret this result is that if we are trying to predict a family's attitude toward energy conservation, we will be more accurate (but not much more) if we use knowledge of its income level in making the prediction. The gamma statistic is based upon a comparison of the errors in predicting the value of one variable (for example, family's attitude toward conservation) with and without knowing the value of another variable (family income). This idea is expressed in the following formula: $\text{gamma} = \frac{\text{prediction errors not knowing income} - \text{prediction errors knowing income}}{\text{prediction errors not knowing income}}$.

The calculation of gamma involves using the information in table 4.2 to determine the number of prediction errors for each of two situations, with and without knowing income. The formula above is actually quite general and applies to a number of measures of association, referred to as PRE (proportionate reduction in error) measures. The more general formulation (Loether and McTavish, 1988) is $\text{PRE measure} = \frac{\text{reduction in errors with more information}}{\text{original amount of error}}$. PRE measures vary, depending upon the definition of prediction error.

Nominal Variables:
Lambda

With two nominal variables, the idea of an association is similar to that between ordinal variables but the approach to determining the extent of the association is a little different. This is so because, according to definition, the attributes of a nominal variable are not

Chapter 4
Determining Association Among
Variables

ordered. The consequences can be seen by looking at another cross-tabulation.

Suppose we have data with which to answer the question about the association between whether the prices farmers receive are affected by government crop supports and the region of the country in which they live. Then the variables and attributes might be as follows: crop supports (yes, no); region of the country (Northeast, Southeast, Midwest, Southwest, Northwest). Some hypothetical data for these variables are displayed in table 4.4.

Table 4.4:
Cross-Tabulation of Two
Nominal Variables

Region	Prices affected by crop		Total
	Yes	No	
Northeast	322	672	994
Southeast	473	287	760
Midwest	366	382	748
Southwest	306	297	603
Northwest	342	312	654
Total	1,809	1,950	3,759

If we start to look for a pattern in this cross-tabulation, we have to be careful because the order in which the regions are listed is arbitrary. We could just as well have listed them as Southwest, Northeast, Northwest, Midwest, and Southeast or in any other sequence. Therefore, the pattern we are looking for cannot depend upon the sequence as it does with ordinal variables.

Chapter 4
Determining Association Among
Variables

Lambda is a measure of association between two nominal variables. It varies from 0, indicating no association, to 1, indicating perfect association.² The calculation of lambda, which is another PRE measure like gamma, involves the use of the mode as a basis for computing prediction errors. For the crop support example, the computed value of lambda is .08.³ This small value indicates that there is not a very large association between crop-support effects and region of the country.

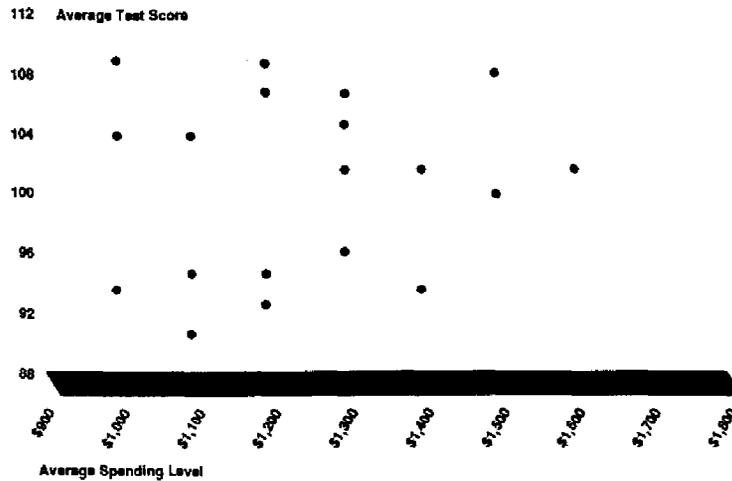
²A definition of perfect association is beyond the scope of this paper. Different measures of association sometimes imply different notions of perfect association.

³There are actually three ways to compute lambda. The numerical value here is the symmetric lambda. There is some discussion of symmetric and asymmetric measures of association later in this paper.

Interval-Ratio
Variables: Correlation
and Regression
Coefficients

A Pearson product-moment correlation coefficient is a measure of linear association between two interval-ratio variables.⁴ The measure, usually symbolized by the letter r , varies from -1 to $+1$, with 0 indicating no linear association. The square of the correlation coefficient is another PRE measure of association.

Figure 4.1: Scatter Plots for Spending Level and Test Scores



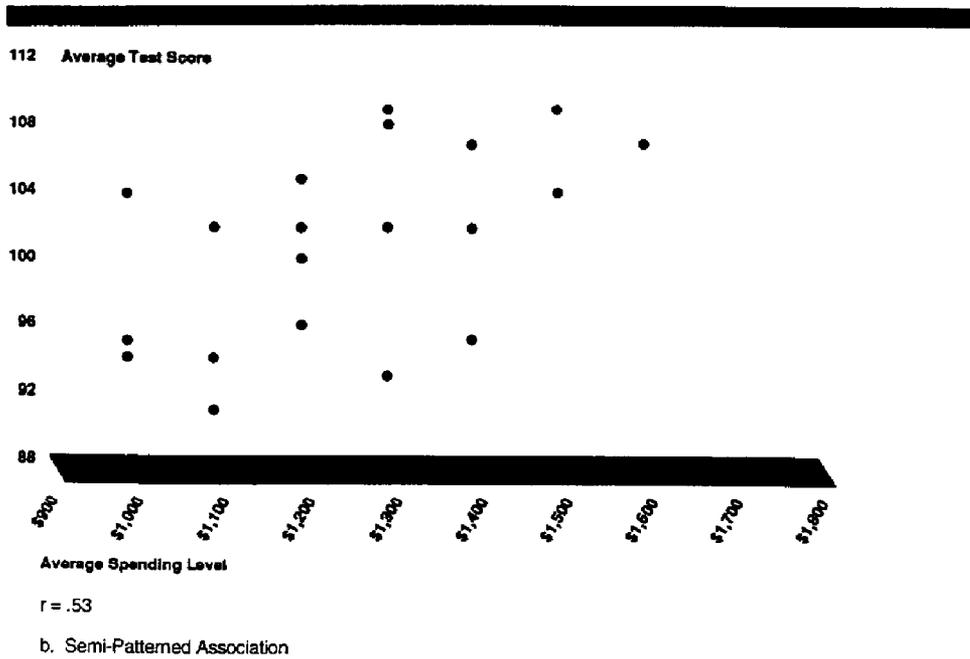
$r = .12$

a. No Pattern of Association

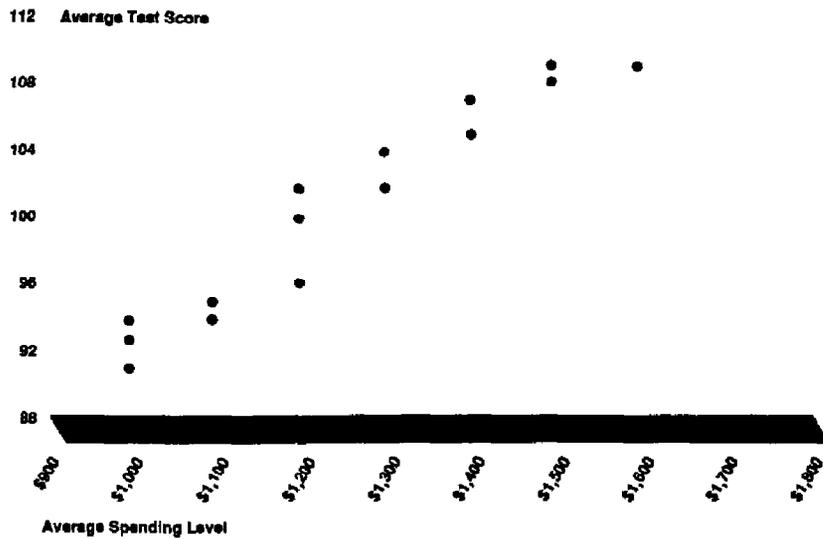
⁴The word "correlation" is sometimes used in a nonspecific way as a synonym for "association." Here, however, the Pearson product-moment correlation coefficient is a measure of linear association produced by a specific set of calculations on a batch of data. It is necessary to specify linear because if the association is nonlinear, the two variables might have a strong association but the correlation coefficient could be small or even zero. This potential problem is another good reason for displaying the data graphically, which can then be inspected for nonlinearity. For a relationship that is not linear, another measure of association, called "eta," can be used instead of the Pearson coefficient (Loether and McTavish, 1988).

Chapter 4
Determining Association Among
Variables

The Pearson product-moment correlation coefficient can be illustrated by considering the question about the association between students' achievement level in the 12th grade and school district spending level, regarding both variables as measured at the interval-ratio level. Such data are often displayed in a scatter plot, an especially revealing way to look at the association between two variables measured at the interval-ratio level. Figure 4.1 shows three scatter plots for three sets of hypothetical data on two variables: average test scores for 12th graders in a school district and the per capita spending level in the district. Each data point represents two numbers: a districtwide test score and a spending level.



Chapter 4
Determining Association Among
Variables



$r = .96$

c. Linear Pattern of Association

In figure 4.1a, which shows essentially no pattern in the scatter of points, the correlation coefficient is .12. In figure 4.1b, the points are still widely scattered but the pattern is clear—a tendency for high test scores to be associated with high spending levels and vice versa; the correlation coefficient is .53. And finally, in figure 4.1c the linear pattern is quite pronounced and the correlation coefficient is .96.

The regression coefficient is another widely used measure of association between two interval-ratio variables and it can be used to introduce the idea of an asymmetric measure. First, we use the scatter plot data from figure 4.1b and replot them in figure 4.2. Using the set of data represented by the scatter plot, we can use the method of regression analysis to

Chapter 4
Determining Association Among
Variables

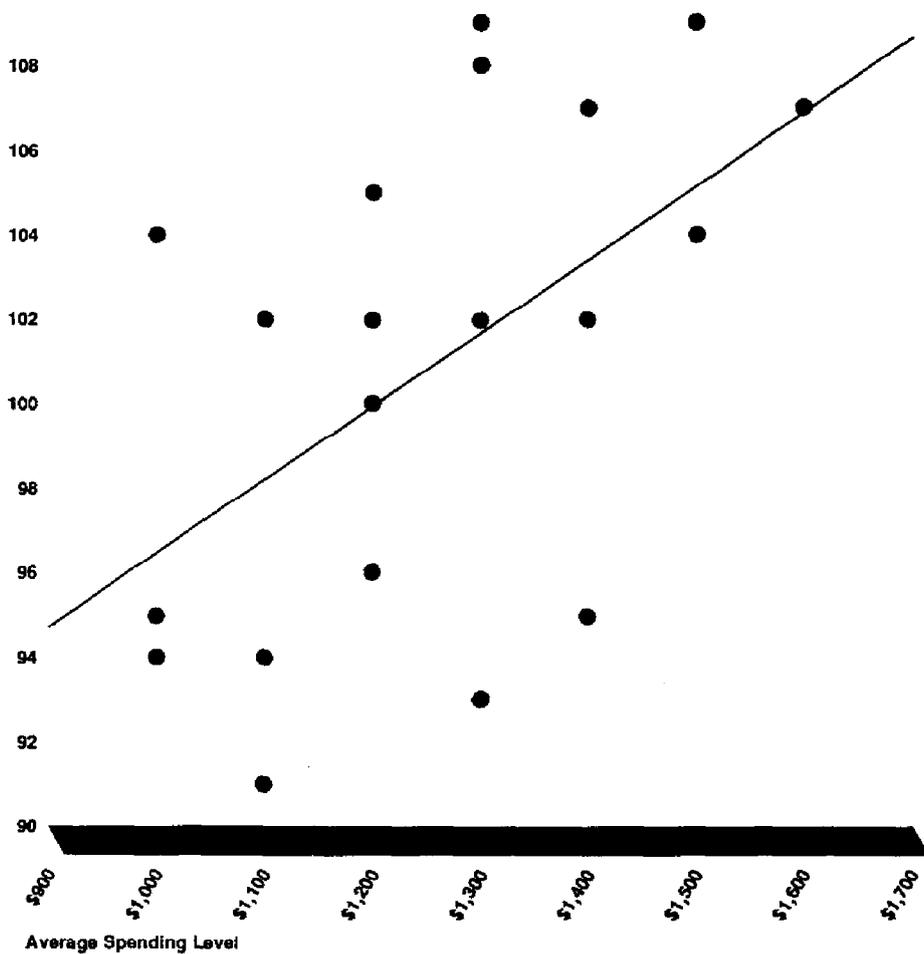
“regress Y on X,” which tells us where to position a line through the scatter plot.⁵ How the analysis works is not important here, but the interpretation of the line as a measure of association is. The slope of the line is numerically equal to the amount of change in the Y variable per 1 unit change in the X variable. The slope is the regression coefficient, an asymmetric measure of association between the two variables. Unlike many other commonly used measures, the regression coefficient is not limited to the interval from -1 and +1.⁶ The regression coefficient for the data displayed in figure 4.2 is 1.76, indicating that a \$100 change in spending level is associated with a 1.76 change in test scores.

