BIBLIOGRAPHIC DATA SHEET

1. CONTROL NUMBER
PN-AAJ-250

2. SUBJECT CLASSIFICATION (695)
JE30-0000-0000

3. TITLE AND SUBTITLE (240)
Planning education for development, vol. 4: Data analysis for educational planning

4. PERSONAL AUTHORS (100)
Lewis, Gary

5. CORPORATE AUTHORS (101)
Harvard Univ. Ctr. for Studies in Education and Development

6. DOCUMENT DATE (110)
1980

7. NUMBER OF PAGES (120)
234p.

8. ARC NUMBER (170)
370.H339

9. REFERENCE ORGANIZATION (130)
Harvard

10. SUPPLEMENTARY NOTES (500)
(vol. 1, 471p.: PN-AAJ-247; vol. 2, 432p.: PN-AAJ-248; vol. 3, 528p.: PN-AAJ-249)

11. ABSTRACT (950)

12. DESCRIPTORS (920)
Educational planning
Education for development
Data processing
Methodology
Analysis
Educational research
Regression analysis

13. PROJECT NUMBER (150)
931008900

14. CONTRACT NO.(140)
AID/ta-C-1336

15. CONTRACT TYPE (140)

16. TYPE OF DOCUMENT (160)

AID 590-7 (10-79)

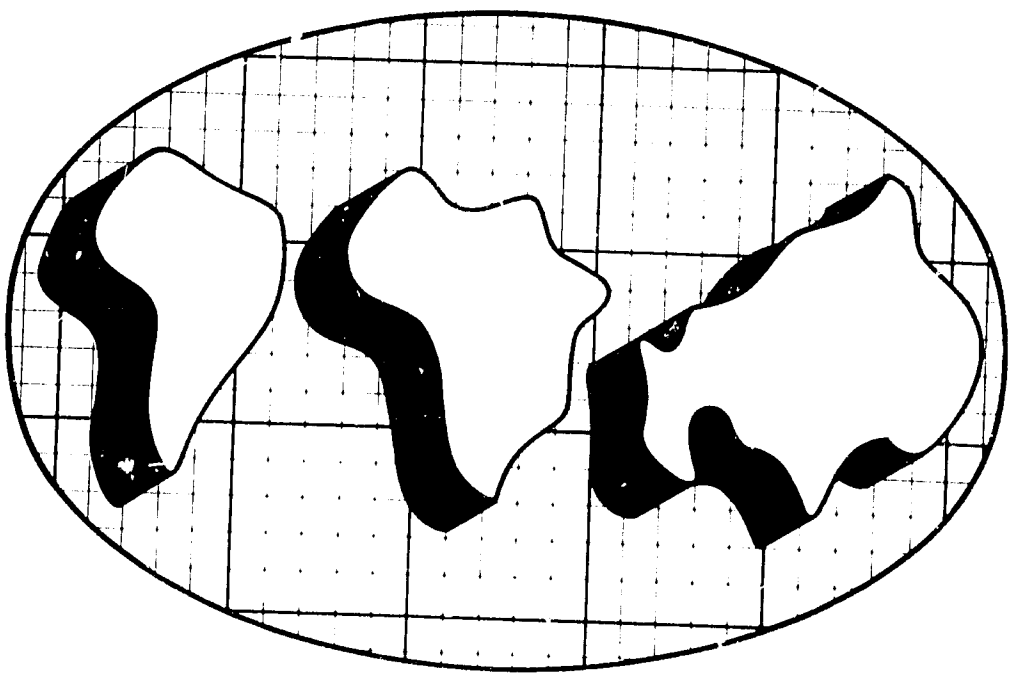
370
H 339
v.4

PN-AAJ-250

Planning Education for Development:

Volume IV
Data Analysis for Educational Planning

GARY LEWIS



Center for Studies in Education and Development
Harvard University, Cambridge, Massachusetts

Planning Education For Development

Volumes I-IV

Volume IV

Data Analysis for Educational Planning

Gary Lewis
with
contributions
by

Noel McGinn
Ernesto Schiefelbein

AID/ta-C-1336
931008900

Center for Studies in Education and Development
Harvard Graduate School of Education

Cambridge, Massachusetts
1980

To present and former students at the
Center for Studies in Education and Development

Preface

The term "data analysis for educational planning" may have different meanings for different people. It will be useful at the outset to elaborate some of the assumptions that the authors of this text have used with regard to the term. In what contexts do we see data analysis and the planning of education?