⁵Regression analysis is not covered in this paper. For extensive treatments, see Draper and Smith, 1981, and Pedhazur, 1982.

⁶The regression coefficient is closely related to the Pearson product moment correlation. In fact, when the observed variables are transformed to so-called z-scores, by subtracting the mean from each observed value of a variable and dividing the difference by the standard deviation of the variable, the regression coefficient of the transformed variables is equal to the correlation coefficient.

Figure 4.2: Regression of Test Score on Spending Level

110 Average Test Score



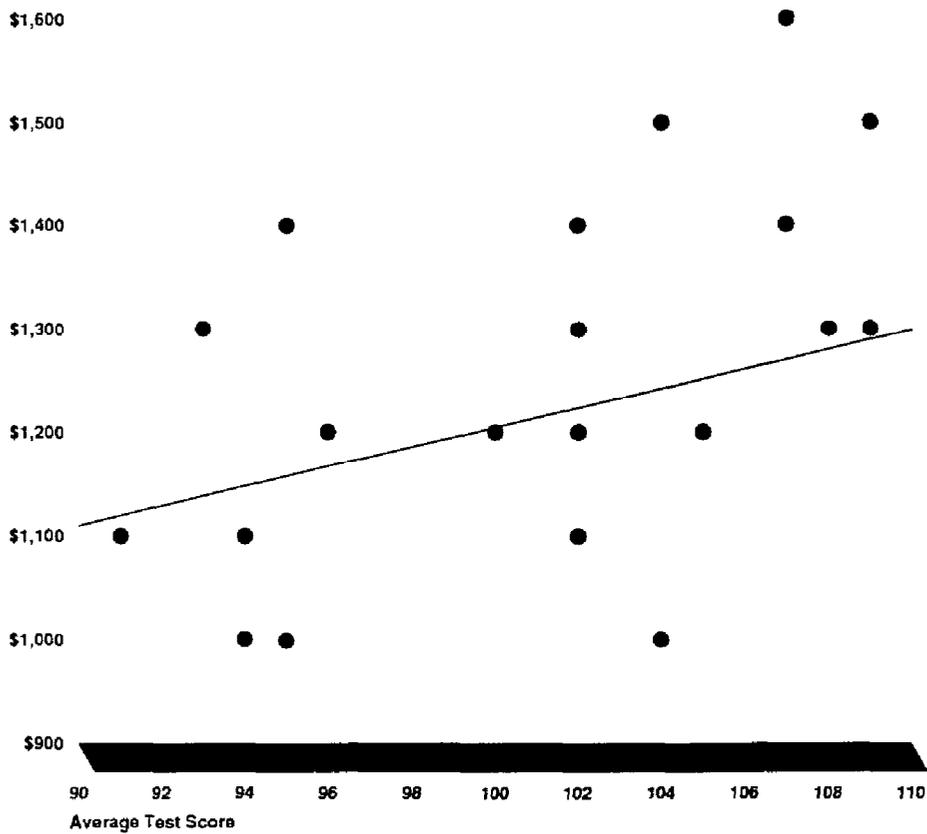
Why the regression coefficient is asymmetric can be understood if we turn the scatter plot around, as in figure 4.3, so that X is on the vertical axis and Y is on

Chapter 4
Determining Association Among
Variables

the horizontal. The pattern of points is a little different now, and if we reverse the roles of the X and Y variables in the regression procedure (that is, "regress X on Y"), the resulting line will have a different slope. Consequently the Y-on-X regression coefficient is different from the X-on-Y coefficient and that is why the measure is said to be asymmetric. Measures of association in which the roles of the X and Y variables can be interchanged in the calculation procedures without affecting the measure are said to be symmetric and measures in which the interchange produces different results are asymmetric.

Figure 4.3: Regression of Spending Level on Test Score

\$1,700 Average Spending Level



When we use asymmetric measures of association, we also use special language to characterize the roles of the variables. Or, put in a more direct way, if we take the view that the variables are playing different roles,

we give them names indicative of the roles. One is called the dependent variable and the other is the independent variable. The language is applied to two kinds of application: (1) when we are trying to establish that the independent variable causes changes in the dependent one and (2) when we are trying to use the independent variable to predict the dependent one, without necessarily supposing the association is causal. In either case, the dependent variable in some sense depends upon the independent one. Graphically, the convention is to plot the dependent variable along the vertical axis and the independent variable along the horizontal axis.

Whether evaluators should use an asymmetric measure of association or a symmetric one depends upon the application. If there is no reason to label variables as dependent and independent, then they should use a symmetric measure. But when they are predicting one variable from another or believe that one has a causal effect on the other, an asymmetric measure is preferred.

In each of the foregoing examples, both variables were measured at the same level. That will not always be the case. One common circumstance in which the variables are at different levels is discussed in a section below, entitled "The Comparison of Groups."

Examples

An example of a measure of association between ordinal variables comes from a GAO report on the use of medical devices in hospitals (U.S. General Accounting Office, 1986). In reporting on the association between the seriousness of a device problem and hospitals' actions to contact the manufacturer or some other party outside the hospital, the evaluators displayed the results in cross-tabular array and summarized them using a symmetric measure.

In a study of election procedures (U.S. General Accounting Office, 1990d), some of the major findings were reported as a series of correlation coefficients that showed the association between voter turnout and numerous factors characterizing absentee ballot rules and voter information activities. The same study used a regression coefficient to show the association between voter turnout and the registration deadline, expressed as number of days before the election.

The Comparison of Groups

A situation of special interest arises when evaluators want to compare two groups on some variable to see if they are different. For example, suppose the evaluators want to compare government benefits received by farmers who live east of the Mississippi to those who live west of the Mississippi. Questions about the difference between two groups are very common. In this instance, it would probably be best to answer the question by computing the mean benefits for each group and looking at the difference.

Equivalently, the comparison between these two groups of cases can be seen as a measure of association. With government benefits measured at the interval-ratio level (in dollars) and region of the country measured at the nominal level (for example, 0 for East and 1 for West), we can compute a measure of association called the point biserial correlation between benefits and region.⁷ If we then multiply this correlation by the standard deviation of benefits and divide by the standard deviation of region, we will get the difference between the means of the two groups. The same result would be obtained if we regressed benefits on region; the regression coefficient is equal

⁷The point biserial correlation is analogous to the Pearson product-moment correlation that applies when both variables are measured at the interval-ratio level.

Chapter 4
Determining Association Among
Variables

to the difference between the means of the two groups. We thus have three different, but statistically equivalent, methods of comparing the two groups: (1) computing the difference between means of the groups, (2) computing the point biserial correlation (and then adjusting it), and (3) computing the regression coefficient.

The point is that when evaluators compare two groups, they are examining the extent of association between two variables: one is group membership and the other is the response variable, the characteristic on which the groups might differ. Such comparisons are the main method for evaluating the effect of a program. For example, a question might be: What is the effect of the Special Supplemental Food Program for Women, Infants, and Children (WIC) on birthweight? The answer is partly to be found in the association, if any, between group membership (program participation or not) and birthweight.

Knowing the association is only part of the answer, however, because the question about effect is about the causal association between program participation and birthweight. As we show in chapter 6, the existence of an association is one of three conditions necessary to establish causality.

A comparison of means is but one among many ways in which it might be appropriate to compare two groups. Other possibilities include the comparison of (1) medians, (2) proportions, and (3) distributions. For example, if two groups are being compared on an ordinal variable and the distribution is highly asymmetric, then an analysis of the medians may be preferable to an analysis of the means. Or, as noted in chapter 3, sometimes the question the evaluators are attempting to answer is focused on the spread of a distribution and so they might be interested in comparing a measure of spread in two groups. For the hospital mortality study, we could compare the

Chapter 4
Determining Association Among
Variables

spread of mortality rates between two categories of hospitals, say teaching and nonteaching ones.

Statistical methods for comparing groups are important to GAO in at least three situations: (1) comparison of the characteristics of populations (for example, farmers in the eastern part of the country with those in the western), (2) determination of program effects (for example, the WIC program), and (3) the comparison of processes (for example, different ways to monitor thrift institutions). The questions that arise from these situations lead to a variety of data analysis methods. Factors that determine an appropriate data analysis methodology include (1) the number of groups to be compared, (2) how cases for the groups were selected, (3) the measurement level of the variables, (4) the shape of the distributions, and (5) the type of comparison (measure of central tendency, measure of spread, and so on). A further complexity is that, when sampling, evaluators need to know if the observed difference between groups is real or most likely stems from sampling fluctuation. For making that determination, the methods of statistical inference are required.

A study of changes to the program called Aid to Families with Dependent Children illustrates the use of group comparisons on factors such as employment status to draw conclusions about effects of the changes (U.S. General Accounting Office, 1985). In another example, two groups of farmers, ones who specialized in a few crops and ones who diversified, were compared on agricultural practices (U.S. General Accounting Office, 1990a).

**Analyzing and
Reporting the
Association
Between
Variables**

Answering a question about the association between two variables really involves four subquestions: Does an association exist? What is the extent of the association? What is the direction of the association? What is the nature of the association? Analysis of a batch of data to answer these questions usually involves the production of tabular or graphic displays as well as the calculation of measures of association.

With nominal or ordinal data presented in tabular form, evaluators can check for the existence of an association by inspection of the tables. If the conditional distributions are identical or nearly so, they can conclude that there is no association. Table 4.2 illustrates a data set for which an association exists between income level and attitude toward energy conservation. Table 4.5 shows another set of 341 cases—one in which there is virtually no association. The marginal distributions are the same for tables 4.2 and 4.5, so the pattern of observations can change only in the nine interior cells.

**Table 4.5: Two Ordinal
Variables Showing No
Association**

Attitude toward energy conservation	Family income level			Total
	Low	Medium	High	
Indifferent	37	38	45	120
Somewhat positive	35	37	43	115
Positive	33	34	39	106
Total	105	109	127	341

Chapter 4
Determining Association Among
Variables

Most bivariate data show the existence of association. The question is really whether the association is large enough to be important.⁸ A measure of association is calculated to help answer this question, and evaluators must make a judgment about importance, using the context of the question as a guide.

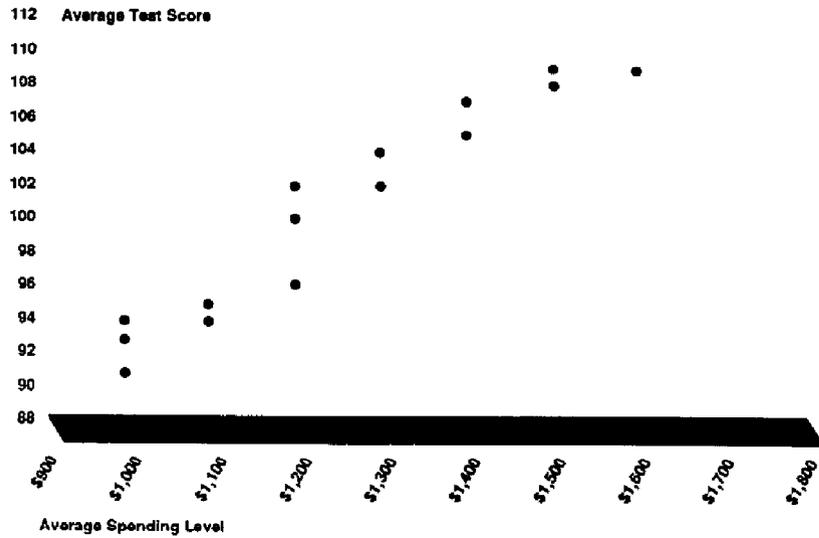
The direction of an association is also given by a measure of association unless the variables are nominal, in which case the direction is not meaningful. Most measures are defined so that a negative value indicates that as one variable increases the other decreases and that a positive value indicates that the variables increase or decrease together.

While the existence, extent, and direction of an association can be revealed by a measure of association, determining the nature of the association requires other methods. Usually it is done by inspecting the tabular or graphic display of a bivariate distribution. For example, a scatter plot will show if the association is approximately linear, a constant amount of change in one variable being associated with a constant amount in the other variable, as in figure 4.4a. However, the scatter plot may show that the association is nonlinear, as in figure 4.4b. Interpretations of the data are usually easier if the data are linear and, of course, interpolations and extrapolations are more straightforward.

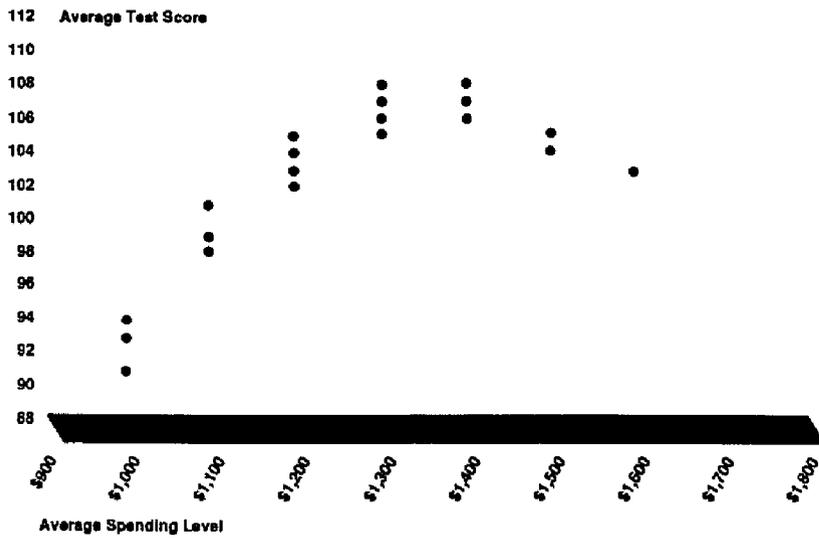
⁸If we are trying to draw conclusions about a population from a probability sample, then we must additionally be concerned about whether what seems to be an association really stems from sampling fluctuation. The data analysis then involves inferential statistics.

Chapter 4
Determining Association Among
Variables

Figure 4.4: Linear and Nonlinear Associations



a. Linear Association



b. Nonlinear Association

Chapter 4
Determining Association Among
Variables

In comparisons between groups of cases, regression analysis is an important tool when the dependent variable is interval-ratio. When the assumptions necessary for regression are not satisfied, other techniques are necessary.⁹

Overall, there are many analysis choices. Evaluators can always find the extent of the association, if any, between the variables and, unless one or both the variables are measured at the nominal level, they can also determine the direction of the association. The appropriateness of a given procedure depends upon the measurement level of the variables and the definition of association believed best for the circumstances. It is also wise to display the data in a table or a graph as a way to understand the form of the association.

How much information from the analysis should be included in a report? The answer depends on how strongly the conclusions are based upon the association that has been determined. If the relationship between the two variables is crucial, then probably both measures of association and tabular or graphic displays should be presented. Otherwise, reporting only the measures will probably suffice. In either case, evaluators should be clear about the level of measurement assumed and analysis methods used.

⁹The assumptions are not very stringent for descriptive statistics but may be problematic for inferential statistics.

Estimating Population Parameters

Many questions that GAO seeks to answer are about relatively large populations of persons, things, or events. Examples are

- What is the average student loan balance owed by college students?
- Among households eligible for food stamps, what proportion receive them?
- How much hazardous waste is produced in the nation annually and how much variation is there among individual generators?
- What is the relationship between the receipt of Medicaid benefits and size of household?

In chapters 2, 3, and 4, we focused on descriptive statistics—ways to answer questions about just those cases for which we had data. We now consider inferential statistics—methods for answering questions about cases for which we do not have observations. The procedures involve using data from a sample of cases to infer conclusions about the population of which the sample is a part.

The shift to inferential statistics is necessary when evaluators want to know about large populations but, for practical reasons, do not try to get information from every member of such populations. The most obvious obstacle to collecting data on many cases is cost, but other factors such as deadlines for producing results may play a role.

To generalize findings from a sample of cases to the larger population, not just any sample of cases will do—a probability sample is required. Random processes for drawing probability samples are

detailed in the transfer paper in this series entitled Using Statistical Sampling.¹ Under such methods, each member of a population has a known, nonzero probability of being drawn.

The methods collectively called inferential statistics are based upon the laws of probability and require samples drawn by a random process. Attempts to draw conclusions about populations based upon nonprobability samples are usually not very persuasive, so we do not consider them here.

From the perspective of inferential statistics, the illustrative questions above need two-part answers: a point estimate of a parameter that describes the population and an interval estimate of the parameter. (Other forms of statistical inference, such as hypothesis testing, are appropriate to other kinds of questions. They are not covered in this paper.) Full understanding of inferential statistical statements requires a thorough knowledge of probability, the development of which is beyond the scope of this introductory paper. For our brief treatment here, we use the concept of the histogram and illustrate how probability comes into play through sampling distributions.

Some notions discussed in earlier chapters, involving data on all cases in a batch, are extended in this chapter to show how statistics computed from a probability sample of cases are used to estimate parameters such as the central tendency of a population (see chapter 2). The notable difference between describing a batch, using statistics from all cases in the batch, and describing a population, using statistics from a probability sample of the population,

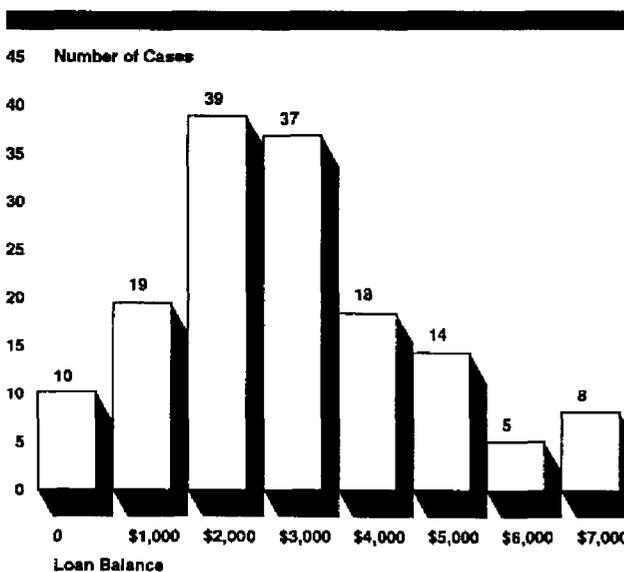
¹Probability sampling is sometimes called statistical sampling or scientific sampling.

is that we will necessarily be somewhat uncertain in describing a population. However, the data analysis methods for inferential statistics allow us to be precise about the degree of uncertainty.

Histograms and Probability Distributions

A key concept in statistical inference is the sampling distribution. The histogram, which was introduced in chapter 1, is a way of displaying a distribution, so we begin there. Expanding on the first example from chapter 1, suppose that instead of information on a batch of 15 college students, we have collected information on loan balances from 150 students. If we round numbers to the nearest \$1,000 for ease of computation and display, our observations produce the distribution of loan balances shown in figure 5.1. For example, the height of the third bar corresponds to the number of students who reported loan balances between \$1,500 and \$2,499. The distribution is somewhat asymmetrical and has a mean of \$2,907.

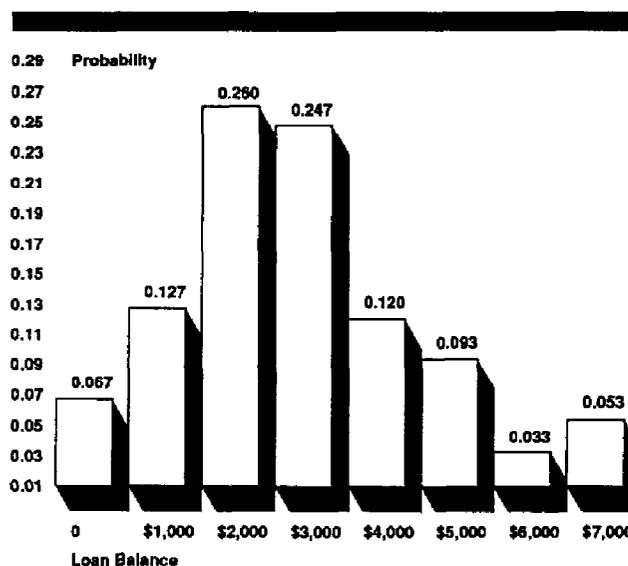
Figure 5.1: Frequency Distribution of Loan Balances



Probability is a numerical way of expressing the likelihood that a particular outcome, among a set of possibilities, will occur. Suppose that we do not have access to the responses from individual students in the survey but that we want to use the distribution in figure 5.1 to make a wager on whether a student to be selected at random from this sample of 150 will have a loan balance between \$1,500 and \$2,499. To make a reasonable bet, we need to know the probability that a particular outcome—a loan value between \$1,500 and \$2,499—will be reported when we make a phone call to the student. The information we need is in the figure but the answer will be more evident if we make a slight change in the display.

We can describe the students' use of loans in probability terms if we convert the frequency distribution to a probability distribution. The frequency distribution shows the number of students who reported each possible outcome (that is, loan balances between \$1,500 and \$2,499 and so on). We can present the same information in terms of percentages by dividing the number of students reporting each outcome (the height of a bar) by the total number in the sample (150). The percentages, expressed in decimal form, can be interpreted as probabilities and are displayed in figure 5.2.

Figure 5.2: Probability
Distribution of Loan
Balances



We now have a probability distribution for the loan balance variable for the sample of 150 students. Picking the outcome we want to make a wager on, we can say that the probability is .26 (39 divided by 150) that a student selected randomly from the sample will report a balance between \$1,500 and \$2,499.

The shape of the probability distribution is the same as the frequency distribution; we have just relabeled the vertical axis. But the probability distribution has two important characteristics not possessed by the frequency histogram: (1) the height of each bar is equal to or greater than 0 and equal to or less than 1 and (2) the sum of the heights of the bars is equal to 1. These characteristics qualify the new display as a

probability distribution for nominal or ordinal variables.² The probability of an outcome is defined as ranging between 0 and 1, and the sum of probabilities across all possible outcomes is 1.

The probability distribution in figure 5.2 is an empirical distribution because it is based on experience. "Theoretical" probability distributions are also important in drawing conclusions from data and deciding actions to take. An example relevant to the decisions that gamblers make is the distribution of possible outcomes from throwing a six-sided die. In theory, the probability distribution for the six possible outcomes could be displayed with six bars, each having a height of 1/6.