Data analysis is viewed as but one part in a much larger research process. Perhaps "research" is too grandiose. It is really just a question-asking and answering process which starts with a clearly defined question of interest. Next, information is collected in response to this question. What kind of data are to be collected and how they shall be collected are typical concerns here. These matters are discussed in Volume III. The collected information must then be stored for easy retrieval. Depending upon the circumstances, this may mean nothing more than a few loose-leaf pages stored in a folder, or it may mean the creation and management of computer data bases. During the fourth stage in this simplified research process, analysis of the data takes place. It is here that the collected data are subjected to a series of techniques and procedures designed to address the questions of concern. In the fifth and final stage, results of the analysis are interpreted and recommendations made. Only the fourth, analysis stage in the research process is of concern in this volume.

In this discussion we more narrowly define data analysis as the set of techniques and procedures used to increase the amount of information from quantitative data. This is not to say that qualitative data are not important or are not used by educational planners. It only means that the analysis of such data is sufficiently different from the analysis of quantitative data that it would warrant its own extended discussion.

There is also an organizational context implicit in the materials selected for inclusion here. The primary audience is assumed to be people concerned with the planning of education. This includes program, institutional, and system level planning. But the common core assumptions are that the planner works in an organizational setting where at least some of the following features are important aspects of the work:

- decision-makers and managers need quickly supplied analyses, even if limited in scope, rather than more elaborate studies that take a longer time to prepare;
- decision-makers and managers are concerned with specific problems and are not concerned about generalizing to all possible situations;
- the planner or analyst must think in terms of the operating system to which the data refer, and the policy instruments available to the decision-maker, and
- simplicity of analysis is important for ease of interpretation and transmission of results, but also because it permits increased local control over the direction and conduct of the analysis.

These assumptions have affected the selection of materials for this volume, of which there are three basic types. First there is a series of chapters which

are instructional in nature. These are designed to teach basic and practical techniques and procedures for the analysis of quantitative educational data. The vast majority of the techniques described in the text can be performed with paper and pencil, although an inexpensive hand calculator removes much of the drudgery from the work. The level of the material is introductory, assuming no prior knowledge of the subject.

The rudiments of each technique are first presented, and are then followed by a series of refinements. Whether one uses the simplest form of a technique or a more refined form depends partly on the situation. Scratching the surface, or looking for only the most obvious points of an analysis, may require only the most elementary forms of analysis. Going beneath the surface often requires a little more work, and many of the refinements in techniques allow the analyst to probe research questions in more depth than is possible with the simpler forms of a technique. What level of sophistication in technique to choose is situationally dependent, and should be based on decisions as to purpose of the analysis, time available, and other resource constraints.

We are not offering here just another standard statistical treatise, even though there is considerable overlap between what is presented and similar material found in introductory and intermediate statistics texts. The reader, for example, will find little or no discussion about probability theory, differing distributional forms, hypothesis testing and other similar items found in statistics textbooks. The materials in this book respond to two basic questions in educational planning: What can be done to increase the amount of information available from educational data, and what techniques can be used in given situations?

Chapter 1 introduces a schema of analysis techniques and discusses which are appropriate for use with different kinds of data. Simple visual displays of data, in the form of stem-and-leaf diagrams, histograms, bar graphs, box plots, statistical maps, and graphs are described in Chapter 2. Chapters 3 and 4 discuss common forms of tabled data and how to obtain more information from the data than mere inspection permits. Chapter 5 is devoted to regression analysis, one of the more powerful techniques for exploring relationships between variables. Examples are liberally used in all chapters. In sum, the techniques in these chapters will provide the analyst with a basis for dealing with the major types of quantitative data occurring in educational planning situations.