Theoretical distributions that play key roles in the methods of inferential statistics are the binomial, normal, chi-square, t, and F distributions. Actually, each of these names refers to a whole family of distributions. The distributions are described in widely available tables that give numerical information about the distributions. For tables and discussions of the distributions, consult a statistics text such as Loether and McTavish (1988). For example, one could use a table of the normal distribution (with mean of 0 and standard deviation of 1) to find the probability that an observation from a population with this distribution could exceed a specified value. Before computers became commonplace for statistical calculations, tables of the distributions were indispensable to the application of inferential statistics.

²Nominal and ordinal variables take on a finite set of values. Interval-ratio variables have a potentially infinite set of values, so the corresponding probability distribution is defined a little differently. (These variables are introduced under "Level of Measurement" in chapter 1.)

Sampling Distributions

The distribution of responses from 150 college students in the example above is the distribution of a sample. If we were to draw another sample of 150 students and plot a histogram, we would almost surely see a slightly different distribution and the mean would be different. And we could go on drawing more samples and plotting more histograms. Differences among the resulting distributions of samples are inherent in the sampling process.

The aim is to be precise about how much variation to expect among statistics computed from different samples. For example, if we use the mean of a sample to describe the distribution of loan balances in a student population, how much uncertainty derives from using a sample? New kinds of distributions called sampling distributions of statistics, or just sampling distributions for short, provide the basis for making statements about statistical uncertainty.

To this point, we have computed statistics without concern for how we produced the data but now we must use probability sampling, which requires that data be produced by a random process. In particular, suppose that we were to draw 100 different simple random samples, each with 150 students, and compute sample statistics, such as the mean, for each sample.³ This would give us a data sheet like that in table 5.1. Since the computed sample means vary across the samples, we could draw a histogram showing the distribution of the sample means (figure 5.3). The midpoint of each bar along the X axis is the midpoint of an interval centered on the number shown. Such a distribution is what we mean by a sampling distribution—one that tells us the probability of obtaining a sample in which a computed statistic,

³There are many kinds of probability samples. The most elementary is the simple random sample in which each member of the population has an equal chance of being drawn to the sample.

such as the mean, will have certain values.⁴ Using figure 5.3, we can say that 25 percent of the sample means had values in the interval \$3,000 plus or minus 50. Using such information, we will be able to make a statement about the probability that a given interval includes the value of the population mean.⁵ This idea is developed further in a later section on interval estimation.

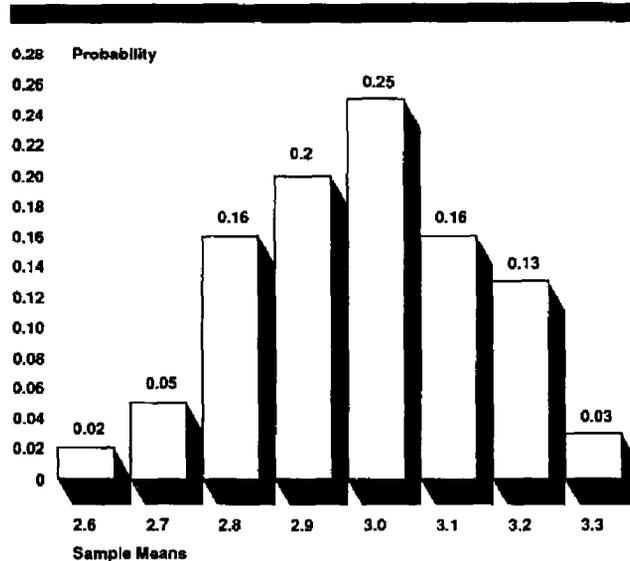
Table 5.1: Data Sheet for 100 Samples of College Students

Sample	Computed mean loan balance
1	\$2,907
2	2,947
3	2,933
4	3,127
5	3,080
...	...
100	3,227

⁴Notice the difference between a sample distribution (the distribution of a sample) and a sampling distribution (the distribution of a sample statistic).

⁵The mean either lies in a given interval or it does not. No probability is involved in that respect. However, the probability statement is appropriate since the population mean is usually unknown and we use the confidence interval as a measure of the uncertainty in our estimate of the mean that stems from sampling.

Figure 5.3: Sampling
Distribution for Mean
Student Loan Balances



Speaking practically, of course, we would not want to draw many samples of college students because a principal reason for sampling, after all, is to avoid having to make a large number of observations. Therefore, we cannot hope actually to produce a distribution like that of figure 5.3 from empirical evidence. But if our sample is a probability sample, we can usually determine the amount of uncertainty associated with sampling and yet draw only one sample. With a probability sample, the laws of probability often enable us to know the theoretical distribution of a sample statistic so that we can use that instead of an empirical distribution obtained by drawing many samples.⁶

⁶This is where families of distributions like the chi-square and the t come into play to help us estimate population parameters. They are the theoretical distributions that we need.

The sample displayed in figure 5.1, as well as the other 99, was, in fact, drawn randomly from a population with a mean loan balance of \$3,000. It can therefore be used to estimate population parameters for the distribution of students.

Population Parameters

A population parameter is a number that describes a population. Consider again the question of the mean student loan balance for college students. We want to know about the population of all college students—specifically, we want to know the mean loan balance—but we do not want to get information from all. We describe the situation by saying that we want to estimate a population parameter—in this case, the mean of the distribution of loan balances for all students. We want a reasonably close estimate but we are willing to tolerate some uncertainty in exchange for avoiding the cost and time of querying every college student.

The idea of a population parameter applies to any variable measured on a population and any single number that might be used to describe the distribution of the variable. For example, if we want to estimate the proportion of households that use food stamps among those eligible to receive them, the population is all the eligible households. The response variable, use of food stamps, is measured at the nominal level and can have only two values: no or yes. (For purposes of statistical analysis, the variable can be coded as no = 0 and yes = 1.) The population parameter in question is the proportion of all eligible households that use food stamps. The proportion of food stamp users is a way of describing the population so it qualifies as a population parameter.

A population might also be described by two or more variables. For example, we might wish to describe the population of U.S. households by “use of Medicaid benefits” and “size of the household.” We can deal

with the two variables individually, estimating the proportion of households that receive Medicaid benefits on the one hand and estimating the median household size on the other, but we can also estimate a measure of association between two variables.

Parameters are associated with populations, and statistics are associated with samples, but the two concepts are linked in that statistics are used to estimate parameters. Two kinds of estimates for population parameters are possible: point estimates and interval estimates. Both kinds of estimates are statistics computed from probability samples. In the following sections, we first give examples of parameter estimates and then discuss what they mean and how they are computed from samples.

Point Estimates of Population Parameters

A point estimate is a statistic, our “best” judgment about the value of the population parameter in question. In the student loan example, we would like to know the mean loan balance for all students. We draw a simple random sample and use the mean of the sample, a statistic, to estimate the unknown population mean. The value of the sample mean, \$2,907, from the first sample of students is a point estimate of the mean of the population.

The statistical practice is that the sample mean is used to estimate the population mean when a simple random sample is used to produce the data. The procedure has intuitive appeal because the sample mean is the analogue to the population mean. That is, the population mean would be the arithmetic average of all members of the population while the sample

mean is the arithmetic average of all members of the sample.⁷

Point estimates are not based on intuition, however. When a sample has been produced by a random process, statistical theory gives us a good way to estimate a population parameter (that is, theory gives us an appropriate sample statistic for estimating the parameter). That is one of the advantages of randomness; by means of statistical theory, the process provides us with a way to make point estimates of parameters. And it should be noted that intuition does not always suggest the best statistic. For example, intuition might say to estimate the standard deviation of a normally distributed population with a sample standard deviation. However, theory tells us that with small samples, the sample standard deviation should not be used to estimate the standard deviation of the population.

Like the mean and standard deviation, other population parameters are estimated from sample statistics. For example, to answer the question about the proportion of households that are eligible for food stamps, we could use the proportion eligible from a simple random sample of households to make a point estimate of the proportion eligible in the population. Study questions might require that we estimate a variety of population parameters, including the spread of a distribution and the association between two variables.

⁷Note that the use of the sample mean to estimate the population mean does not deal with the question, raised in chapter 2, as to the circumstances under which the mean is the best measure of central tendency. When the population distribution is highly asymmetric, the population median may be a better measure of central tendency for some purposes. We would then want point and interval estimates of the median.

One of the important factors determining the choice of a statistic to estimate a population parameter is the procedure used to produce the data—that is, the sample design. To use the methods of statistical inference, the sampling procedure must involve a random process but that still leaves many options (see the transfer paper entitled Using Statistical Sampling). In the student loan example, the sample design was a simple random sample, and that allowed the use of the sample mean to estimate the population mean. With the simple random sample, each student in the population had an equal probability of being drawn to the sample, but as a practical matter such samples are not often used. Instead, commonly used sample designs such as a stratified random sample imply unequal, but known, probabilities so that a weighted sample mean is used to estimate the population mean. The procedures for estimating the population mean then become a little more complicated (for example, we have to determine what weights to use) but the statistical principles are the same.

A point estimate provides a single number with which to describe the distribution of a population. But as we have seen in table 5.1, different samples yield different numerical values that are not likely to correspond exactly to the population mean. We sample because we are willing to trade off a little error in the estimate of the population parameter in exchange for lower cost. But how much error should we expect from our sampling procedure? Interval estimates, the subject of the next section, enable us to describe the level of sampling variability in our procedures.

GAO reports provide numerous illustrations of point estimates of population parameters. The most commonly estimated parameters are probably the mean of a normal distribution and the probability of an event in a binomial distribution. In a study of the

Food Stamp program, for example, the probability of program participation was estimated for all eligible households and many subcategories of households (U.S. General Accounting Office, 1990b). The estimates were based upon a nationally representative sample of 7,061 households. A study of bail reform estimated the mean number of days in custody for two groups of felony defendants in four judicial districts (U.S. General Accounting Office, 1989b). Population means for two 6-month periods in 1984 and 1986 were estimated from two probability samples of 605 and 613 defendants, respectively.

Interval Estimates of Population Parameters

Point estimates of population parameters are commonly made and, indeed, sometimes only point estimates are made. That is unfortunate because point estimates are apt to convey an unwarranted sense of precision. A point estimate should be accompanied by interval estimates to show the amount of variability in the point estimate.

An interval estimate of a population parameter is composed of two numbers, called lower and upper confidence limits, each of which is a statistic. For example, an interval estimate of mean student loan balance is \$2,625 and \$3,189, corresponding to the two limits. Formulas for computing confidence limits are known for many population parameters (see statistical texts such as Loether and McTavish, 1988).

To interpret an interval estimate properly, we need to imagine drawing multiple samples. Following our student loan example, we can suppose that we have 100 samples and construct an expanded version of the data sheet in table 5.1. The interval based on the first sample is in row 1 of table 5.2 and we have computed intervals for each of the 5 other samples in the display. If the table were completely filled out, we would have estimates for 100 intervals just as we have 100 point estimates.

Table 5.2: Point and Interval Estimates for a Set of Samples

Sample	Point estimate	Interval estimate
1	\$2,907	\$2,625-3,189
2	2,947	2,667-3,227
3	2,933	2,647-3,219
4	3,127	2,947-3,397
5	3,080	2,810-3,350
...
100	3,227	2,959-3,595

An interval estimate has the following interpretation: among all the interval estimates made from many samples of a population, approximately P percent will enclose the true value of the population parameter. The value of P is the confidence level and is frequently set at .95. With respect to the interval estimates in table 5.2, this means that approximately 95 out of 100 intervals are boundaries of the true value of the population parameter.

Because we do not actually draw 100 samples, we must now translate the foregoing reasoning to the situation in which we draw a single sample. Suppose it is sample 1 in table 5.2. This sample produced lower and upper bounds of \$2,625 and \$3,189. Following the reasoning above and assuming this is the only sample drawn, we would say that we are 95-percent confident that the mean loan balance is between \$2,625 and \$3,189. That is, applying the interval-estimating procedure to all possible samples, a statement that a given interval enclosed the mean would be correct 95 percent of the time. Therefore, for our single sample, we are justified in claiming that we are 95-percent sure that it embraces the true population mean. We must always admit that if we are unlucky, our estimate based upon sample 1 might be one of the 5 percent that does not bound the population mean.

To make an interval estimate, we choose a confidence level and then use the value to calculate the confidence limits. P can be any percentage level, up to almost 100, but by convention it is usually set at 90 or 95. The larger the value of P , the wider will be the interval estimate. In other words, to increase the likelihood that an interval will “cover” the population parameter, the interval must be widened.

The interval estimate has intuitive appeal because when the confidence level is high, say 95 percent, we feel that the population parameter is somewhere within the interval—even though we know that it might not be.

As in our discussion of sampling distributions, the idea of drawing multiple samples is only to further our understanding of the underlying principle. To actually make an interval estimate, we draw one sample and use knowledge of probability and theoretical sampling distributions to compute the confidence limits. For example, we know from the central limit theorem of probability theory that if the sample size is relatively large (say greater than 30), then the sampling distribution of sample means is distributed approximately as a normal distribution, even if the distribution of the population is not.⁸ Then we can use formulas from probability theory and published tables for the t distribution to compute the lower and upper confidence limits. Of course, in practice the calculations are usually carried out on a computer that is simply given instructions to carry out all or most of the steps necessary to produce an interval estimate from the sample data. It should be noted, however, that the computer does not know whether the data were produced by a random process. It will

⁸Notice that although the distribution of loan balances in figure 5.2 is somewhat asymmetric, the sampling distribution is more symmetric.

analyze any set of data; the analyst is responsible for ensuring that the assumptions of the methodology are satisfied.

Statistical analyses similar to the one just outlined for estimating a population mean can be used to estimate other parameters such as the spread in the amount of hazardous waste produced by generators or the association between the use of Medicaid benefits in a household and the size of the household. For the hazardous waste question, we might obtain an interval estimate of the standard deviation (see "Measures of the Spread of a Distribution" in chapter 3), and for the Medicaid question we probably would make an interval estimate for the point biserial correlation (see "The Comparison of Groups" in chapter 4).⁹

An interval estimate allows us to express the uncertainty we have in the value of a population parameter because of the sampling process but it is important to remember that there are other sources of uncertainty. For example, measurement error may substantially broaden the band of uncertainty regarding the value of a parameter.

The GAO studies cited earlier as illustrating point estimates also provide examples of interval estimates. Confidence intervals were estimated for the probability of Food Stamp program participation (U.S. General Accounting Office, 1990b) and for the mean days spent in custody by felon defendants (U.S. General Accounting Office, 1989b).

⁹Obtaining an interval estimate for the standard deviation is highly problematic because, unlike the case of the mean, the usual procedures are invalid when the distribution of the variable is not normal.

Determining Causation

“Correlation does not imply causation” is a commonly heard cautionary statement about a correlation or, more generally, an association between two variables. But causation does imply association. That is, if two variables are causally connected, they must be associated—but that is not enough. In this chapter, we consider the evidence that is necessary to answer questions about causation and, briefly, some analytical methods that can be brought to bear. In other words, we address the fourth and final generic question in table 1.1.

The following example, similar to one given in chapter 4, is a question framed in causal terms:

- Are homeowners’ appliance-purchasing decisions affected by government information campaigns aimed at reducing energy consumption? (Note the substitution of “affected by” for “associated with.”)

Some related questions can be imagined:

- Are homeowners’ attitudes about energy conservation influenced by their income level?
- Do homeowners purchase energy-efficient appliances as a consequence of government-required efficiency labels?
- Is the purchase of energy-efficient appliances causally determined by homeowners’ income level?

If it is possible to collect quantitative information on such issues, statistical analysis may play a role in drawing conclusions about causal connections.

What Do We Mean by Causal Association?

What does it mean to say that homeowners’ decisions to purchase energy-efficient appliances are affected by a government information campaign? It means that

the campaign in some sense determines whether the homeowners purchase energy-efficient appliances.¹ There is thus a link between the campaign and the purchase decisions. To claim a causal link is to claim that exposure to the campaign influences the likelihood that homeowners will purchase energy-efficient appliances. Three aspects of causal links have a bearing on how we analyze the data and how we interpret the results.

First, an association between two variables is regarded as a probabilistic one. For most of GAO's work, associations are not certain. For example, most people exposed to the government energy information campaign might purchase energy-efficient appliances but some might not. So knowing the attribute for one variable does not allow us to predict the attribute of the other variable with certainty. In this paper, we assume that the cause-and-effect variables are expressed numerically with the consequence that statistical methods can be used to analyze probabilistic associations. In particular, measures of association indicate the strength of causal connections.

Second, a causal association is temporally ordered. That is, the cause must precede the effect in time. Perhaps income causes attitude about conservation

¹The exact nature of causation, both physical and social, is much debated. We do not delve into the intricacies in this paper. There are many detailed discussions of the issues; Bunge (1979) and Hage and Meeker (1988) are two.

Chapter 6
Determining Causation

or, conceivably, attitude causes income—but it does not work both ways at exactly the same time.² This means that, for causation to be established, the relationship between two variables must be asymmetric, whereas a measure of association between two variables can be either symmetric or asymmetric. In statistical language, the direction of causality is expressed by referring to the cause variable as the independent variable and the effect variable as the dependent variable. Measures like those in chapter 4, if they are asymmetric ones, are used to characterize the association between dependent and independent variables.

Third, we must assume that an effect has more than one cause or that a cause has more than one effect. In the real world, a causal process is seldom if ever limited to two variables. It seems likely that a number of factors would influence a homeowner's purchasing decisions—knowledge acquired from the government information program perhaps, but also maybe income and educational level. It is also likely that the decision would vary by the homeowners' age, gender, place of residence, and probably many more factors. In trying to determine the extent of causal association between any two variables, we have to consider a whole network of associations. If we look only at the association between exposure to the government program and the purchase decision, we are likely to draw the wrong conclusion.

²The asymmetry feature does not rule out reciprocal effects in the sense that first attitude affects income, then income affects attitude, and so on.

Evidence for Causation

Thus, determining the causal connection between two variables is a formidable task involving a search for evidence on three conditions: (1) the association between two variables, (2) the time precedence between them, and (3) the extent to which they have been analyzed in isolation from other influential variables.³ In short, analyzing causation requires evidence on the association and time precedence of isolated variables.⁴

When we speak of evidence about the association between two variables, we mean simply that we can show the extent to which a variable X is associated with another variable Y. Asymmetric measures of association provide the necessary evidence. If we treat X as the independent variable and Y as the dependent one, compute an appropriate measure, and find that it

³The three conditions are almost uniformly presented as those required to "establish" causality. However, the language varies from authority to authority. This paper follows Bollen (1989) in using the concept of isolation rather than that of nonspuriousness, the more usually employed concept.

⁴In this chapter, we discuss evidence for a causal relationship between quantitative variables and methods, as used in program evaluation and the sciences generally, for identifying causes. The word "cause" is used here in a more specific way than it is used in auditing. There, "cause" is one of the four elements of a finding, and the argument for a causal interpretation rests essentially on plausibility rather than on establishing time-ordered association and isolating a single cause from other potential ones. The methods described in this paper may help auditors go beyond plausibility arguments in the search for causal explanations. See U.S. General Accounting Office, *Government Auditing Standards* (Washington, D.C.: 1988), standard 11 on page 6-3 and standards 21-24 on page 7-5.

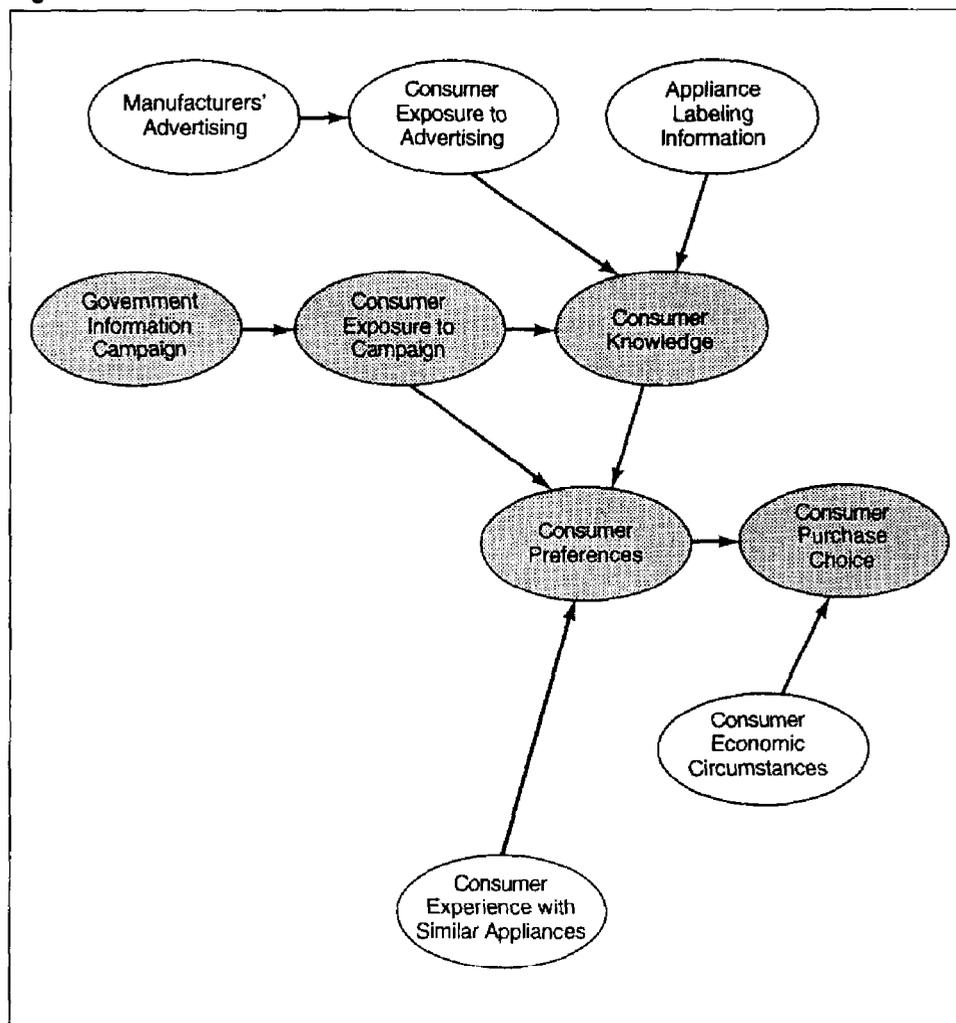
is sufficiently different from zero, we will have evidence of a possibly causal relation.⁶ For example, if we had data on whether homeowners were exposed to a government program that provided energy information and the extent to which they have purchased energy-efficient appliances, we could compute a measure of association between the two variables. A simple association between two variables is usually not sufficient, however, because other variables are likely to influence the dependent variable, and unless we take them into account, our estimate of the extent of the causal association will be wrong.

Taking account of the other variables means determining the association between X and Y in isolation from them. This is necessary because in the real world, as we have noted, the two variables of interest are ordinarily part of a causal network—with perhaps many associated variables including several causal links.

Figure 6.1 shows a relatively small network of which our two variables, consumer-exposure-to-campaign and consumer-purchase-choice, are a part. The arrows in the network indicate possible causal links. The government information campaign plus variables that may be affected by it are indicated by shaded areas. Other variables that may influence the consumer's choice of appliance are represented by unshaded areas.

⁶Judgment is applied in deciding the magnitude of a "sufficient difference."

Figure 6.1: Causal Network



Consumer-exposure-to-campaign and consumer-purchase-choice may indeed have an underlying causal association, but the presence of the

other variables will distort the computed association unless we isolate the variables. That is, the computed amount of the association between X and Y may be either greater or less than the true level of association unless we take steps to control the influence of the other variables. Control is exerted in two ways: by the design of the study and by the statistical analysis.