There is also a series of exercises and problem sets which permit the reader to practice the analysis techniques described in the text. The problem sets appearing at the end of each chapter use data taken from an actual educational planning situation which has been disguised for inclusion here. The situation is simple and widely relevant: migration and demographic factors have caused a region of a country to experience fluctuating enrollments. The educational authorities are concerned with the impact that changing enrollments will have upon the cost and quality of education. The data set used to address some of these concerns are introduced in Chapter 1 and then used in the practice exercises at the end of each succeeding chapter. In addition,

there are several exercises included in the chapters that can be used to provide additional practice in the techniques described.

The third and final type of material included in this volume is found in Chapter 6. This chapter is a case study of how even simple analysis of educational data can be used to support policy. The setting of the case is the Ministry of Education planning unit in Paraguay. Of particular interest in the analysis is the usefulness of multiple-grade classes and double shifts in that country.

In total, it is hoped that the instructional materials, practice exercises, and case study will provide complementary introductions into the use of data analysis for educational planning purposes.

Acknowledgements

In addition to those individuals mentioned in the general acknowledgements to the entire series of volumes, we would especially like to add the name of Lascelles Anderson, who wrote an initial draft of the material on regression analysis.

VOLUME IV

Table of Contents

Preface

Acknowledgements

Chapter 1: Categories of Analysis Techniques

1.0 What's Ahead? 1
2.0 Techniques for Data Analysis 1
3.0 Basic Terms 2
3.1 Universe of Interest 3
3.2 Data Collection 3
3.3 Single Variable Descriptive Analysis 4
3.4 Data Base Classification 7
3.5 Determination of Techniques 8
4.0 Summary 10
5.0 Terms for Review 10
6.0 Practice Exercises 11

Chapter 2: Visual Data Displays

1.0 What's Ahead? 19
2.0 Displays for Interval Data 19
2.1 The Stem-and-Leaf Diagram 19
2.1.1 Measures of Central Tendency 24
2.1.2 Measures of Spread 25
2.1.3 The Shape of the Distribution 27
2.2 Data Transformations 28
2.3 Comparisons Among Groups 35
3.0 Displays for Ordinal Data 41
3.1 Histograms with Interval Data 41
3.2 Cumulative Frequency Graphs 46
3.3 Histograms with Ordinal Data 48
4.0 Displays for Nominal Data 48
5.0 Displays of Data Categorized in Other Manners 51
5.1 Statistical Maps for Use with Geographical Categories 51
5.2 Displays of Time Series Data 54
6.0 Summary 59
7.0 Terms for Review 60
8.0 Practice Exercises 60

Chapter 3: Tables of Counts

1.0 What's Ahead? 65
2.0 Independence and Association 66
  2.1 Independence 66
  2.2 Association 68
  2.3 The Chi-square Measure of Independence 72
3.0 Measures of Association 80
  3.1 Cramer's V 80
  3.2 Yule's Q 80
  3.3 Gamma 83
  3.4 Somer's D 91
4.0 The Nature of an Association 92
  4.1 Percentaged Tables 92
  4.2 Prediction Analysis 94
5.0 Summary 98
6.0 Terms for Review 98
7.0 Practice Exercises 99

Chapter 4: Tables of Central Tendency

1.0 What's Ahead? 103
2.0 The Additive Fit 104
  2.1 The General Idea 104
  2.2 The Mechanics of an Additive Fit 107
  2.3 Polishing an Additive Fit 110
  2.4 Goodness of Fit 111
3.0 The Additive Fit Extended 115
  3.1 Adding an Interaction Term 116
  3.2 Data Transformations 120
4.0 Summary 127
5.0 Terms for Review 127
6.0 Practice Exercises 127

Chapter 5: Regression Analysis

1.0 What's Ahead? 131
2.0 The Basic Idea of Regression Analysis 131
3.0 Fitting Procedures 135
  3.1 The Inspection Fit 138
  3.2 The Median Fit 140
  3.3 The Least Squares Fit 140
4.0 Examination of Residuals 142
  4.1 A Plot of Residuals 143
5.0 Measure of Association 146
6.0 Data Transformations 148
  6.1 Selection of the Appropriate Transformation 152

7.0 Multiple Regression 159
7.1 The Fitting Procedure 161
7.2 Examining Residuals 164
7.3 The Interpretation of Multiple Regression Coefficients 168
8.0 Summary 171
9.0 Terms for Review 171
10.0 Practice Exercise 172