Finally, we must also have evidence for time precedence, which means that we must show that X precedes Y in time. If we can show that the appliance purchases always came after exposure to the information program, then we have evidence that X preceded Y. Note that the use of asymmetric measures of association does not ensure time precedence. We can compute asymmetric measures for any pair of variables. Evidence for time precedence comes not from the statistical analysis but, rather, from what we know about how the data on X and Y were generated.

Determining the association between two variables is usually not much of a technical problem because computer programs are readily available that can calculate many different measures of association. Establishing time precedence can sometimes be difficult, depending in part upon the type of design employed for the study. (See the transfer paper entitled Designing Evaluations.) For example, with a cross-sectional survey, it may not be easy to decide which came first—a consumer's preference for certain appliances or exposure to a government information program. But with other designs, like an experiment that exposes people to information and then measures their preference, the evidence for time precedence may be straightforward.

Most difficulties in answering causal questions (sometimes called "impact" questions) stem from the requirement to isolate the variables. In fact, it is never possible to totally isolate two variables from all other possible influences, so it is not possible to be

absolutely certain about a causal association. Instead, the confidence that we can have in answering a causal question is a matter of degree—depending especially upon the design of the study and the kind of data analysis conducted.

The key task of isolating two variables—or, in other words, controlling variables that confound the association we are interested in—can be approached in a variety of ways. Most important is the design for producing the data. For simplicity, consider just two broad approaches: experimental and nonexperimental designs.

In the most common type of experiment, we form two groups of subjects or objects and expose one group to a purported cause while the other group is not so exposed. For example, one group of homeowners would be exposed to a government information campaign about energy conservation and another group would not be exposed. In data analysis terms, we would thus have a nominal, independent variable (X), usually called a treatment, that has two attributes: exposure-to-the-campaign and nonexposure-to-the-campaign. If the groups are formed by random assignment, the design is called a true experiment; otherwise, it is called a quasi-experiment.

In answering a causal question based upon experimental data, our basic logic is to compare what happens to the dependent variable Y when the purported cause is present ($X = 1$) with what happens when it is absent ($X = 0$). For example, we could compare the overall proportion of energy-efficient appliances purchased by the two groups. In a true experiment, isolation is achieved by the process of random assignment, which ensures that the two groups are approximately equivalent with respect to all variables, except X , that might affect the purchase of appliances. In this sense, the variables Y

and X have been isolated from the other variables and a measure of association between Y and X can be taken as a defensible indicator of cause and effect.

In a quasi-experiment, random assignment is not used to form the two groups but, rather, they are formed or chosen so that the two groups are as similar as possible. The quasi-experimental procedure, while imperfect, can isolate X and Y to a degree and may provide the basis for estimating the extent of causal association.

In a nonexperimental design, there is no effort to manipulate the purported cause, as in a true experiment, or to contrive a way to compare similar groups, as with a quasi-experiment. Observations are simply made on a collection of subjects or objects with the expectation that the individuals will show variation on the independent and dependent variables of interest. Sample surveys and multiple case studies are examples of nonexperimental designs that could be used to produce data for causal analysis.⁶ For example, we might conduct telephone interviews with a nationally representative sample of adults to learn about their attitudes toward energy conservation and the extent to which they are aware of campaigns to reduce energy use. The designs for sample surveys and case studies do not isolate the key variables, so the burden falls on the data analysis, a heavy burden indeed.

Two broad strategies for generating evidence on the association and time precedence of isolated variables are available: experimental and nonexperimental. Using such evidence, data analysis aimed at

⁶Sample surveys and case studies can be used in conjunction with experimental designs. For example, a sample survey could be used to collect data from the population of people participating in an experiment.

determining causation can be carried out in a variety of ways. As noted, we are here assuming that the data are quantitative. Other approaches are necessary with qualitative data. (Tesch, 1990, describes computer programs such as AQUAD and NUDIST that have some capability for causal analysis of qualitative data.)

**Causal Analysis of
Nonexperimental
Data**

All analysis methods involve determining the time order and the extent of any association between two variables while attempting to isolate those two variables from other factors. While it might be tempting just to compute an asymmetric measure of association between the variables—for example, by determining the regression coefficient of X when Y is regressed on X—such a procedure would almost always produce misleading results. Rather, it is necessary to consider other variables besides X that are likely to affect Y.

The preferred method of analysis is to formulate a causal network—plausible connections between a dependent variable and a set of independent variables—and to test whether the observed data are consistent with the network.⁷ There are many related ways of doing the testing that go by a variety of names, but structural equation modeling seems the currently preferred label.⁸

Two distinct steps are involved in structural equation modeling. The first is to put forth a causal network that shows how the variables we believe are involved

⁷Unless the causal network is an unusually simple one, just adding additional variables to the regression equation is not an appropriate form of analysis.

⁸Other common names for the methods are “analysis of covariance structures” and “causal modeling.”

in a causal process relate to one another (figure 6.1). The network, which can take account of measurement error, should be based on how we suppose the causal process works (for example, how a government program X is intended to bring about a desirable outcome Y). This preliminary understanding of causation is usually drawn from evidence on similar programs and from more general research evidence on human behavior and so on.

The second step is to analyze the data, using a series of linear equations that are written to correspond to the network. Computer programs, such as LISREL and EQS, are used to compare the data on the observed variables with the model and to produce measures of the extent of causal association among the variables. The computer programs also produce indicators of the degree of "fit" between the model and the data. If the fit is not "good" enough, the causal network may be reformulated (step 1) and the analysis (step 2) carried out again. The analyst may cycle through the process many times.

A good fit between the model and the data implies not that causal associations estimated by structural equation modeling are correct but just that the model is consistent with the data. Other models, yet untested, may do as well or better.

With data generated from nonexperimental designs, the statistical analysis is used in an effort to isolate the variables. With experimental designs, an effort is made to collect the data in such a way as to isolate the variables.

**Causal Analysis of
Experimental Data**

As noted earlier, in a true experiment, random assignment of subjects or objects to treatment and comparison groups provides a usually successful way to isolate the variables of interest and, thus, to

produce good answers to causal questions.⁹ In a quasi-experiment, the comparison group is not equivalent (in the random assignment sense) to the treatment group, but if it is similar enough, reasonably good answers to causal questions may be obtained.

The usual ways to analyze experimental data are with techniques such as analysis of variance (ANOVA), analysis of covariance (ANCOVA), and regression. Regression subsumes the first two methods and can be used when the dependent variable is measured at the interval level and with independent variables at any measurement level. If the dependent variable is measured at the nominal or ordinal level, other techniques such as logit regression are required.¹⁰

Although the experimental design is used in an effort to isolate the variables, the objective is never perfectly achieved. Quasi-experimental designs, especially, may admit alternative causal explanations. Therefore, structural equation modeling is sometimes used to

⁹An experiment ordinarily provides strong evidence about causal associations because the process of random assignment ensures that the members of treatment and control groups are approximately equivalent with respect to supplementary variables that might have an effect on the response variable. Being essentially equivalent, almost all variables except the treatment are neutralized in that treatment and control group members are equally affected by those other variables. For example, even though a variable like a person's age might affect a response variable such as health status, random assignment would ensure that the treatment and control groups are roughly equivalent, on the average, with respect to age. In estimating the effect of a health program, then, the evaluator would not mistake the effect of age on health condition for a program effect.

¹⁰The line between ordinal and interval data is not hard and fast. For example, many analysts with a dependent variable measured at the ordinal level use regression analysis if they believe the underlying variable is at the interval level (and limited only to ordinal because of the measuring instrument).

analyze experimental data to further control the variables.

Limitations of Causal Analysis

Statistics texts cover the many assumptions and limitations associated with quantitative analysis to determine causation. The bibliography lists several that give detailed treatments of the methods. However, two more general points need to be made.

First, some effects may be attenuated or changed because of the settings in which they occur—that is, whether the causal process happens in a natural way or is “forced.” In a natural setting, X may have a strong causal influence on Y, but if the setting is artificial, the connection may be different. For example, homeowners who are provided information indicating the advantages of conserving energy (X) may decline to take energy-saving steps (Y) if they are part of a designed experiment. However, the same homeowners might adopt conservation practices if they sought out the information on their own. Strictly speaking, the nature of the X variable is different in these two situations but the point is still the same: the causal process may be affected by differences, sometimes subtle, between the experimental and natural conditions. For some variables, the causal link might be strongest in the natural setting, but for other variables it might be strongest in the experimental setting.

The second point is that causal processes may not be reversible or they may revert to an original state slowly. To illustrate, suppose that laws to lower the legal age for drinking alcohol have been shown to cause a higher rate of automobile accidents. It does not necessarily follow that subsequent laws to raise the drinking age will produce lower accident rates. Evidence to show the effect of increasing (decreasing) a variable cannot, in general, be used to support a

Chapter 6
Determining Causation

claim about the effects of decreasing (increasing) the variable.

Avoiding Pitfalls

Basic ideas about data analysis have been presented in the preceding chapters. Several methods for analyzing central tendency, spread, association, inference from sample to population, and causality have been broadly described. In keeping with the approach in the rest of the paper, this chapter offers advice at a general level with the understanding that specific strategies and cautions are associated with particular methods.

Attention to data analysis should begin while evaluators are formulating the study questions, and in many instances it should continue until they have made the last revisions in the report. Throughout this time, they have many opportunities to enhance the analysis or to make a misstep that will weaken the soundness of the conclusions that may be drawn.

Analysis methods are intertwined with data collection techniques and sampling procedures so that decisions about data analysis cannot be made in isolation. During the planning stages of a study, evaluators must deal with all three of these dimensions simultaneously; after samples have been drawn and data collected, analysis methods are constrained by what has already happened. If it were necessary to summarize advice in a single word, it would be: anticipate.

In the Early Planning Stages

Be clear about the question. As a study question is being formulated and refined, it helps to think through the implications for data analysis. If evaluators cannot deduce data analysis methods from the question or if the question is so vague as to lead to a variety of possible approaches, then they probably need to restate the question or add some additional statements to elaborate upon the question. For example, a question might be: To what extent have the objectives of the dislocated worker program been achieved?

By one reading of this question, the appropriate analysis would simply be to determine the extent to which dislocated workers have found new employment at a rate in excess of (or less than) program goals. With proper sampling and data collection, the analysis would be a matter of computing the proportion of a pool of workers who found reemployment and compare that proportion to the goal for the program. This analysis would not permit the policymaker to draw conclusions about whether the program contributed to the achievement of the goal, because the influence of other factors that might affect the reemployment rate have not been considered.

By another reading, the question implies making a causal link between the government program and the proportion of displaced workers who find reemployment. This means that the design and the analysis must contend with the three conditions for causality discussed in chapter 6. For example, an effort must now be made to isolate the two variables, the program and the reemployment rate, from other variables that might have a causal connection with the reemployment rate. The two interpretations of the question are quite different, and so the question must be clarified before work proceeds.

Understand the subject matter. Evaluators usually need in-depth knowledge of the subject matter to avoid drawing the wrong conclusion from a data set. Numbers carry no meaning except that which derives from how the variables were defined. Moreover, data are collected in a social environment that is probably changing over time. Consequently, there is often an interplay between the subject matter and data analysis methods.

An example using medical data illustrates the importance of understanding the phenomena behind the numbers. Mortality rates for breast cancer among

younger women show some decline over time. However, it would be wrong to draw conclusions about the efficacy of treatment on this evidence alone. It is necessary to understand the details of the process that is producing the numbers. One important consideration is that diagnostic techniques have improved so that cancers are detected at an earlier stage of development. As a consequence, mortality rates will show a decline even if treatment has not improved. A data analysis aimed at determining change in mortality from changes in treatment must adjust for the “statistical artifact” of earlier detection. (For an elaboration of this example, see U.S. General Accounting Office, 1987a.)