Chapter 6: How Educational Statistics Can Be Utilized for Policy Analysis: A Case Study

1.0 Objectives of this Unit 177
2.0 Can a Planning Office be Useful to a Ministry of Education? 178
2.1 The Educational Planning Office in Paraguay 179
2.1.1 Problems Facing Education in Paraguay 179
2.1.2 Planning Office Response 181
3.0 Why Ministry Statistical Systems Don't Meet the Planner's Needs 182
3.1 Data Collection in Paraguay 184
4.0 How to Choose Variables for Analysis in Statistics 190
4.1 Variables Included in the Analysis in Paraguay 193
5.0 Sampling the Data Source 196
5.1 Sampling the Statistical Data Base in Paraguay 198
5.1.1 Design of the Sample 198
5.1.2 Characteristics of the Obtained Sample 198
5.1.3 Tabulation Procedure 200
6.0 Analysis Procedures 203
6.1 Are Multiple Grade Classes and Double Shifts Useful in Paraguay? 204
7.0 Conclusions 208

Appendix A: Selecting a Simple Random Sample

1.0 Why Sample and What to Look for in a Sample 211
2.0 Tables of Random Numbers 212
3.0 Sample Size 215
4.0 Other Sampling Procedures 216

Coda on the Passing of an Institution
USAID Harvard Planning Project Papers

CHAPTER 1

Categories of Analysis Techniques

Gary Lewis

1.0 *What's Ahead?*

In this introductory chapter, the main purpose is to outline in brief fashion what the remainder of the text will include. To do this, basic terminology that appears frequently must first be defined. A schema of analysis techniques is introduced as a vehicle for defining basic terms and for identifying those techniques discussed in later chapters. The schema can also serve as a guide to the proper selection of a technique, given specific kinds of data and the general purpose of the analysis. Finally, the reader will be introduced to a data set actually used in an educational planning situation. Analysis of this data was used in an effort to assess the impact of fluctuating enrollments on the cost of education within a region of an (unidentified) country. The data set will be used throughout this text as practice problems appearing at the end of each chapter.

2.0 *Techniques for Data Analysis*

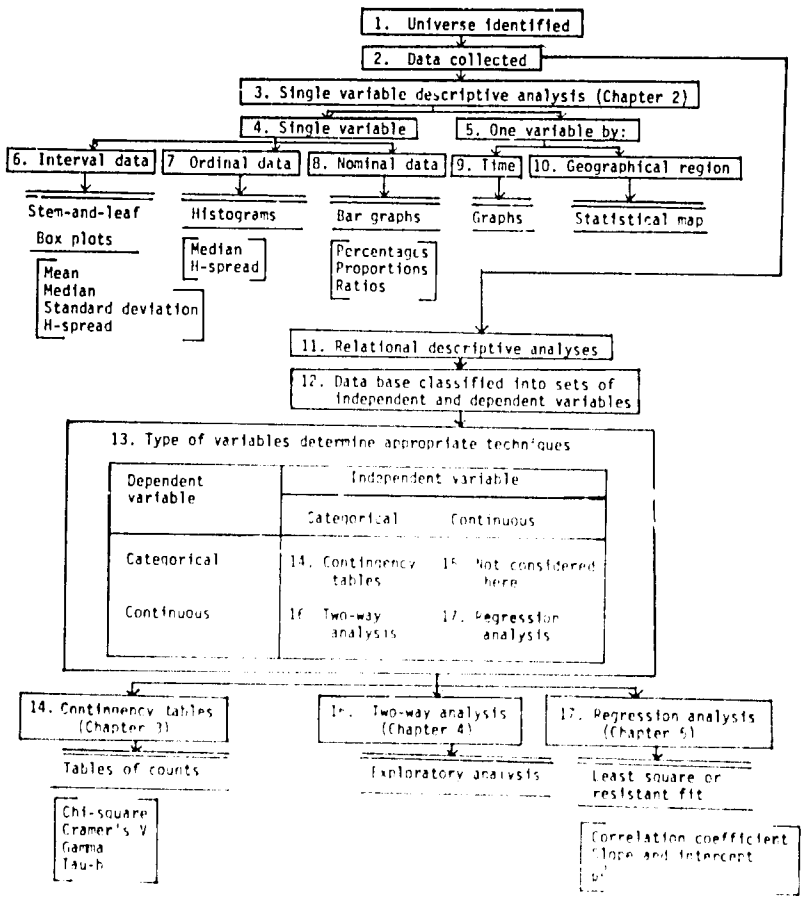
Shown in Figure 1.1 is a schema of procedures for data analysis often found useful by educational planners. While there are additions that could be made to the schema, the identified techniques will permit the planner to address general questions of concern using the kinds of data most commonly available.