The need to understand the subject matter implies a thorough literature review and consultation with diverse experts. It may also mean collecting supplementary data, the need for which was not evident at the outset of the study. For example, in a study of cancer mortality rates, it would be necessary to acquire information about the onset of new diagnostic procedures.

Develop an analysis plan. The planning stage of a project should yield a set of questions to be answered and a design for producing the answers. A plan for analysis of the data should be a part of the design.

Yin (1989) has observed that research designs deal with logical problems rather than logistical problems. So it is with the analysis plan—it should carry forward the overall logic of the study so that the connection between the data that will be collected and the answers to the study questions will become evident. For example, if a sample survey will be used to produce the data, the analysis plan should explain the population parameters to be estimated, the analysis methods, and the form of reporting. Or, if a field experiment will produce the data, the plan should explain the comparisons to be made, the analysis

methods to be used, any statistical adjustments that will be made if the comparison groups are nonequivalent, and the form of reporting that will be used. Another matter that should be considered at this time is the appropriate units of analysis. Whatever the nature of the study, the analysis plan should close the logical loop by showing how the study questions will be answered.

When Plans Are Being Made for Data Collection

Coordinate analysis plans with methods for selecting sources of information. The methods for selecting data sources strongly determine the kinds of analysis that can be applied to the resulting data. As noted in earlier chapters, evaluators can use descriptive statistics in many situations, but inferential techniques depend upon knowledge of sampling distributions, knowledge that can be applied only when the data have been produced by a random process.

Random processes can be invoked in many ways and with attendant variations in analytic methods. Often the choice of sampling procedure can affect the efficiency of the study as well. Evaluators should make a decision on the particular form of random selection in consultation with a sampling statistician in advance of data collection. Unless proper records of the sampling process are maintained, an analyst may not be able to use statistical inference techniques to estimate population parameters.

Coordinate analysis plans with data collection. As data collection methods are firmed up and instruments are developed, variables will be defined and measurement levels will be determined. This is the time to review the list of variables to ensure that all those necessary for the analysis are included in the data collection plans. The measurement level corresponding to a concept is often intrinsic to the concept, but if that is not so, it is usually wise to strive for the higher levels of measurement. There may be analytic advantage to

the higher levels or, if going to a higher level is more costly, the proper trade-off may be to settle for a somewhat weaker analysis method.¹

As the Data Analysis Begins

Check the data for errors and missing attributes. No matter how carefully evaluators have collected, recorded, and transformed the data to an analysis medium, there will be errors. They can detect and remove some by simple checks. Computer programs can be written, or may already exist, for checking the plausibility of attributes. For example, the gender variable has two attributes, male and female, and therefore two possible numerical values, say 0 and 1. Any other value is an error and can be readily detected. In a similar way, evaluators can check all variables to ensure that the attributes in the data base are reasonable.

They can detect other errors by contingency checks. Such checks are based on the fact that the attributes for some variables are conditional upon the attributes of other variables. For example, if a medical case has "male" as an attribute for gender, then it should not have "pregnant" as an attribute of physical condition. These kinds of if-then checks on the attributes are also relatively easy to automate.

A missing attribute, where none of the acceptable attributes for the variable is present, is more difficult to deal with. Evaluators have four options: (1) go

¹Flexibility usually exists on the fuzzy border between ordinal and interval variables. Analysts often treat an ordinal variable as if it were measured at the interval level. In fact, some authorities (see Kerlinger, 1986, pp. 401-3, for example) believe that most psychological and educational variables approximate interval equality fairly well. In any case, instrument construction should take account of the measurement level desired.

back to the data source and try to recover the missing attribute, (2) drop the case from all analyses, (3) drop the case from any analysis involving the variable in question but use the case for all other analyses, and (4) fill in a substitute value for the missing attribute. Considerations involved in dealing with missing attributes are treated by many writers. (See, for example, Groves, 1989; Little and Rubin, 1987; and Rubin, 1987.)

When evaluators have used probability sampling with the aim of estimating population parameters from sample results, overall nonresponse by units from the sample is an especially important problem. If the nonresponse rate is substantial and if it cannot be shown that the respondents and nonrespondents are probably similar on variables of interest, doubt is cast on the estimates of population parameters. Consequently, an analysis of nonrespondents will be needed. See Groves (1989) for an introduction to the literature on nonresponse issues.

Explore the data. A number of statistical methods have been specifically developed to help evaluators get a feel for the data and to produce statistics that are relatively insensitive to idiosyncracies in the data. Some of these, like the stem-and-leaf plot and the box-and-whiskers plot, are graphic and especially useful in understanding the nature of the data. (Details about exploratory data analysis may be found in Tukey, 1977; Hoaglin, Mosteller, and Tukey, 1983, 1985; Velleman and Hoaglin, 1981; and Hartwig and Dearing, 1979.)

Fit the analysis methods to the study question and the data in hand. The appropriateness of an analysis method depends upon a number of factors such as the way in which the data sources were selected, the measurement level of the variables involved, the distribution of the variables, the time order of the variables, and whether the intent is to generalize from

the cases for which data are available to a larger population. Some factors, like the measurement level, must be considered in every data analysis, while others, like time order, may be germane only for certain types of questions—in this instance, a question about a causal association.

When evaluators consider two or more different analysis methods, the choice may not be obvious. For example, with interval level measurement, the median may be preferable to the mean as a measure of central tendency if the distribution is very asymmetrical. But asymmetry is a matter of degree and a little error from asymmetry may be acceptable if there are strong advantages to using the mean. Or it may be easy to transform the variable so that near symmetry is attained. Statistical tests that indicate the degree of asymmetry are available, but ultimately the evaluators have to make a judgment.

“The data don’t remember where they came from.” These words of a prominent statistician underscore the point that the data analyst must be mindful of the process that generated the data. We can blindly apply a host of numerical procedures to a data set but many of them would probably not be appropriate in view of the process that produced the data. For example, the methods of statistical inference apply only to data generated by a random process or one that is “random in effect.” (For a discussion of the circumstances in which statistical inference is appropriate, see Mohr, 1990, pp. 67-74.) Since the data do not remember how they were produced, the analyst has to ensure that the techniques are not misapplied.

Monitor the intermediate results and make analytic adjustments as necessary. Even with good planning, it is not possible to foresee every eventuality. The data in hand may be different from what was planned, or preliminary analyses may suggest new questions to

explore. For example, the distribution of the data may take a form not anticipated so that analytic transformations are necessary. Or, a program may have an unanticipated effect that warrants a search for an explanation. The analyst must scrutinize the intermediate results carefully to spot opportunities for supplementary analyses as well as to avoid statistical procedures that are not compatible with the data.

**As the Results Are
Produced and
Interpreted**

Use graphics but avoid displays that distort the data. The results of quantitative data analysis may be terse to the point of obtuseness. Graphics may help both in understanding the results and in communicating them. There are many excellent examples of how to visually display quantitative information but even more of how to distort and obfuscate. (For introductions to graphic analysis and data presentation, see Cleveland, 1985; Du Toit, Steyn, and Stumpf, 1986; and Tufte, 1983.)

Be realistic and forthright about uncertainty. Uncertainty is inherent in real-world data. All measurements have some degree of error. If sampling is used, additional error is introduced. Data entry and data processing may produce yet more error. While evaluators can and should take steps to reduce error, subject to resource constraints, some error will always remain. The question that must be addressed is whether the level of error present threatens what are otherwise the conclusions from the study.

A complementary question is how to report the nature and extent of error. Reporting issues for some forms of quantitative analysis have been given considerable

Chapter 7
Avoiding Pitfalls

attention and several professional organizations offer guidelines.² The basic rule is to be forthright about the nature of the evidence.

²The Evaluation Research Society (now merged with Evaluation Network to become the American Evaluation Association) published standards that include coverage of reporting issues (Rossi, 1982). Other standards that give somewhat more attention to statistical issues are those of the American Association of Public Opinion Research (1991) and the Council of American Survey Research Organizations (1986). In 1988, the federal government solicited comments on a draft Office of Management and Budget circular establishing guidelines for federal statistical activities. A final version of the governmentwide guidelines, which included directions for the documentation and presentation of the results of statistical surveys and other studies, has not been published.

Bibliography

The statistics texts marked with an asterisk (*) provide more detailed information on the statistical topics mentioned in this paper.

American Association of Public Opinion Research. Code of Professional Ethics and Practices. Ann Arbor, Mich.: 1991.

Arney, W. R. Understanding Statistics in the Social Sciences. New York: W. H. Freeman, 1990.*

Bollen, K. A. Structural Equations With Latent Variables. New York: John Wiley, 1989.

Bornstedt, G. W., and D. Knoke. Statistics for Social Data Analysis. Itasca, Ill.: F. E. Peacock, 1982.*

Bryman, A., and D. Cramer. Quantitative Data Analysis for Social Scientists. New York: Routledge, 1990.*

Bunge, M. Causality and Modern Science, 3rd ed. rev. New York: Dover Publications, 1979.

Cleveland, W. S. The Elements of Graphing Data. Monterey, Calif.: Wadsworth, 1985.

Cohen, J., and P. Cohen. Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1975.*

Council of American Survey Research Organizations. Code of Standards. Port Jefferson, N.Y.: 1986.

Draper, N., and H. Smith. Applied Regression Analysis, 2nd ed. New York: John Wiley, 1981.*

Du Toit, S. H. C., A. G. W. Steyn, and R. H. Stumpf. Graphical Exploratory Data Analysis. New York: Springer-Verlag, 1986.

Bibliography

Freedman, D., R. Pisani, and R. Purves. Statistics. New York: W. W. Norton, 1980.*

Groves, R. M. Survey Errors and Survey Costs. New York: John Wiley, 1989.*

Hage, J., and B. F. Meeker. Social Causality. Boston: Unwin Hyman, 1988.

Hahn, G. J., and W. Q. Meeker. Statistical Intervals: A Guide for Practitioners. New York: John Wiley, 1991.*

Hartwig, F., and B. E. Dearing. Exploratory Data Analysis. Newbury Park, Calif.: Sage, 1979.

Hays, W. L. Statistics, 3rd ed. New York: Holt, Rinehart and Winston, 1981.*

Hildebrand, D. K., J. D. Laing, and H. Rosenthal. Analysis of Ordinal Data. Newbury Park, Calif.: 1977.*

Hoaglin, D. C., F. Mosteller, and J. W. Tukey (eds). Understanding Robust and Exploratory Data Analysis. New York: John Wiley, 1983.

Hoaglin, D. C., F. Mosteller, and J. W. Tukey (eds). Exploring Data Tables, Trends, and Shapes. New York: John Wiley, 1985.

Kerlinger, F. N. Foundations of Behavioral Research, 3rd ed. New York: Holt, Rinehart and Winston, 1986.

Little, R. J. A., and D. B. Rubin. Statistical Analysis with Missing Data. New York: John Wiley, 1987.

Loether, H. J., and D. G. McTavish. Descriptive and Inferential Statistics: An Introduction, 3rd ed. Boston: Allyn and Bacon, 1988.*

Bibliography

McPherson, G. Statistics in Scientific Investigation: Its Basis, Application, and Interpretation. New York: Springer-Verlag, 1990.*

Madansky, A. Prescriptions for Working Statisticians. New York: Springer-Verlag, 1988.*

Madow, W. G., I. Olkin, and D. B. Rubin. Incomplete Data in Sample Surveys, vols. 1-3. New York: Academic Press, 1983.

Mohr, L. B. Understanding Significance Testing. Newbury Park, Calif.: Sage, 1990.

Moore, D. S., and G. P. McCabe. Introduction to the Practice of Statistics. New York: W. H. Freeman, 1989.*

Mosteller, F., S. E. Fienberg, and R. E. K. Rourke. Beginning Statistics with Data Analysis. Reading, Mass.: Addison-Wesley, 1983.*

Pedhazur, E. J. Multiple Regression in Behavioral Research, 2nd ed. New York: Holt, Rinehart and Winston, 1982.*

Reynolds, H. T. Analysis of Nominal Data, 2nd ed. Newbury Park, Calif.: Sage, 1984.*

Rossi, P. H. (ed). Standards for Evaluation Practice. San Francisco: Jossey-Bass, 1982.

Rubin, D. B. Multiple Imputation for Nonresponse in Surveys. New York: John Wiley, 1987.

Siegel, S., and N. J. Castellan, Jr. Nonparametric Statistics for the Behavioral Sciences, 2nd ed. New York: McGraw-Hill, 1988.*

Tesch, R. Qualitative Research: Analysis Types and Software Tools. New York: Falmer Press, 1990.

Bibliography

Tufte, E. R. The Visual Display of Quantitative Information. Cheshire, Conn.: Graphics Press, 1983.

Tukey, J. W. Exploratory Data Analysis. Reading, Mass.: Addison-Wesley, 1977.

U.S. General Accounting Office. Potential Effects of a National Mandatory Deposit on Beverage Containers, GAO/PAD-78-19. Washington, D.C.: December 1977.

U.S. General Accounting Office. States' Experience With Beverage Container Deposit Laws Shows Positive Benefits, GAO/PAD-81-8. Washington, D.C.: December 1980.

U.S. General Accounting Office. An Evaluation of the 1981 AFDC Changes: Final Report, GAO/PEMD-85-4. Washington, D.C.: July 1985.

U.S. General Accounting Office. Medical Devices: Early Warning of Problems Is Hampered by Severe Underreporting, GAO/PEMD-87-1. Washington, D.C.: December 1986.

U.S. General Accounting Office. Cancer Patient Survival: What Progress Has Been Made? GAO/PEMD-87-13. Washington, D.C.: March 1987a.

U.S. General Accounting Office. Noncash Benefits: Methodological Review of Experimental Valuation Methods Indicates Many Problems Remain, GAO/PEMD-87-23. Washington, D.C.: September 1987b.

U.S. General Accounting Office. The H-2A Program: Protections for U.S. Farmworkers, GAO/PEMD-89-3. Washington, D.C.: October 1988.

U.S. General Accounting Office. Children and Youths: About 68,000 Homeless and 186,000 in Shared

Bibliography

Housing at Any Given Time, GAO/PEMD-89-14.
Washington, D.C.: June 1989a.

U.S. General Accounting Office. Criminal Justice: Impact of Bail Reform in Selected District Courts, GAO/GGD-90-7. Washington, D.C.: November 1989b.

U.S. General Accounting Office. Alternative Agriculture: Federal Incentives and Farmers' Opinions, GAO/PEMD-90-12. Washington, D.C.: February 1990a.

U.S. General Accounting Office. Food Stamp Program: A Demographic Analysis of Participation and Nonparticipation, GAO/PEMD-90-8. Washington, D.C.: January 1990b.

U.S. General Accounting Office. Promising Practice: Private Programs Guaranteeing Student Aid for Higher Education, GAO/PEMD-90-16. Washington, D.C.: June 1990c.

U.S. General Accounting Office. Voting: Some Procedural Changes and Informational Activities Could Increase Turnout, GAO/PEMD-91-1. Washington, D.C.: November 1990d.

Velleman, P. F., and D. C. Hoaglin. Applications, Basics, and Computing of Exploratory Data Analysis. Boston: Duxbury Press, 1981.

Wallis, W. A., and H. V. Roberts. Statistics: A New Approach. Glencoe, Ill.: Free Press, 1956.*

Yin, R. K. Case Study Research: Design and Methods, rev. ed. Newbury Park, Calif.: Sage, 1989.

Glossary

Analysis of Covariance	A method for analyzing the differences in the means of two or more groups of cases while taking account of variation in one or more interval-ratio variables.
Analysis of Variance	A method for analyzing the differences in the means of two or more groups of cases.
Association	General term for the relationship among variables.
Asymmetric Measure of Association	A measure of association that makes a distinction between independent and dependent variables.
Attribute	A characteristic that describes a person, thing, or event. For example, being female is an attribute of a person.
Batch	A group of cases for which no assumptions are made about how the cases were selected. A batch may be a population, a probability sample, or a nonprobability sample, but the data are analyzed as if the origin of the data is not known.
Bell-Shaped Curve	A distribution with roughly the shape of a bell; often used in reference to the normal distribution but others, such as the t distribution, are also bell-shaped.
Bivariate Data	Information about two variables.

Glossary

Box-and-Whisker Plot	A graphic way of depicting the shape of a distribution.
Case	A single person, thing, or event for which attributes have been or will be observed.
Causal Analysis	A method for analyzing the possible causal associations among a set of variables.
Causal Association	A relationship between two variables in which a change in one brings about a change in the other.
Central Tendency	General term for the midpoint or typical value of a distribution.
Conditional Distribution	The distribution of one or more variables given that one or more other variables have specified values.
Confidence Interval	An estimate of a population parameter that consists of a range of values bounded by statistics called upper and lower confidence limits.
Confidence Level	A number, stated as a percentage, that expresses the degree of certainty associated with an interval estimate of a population parameter.
Confidence Limits	Two statistics that form the upper and lower bounds of a confidence interval.

Glossary

Continuous Variable	A quantitative variable with an infinite number of attributes.
Correlation	(1) A synonym for association. (2) One of several measures of association (see <u>Pearson Product-Moment Correlation Coefficient</u> and <u>Point Biserial Correlation</u>).
Data	Groups of observations; they may be quantitative or qualitative.
Dependent Variable	A variable that may, it is believed, be predicted by or caused by one or more other variables called independent variables.
Descriptive Statistic	A statistic used to describe a set of cases upon which observations were made. Compare with <u>Inferential Statistic</u> .
Discrete Variable	A quantitative variable with a finite number of attributes.
Dispersion	See <u>Spread</u> .
Distribution of a Variable	Variation of characteristics across cases.
Experimental Data	Data produced by an experimental or quasi-experimental design.

Glossary

Frequency Distribution	A distribution of the count of cases corresponding to the attributes of an observed variable.
Gamma	A measure of association; a statistic used with ordinal variables.
Histogram	A graphic depiction of the distribution of a variable.
Independent Variable	A variable that may, it is believed, predict or cause fluctuation in a dependent variable.
Index of Dispersion	A measure of spread; a statistic used especially with nominal variables.
Inferential Statistic	A statistic used to describe a population using information from observations on only a probability sample of cases from the population. Compare with <u>Descriptive Statistic</u> .
Interquartile Range	A measure of spread; a statistic used with ordinal, interval, and ratio variables.
Interval Estimate	General term for an estimate of a population parameter that is a range of numerical values.
Interval Variable	A quantitative variable the attributes of which are ordered and for which the numerical differences between adjacent attributes are interpreted as equal.

Glossary

Lambda	A measure of association; a statistic used with nominal variables.
Level of Measurement	A classification of quantitative variables based upon the relationship among the attributes that compose a variable.
Marginal Distribution	The distribution of a single variable based upon an underlying distribution of two or more variables.
Mean	A measure of central tendency; a statistic used primarily with interval-ratio variables following symmetrical distributions.
Measure	In the context of data analysis, a statistic, as in the expression "a measure of central tendency."
Median	A measure of central tendency; a statistic used primarily with ordinal variables and asymmetrically distributed interval-ratio variables.
Mode	A measure of central tendency; a statistic used primarily with nominal variables.
Nominal Variable	A quantitative variable the attributes of which have no inherent order.
Nonexperimental Data	Data not produced by an experiment or quasi-experiment; for example, the data may be administrative records or the results of a sample survey.

Glossary

Nonprobability Sample	A sample not produced by a random process; for example, it may be a sample based upon an evaluator's judgment about which cases to select.
Normal Distribution (Curve)	A theoretical distribution that is closely approximated by many actual distributions of variables.
Observation	The words or numbers that represent an attribute for a particular case.
Ordinal Variable	A quantitative variable the attributes of which are ordered but for which the numerical differences between adjacent attributes are not necessarily interpreted as equal.
Outlier	An extremely large or small observation; applies to ordinal, interval, and ratio variables.
Parameter	A number that describes a population.
Pearson Product-Moment Correlation Coefficient	A measure of association; a statistic used with interval-ratio variables.
Point Biserial Correlation	A measure of association between an interval-ratio variable and a nominal variable with two attributes.

Glossary

Point Estimate	An estimate of a population parameter that is a single numerical value.
Population	A set of persons, things, or events about which there are questions.
Probability Distribution	A distribution of a variable that expresses the probability that particular attributes or ranges of attributes will be, or have been, observed.
Probability Sample	A group of cases selected from a population by a random process. Every member of the population has a known, nonzero probability of being selected.
Qualitative Data	Data in the form of words.
Quantitative Data	Data in the form of numbers. Includes four levels of measurement: nominal, ordinal, interval, and ratio.
Random Process	A procedure for drawing a sample from a population or for assigning cases to experimental and control conditions such that no purposeful forces influence the selection of cases and that the laws of probability therefore describe the process.
Range	A measure of spread; a statistic used primarily with interval-ratio variables.

Glossary

Ratio Variable	A quantitative variable the attributes of which are ordered, spaced equally, and with a true zero point.
Regression Analysis	A method for determining the association between a dependent variable and one or more independent variables.
Regression Coefficient	An asymmetric measure of association; a statistic computed as part of a regression analysis.
Resistant Statistic	A statistic that is not much influenced by changes in a few observations.
Response Variable	A variable on which information is collected and in which there is an interest because of its direct policy relevance. For example, in studying policies for retraining displaced workers, employment rate might be the response variable. See <u>Supplementary Variable</u> .
Sample Design	The sampling procedure used to produce any type of sample.
Sampling Distribution	The distribution of a statistic.
Scientific Sample	Synonymous with <u>Probability Sample</u> .
Simple Random Sample	A probability sample in which each member of the population has an equal chance of being drawn to the sample.

Glossary

Spread	General term for the extent of variation among cases.
Standard Deviation	A measure of spread; a statistic used with interval-ratio variables.
Statistic	A number computed from data on one or more variables.
Statistical Sample	Synonymous with <u>Probability Sample</u> .
Stem-and-Leaf Plot	A graphic or numerical display of the distribution of a variable.
Structural Equation Modeling	A method for determining the extent to which data on a set of variables are consistent with hypotheses about causal associations among the variables.
Supplementary Variable	A variable upon which information is collected because of its potential relationship to a response variable.
Symmetric Measure of Association	A measure of association that does not make a distinction between independent and dependent variables.
Transformed Variable	A variable for which the attribute values have been systematically changed for the sake of data analysis.

Glossary

Treatment Variable	In program evaluation, an independent variable of particular interest because it corresponds to a program or a policy instituted with the intent of changing some dependent variable.
Unit of Analysis	The person, thing, or event under study.
Variable	A logical collection of attributes. For example, each possible age of a person is an attribute and the collection of all such attributes is the variable age.

Contributors

Carl Wisler
Lois-ellin Datta
George Silberman
Penny Pickett

Papers in This Series

This is a flexible series continually being added to and updated. The interested reader should inquire about the possibility of additional papers in the series.

The Evaluation Synthesis. Transfer paper 10.1.2.

Content Analysis: A Methodology for Structuring and Analyzing Written Material. Transfer paper 10.1.3, formerly methodology transfer paper 3.

Designing Evaluations. Transfer paper 10.1.4.

Using Structured Interviewing Techniques. Transfer paper 10.1.5.

Using Statistical Sampling. Transfer paper 10.1.6.

Developing and Using Questionnaires. Transfer paper 10.1.7, formerly methodology transfer paper 7.

Case Study Evaluations. Transfer paper 10.1.9.

Prospective Evaluation Methods: The Prospective Evaluation Synthesis. Transfer paper 10.1.10.

Quantitative Data Analysis: An Introduction. Transfer paper 10.1.11.



**United States
General Accounting Office
Washington D.C. 20548**

**Official Business
Penalty for Private Use \$300**

**First Class Mail
Postage & Fees Paid
GAO
Permit No. G100**