Before going through the schema in detail, a prior word concerning the structure of the schema is in order. The items in boxes refer to general steps in the data analysis process. They also identify certain kinds of data with which the analyst often deals. There is a rough order to the numbered boxes in the sense that steps near the top of the figure normally occur prior to those near the bottom. But within a row of numbers (e.g., boxes 6 through 10 appear at the same level in Figure 1.1) there is no significance to the number sequence, which is used merely for identification purposes.

At the end of a tracing of boxes and arrows, the reader will find a double horizontal line. The items below these lines are specific techniques to be used in situations described by the tracing of boxes and arrows to that point. For example, find *stem-and-leaf displays* in Figure 1.1, located below Box #6 (*Interval data*). It will become clearer momentarily what is meant by all the terms, but for now it is sufficient to indicate that the structure of the schema says that stem-and-leaf displays are suitable for use with interval data (#6) on a single variable (#4) for descriptive analysis (#3) of the data collected (#'s 1-2). Beneath the listing of analysis procedures appropriate for a given

Figure 1.1:

Schema of Techniques for Data Analysis



situation is a listing of calculations which can be used to summarize features of the analysis. These are separated from the list of analysis procedures by a single horizontal line. For example, means, medians, standard deviations, and H-spreads could all be used to describe features of stem-and-leaf displays.

3.0 *Basic Terms*

To fully understand the use of Figure 1.1, it is necessary to introduce some basic terms which will appear throughout the remaining text. Where the terms are not assumed to be familiar, definitions are provided, and these have been underlined.

3.1. *Universe of interest* (Box 1 in Figure 1.1)

The *universe of interest* is the entire set of units (people, schools, classes, geographic regions, etc.) that the analyst wishes to study. If we wish to learn something about educational expenditures in countries of the world, then the universe is the entire set of countries. If we wish to learn something about student achievement in grade 6 in school Z, then the universe of interest is the entire set of students in grade 6 in that school. If we wish to learn something about school facilities in a given geographic region of a country, then the universe of interest is the entire set of school facilities in that region.

The important feature to note is that a universe is determined by the subject of study. Sometimes it may include many units; at other times only a few. What may be a universe of interest in one study may be only a subset of interest in another. In other words, a description of the universe is situationally dependent, to be defined by the topic of interest.

The universe is also commonly referred to as the *population* of interest. Both terms are used interchangeably here.

3.2. *Data collection* (Box 2 in Figure 1.1)

After having defined the universe of interest, it is necessary to collect relevant information. If the universe of interest is small, it is usually possible to collect information on all the units. For example, if a supervisor is interested in total enrollments in the 14 schools under his or her direction, then it is both possible and desirable to collect this information about each of the schools. In this case the entire universe is used to complete the remaining analysis.

However, if the universe is large, collection of information on all units may require an excessive expenditure of resources, in terms of time, people, or money. In such instances it is useful to choose from the universe a smaller sample of units. Data are then collected for all the units in the sample, and the analysis is done using only these units. For example, suppose we are interested in student achievement in primary grade 6 for an entire country. It would be a large undertaking to administer achievement tests to all the relevant students, unless it were a small country indeed. In such cases a much smaller sample would be drawn from the entire universe, and the test would be administered to this smaller group of individuals. We would then be interested in knowing if the information obtained from the analysis of the sample data also applies to the larger groups of students in grade 6 in the entire country. In other words, *can we believe* the results from the sample if we wish to make decisions that will affect the universe from which the sample is drawn?

There are additional terms that appear frequently when discussing samples and universes. Typically, one refers to population, or universe, *parameters*, and to sample *statistics*. For example, the average test score among all primary 6 students (our universe) is a *population parameter*, a description of some feature concerned with all the individuals or units in the universe. On the other hand the average test score computed for a sample of primary 6 students is called a *statistic*. If the sample is representative of the entire

universe, then our sample statistic and the actual population parameter will be fairly similar. Typically, we do not know what the population parameters are, although it is often possible to determine with some degree of confidence that our statistics, based on sample data, reflect the population.

There are various methods that can be used for selecting a sample from a universe of units. These will not be detailed in this text, since they more properly belong in a unit on research methods. However, one method of sample selection is of central importance and should be mentioned. This method, called *simple random sampling*, is important because it underlies so many efforts on the part of statisticians to answer the "Can I believe it?" question. For other methods of sampling, it is often difficult to answer this question. For this reason, and because it is easy to learn and apply, an introduction to simple random sampling is provided in the Appendix.

3.3. *Single-Variable Descriptive Analysis* (Boxes 3 through 10 in Figure 1.1)

A central term in data analysis is the concept of a *variable*. Suppose a supervisor of several schools wishes to learn something about teacher salaries in the schools in his or her jurisdiction. Data on the salaries of teachers would be collected, either through a sample if that is necessary, or from the entire universe if that is possible and easily obtained. An item for which information is collected is called a *variable*. Technically, a variable is a quantity that may take on one of many possible values; that is, a variable has varying values. Teachers' salaries is the variable of interest in the above example, and, of course, salaries do vary from teacher to teacher. If one is concerned with student achievement scores, then the variable of interest is achievement scores, obviously varying from student to student. The age, sex, height, race, and educational attainment level of a group of individuals are all variables that might be of interest.

When data have been collected on a single variable, the first analysis is normally used to describe important features of the variable. Typically, these analyses address the decision-maker's need to know answers to the "How much?" questions. We might, for example, wish to know what the minimum salary was in a sample of school administrators; or what the range between minimum and maximum salaries was; or perhaps something about the salary most commonly earned by the individuals in the group.

While there are several different techniques that can be employed for such single-variable analyses, their legitimate use depends upon the kind of data. More precisely, it depends upon the *level of measurement* involved in the data. Researchers have traditionally distinguished four different levels of measurement, these being *nominal*, *ordinal*, *interval*, and *ratio*. The last two levels are similar enough so that in practical usage they can be considered identical.[1] For our purposes here, we will be concerned only with nominal, ordinal, and interval scales.

Perhaps the most fundamental level of measurement is simple classification into categories. It is possible, for example, to classify people by their sex,

by their race, by their religion, by the color of their eyes, and by a host of other variables. A *nominal* scale of measurement occurs when we classify units (people, schools, classrooms, etc.) into these kinds of categories, *with no ordered relationship between the categories*. For example, students can be classified as going to School X, School Y, or School Z, without any implication of order between the schools — that, for example, School X is better or higher on some underlying scale than the others. Such classification is an example of a nominal scale.

When we classify units into categories *and there is an order* that can be perceived among the categories, then the scale of measurement is said to be *ordinal*. We might, for example, classify a sample of individuals as having attended either primary, secondary, or tertiary schools. In this case the three categories do have an underlying order to them: primary school must usually be completed before secondary school, which in turn is completed before tertiary school. The scale has the dimension of length of schooling: those in the primary classification have attended school for less time than those in secondary or tertiary categories, and those in the secondary classification less than the individuals with tertiary education. It should be pointed out that an ordinal scale does not permit one to discuss the *distance* between points of the scale. We can say that the tertiary education classification is *greater* than the secondary or primary one, but we cannot say how much greater.

When there is both an ordered nature to the data, *and* it is possible to specify the distance between any two points on the scale, then the level of measurement is said to be *interval*. An example of such an interval scaled variable would be the age of individuals. This meets both of the criteria listed above. First, points on the scale are ordered: a person whose age is 18 is *higher* on the scale than someone whose age is 8. The scale, of course, measures the length of time the individual has been alive. And, secondly, the distance between any of the points on the scale can be specified: someone 18 years of age is 10 years older than someone who is 8. The two conditions of ordered data and distance specification are both required for a scale to be interval.

Note that the conditions required for a nominal scale are minimal, while those for an ordinal scale are somewhat more restricted, and those for an interval scale the most restrictive considered here. We often speak of the *strength* of the level of measurement, with nominal the weakest (least restrictive), interval the strongest (most restrictive) and the ordinal scale intermediate in strength.

There are some additional terms closely associated with the level of measurement which appear commonly in analytical studies. A variable is said to be *discrete* or *categorical* when the level of measurement used is nominal or ordinal, that is, based upon classification into categories. A variable is said to be *continuous* when the level of measurement is based upon an interval level of measurement.[2]

Having distinguished between levels of measurement and kinds of variables, it is possible to finish discussing the variety of analytical tech-

niques possible for descriptive analyses of single variables. In Figure 1.1 the reader will see that if an analyst is concerned with only one variable, the kind of technique to choose is determined by the level of measurement involved. If the variable is measured with a nominal scale, then bar charts can be used, with summaries like percentages, proportions, and ratios used to describe features of the data. If an ordinal scaled variable is to be described, histograms are of use, with medians and H-spreads for further quantitative description. If the data are interval, stem-and-leaf displays can be used along with means and standard deviations to summarize the analysis. It is not expected at this point that the reader will be familiar with any of these techniques nor the summary calculations to be used with them. That is the purpose of the remaining chapters. The important point is the fact that different techniques are applicable to different kinds of variables or levels of measurement. This also means that techniques can be misapplied by using them in situations where the conditions for their use have not been met.

It is worth pointing out that techniques and statistics that apply to weaker levels of measurement can always be used with stronger levels of measurement. For example, percentages or proportions which are useful in describing nominal data can also be used for interval level data. On the other hand, it is *not* appropriate to use means or standard deviations with nominal data. These statistics can be used when the conditions necessary for interval level measurement have been met; and these conditions are not met with nominal or ordinal data.

As we shall see, however, the full utility of a technique is gained when it is used with the type of data for which it was intended and not for data at stronger levels of measurement. To use percentages for interval level data, for example, would require that somewhat arbitrary classification categories be established, with a resulting loss in information. We could classify incomes of individuals, for example, into low, medium, and high, and then calculate the percentage of people in each category. But where the cut-off points between low, medium, and high incomes should occur is a problem. We would also not be able to determine from this analysis alone what the minimum or maximum incomes were, or what incomes most commonly occurred. So there is a price to be paid when techniques appropriate for weaker levels of measurement are used with data available at a stronger level. The price paid is usually a loss of information and an introduction of increased ambiguity.

Techniques appropriate for use with interval, ordinal, and nominal variables are discussed in Chapter 2. Also presented in that chapter are two closely related techniques, each for use with a single variable, but in instances where the variable is available at several points in time or is available for different geographical regions. If, for example, we collect data on average teacher salaries over the past several years, a graphical display may be very effective in presenting a visual impression of any trends. Such displays are also effective when comparing trends for the same variable but for different groups. Similarly, it may be that we have data available on several different geographical regions. For example, we may have informa-

tion on average teacher salaries for each of several school districts within a region of a country. In such circumstances, statistical mapping techniques can be used to show the variable of interest across the geographical regions displayed. Both graphical and statistical mapping procedures are also discussed in Chapter 2. The emphasis in that chapter is on the "How much?," "Is it different from?," and "Can I believe it?" kind of questions that confront educational decision-makers and must be addressed by planners in their analyses.

3.4. *Data base classification* (Boxes 11 and 12 in Figure 1.1)

It is often the case that planners are concerned with questions of relationships between variables. How, for example, do educational expenditures rise as the number of students increase? How are program enrollments related to the sex of the student enrolled; i.e., are females apt to be in one kind of program while males are in others? How does social class affect the opportunity for attending school? These and countless other questions are of potential interest. They all involve assessing the relationship between two or more variables.

When doing analyses which relate one variable to another, it is customary to classify variables as being either *dependent* or *independent*. Dependent variables are those believed to be influenced by other variables. It is the variation in dependent variables that is typically of interest. Ultimately we would like to be able to understand what causes or what accounts for this variation. The independent variables,[3] on the other hand, are those variables which are believed to do the influencing of the dependent variables and thus are believed to account for some of the variation observed in the dependent variable. These definitions become clearer with a few examples.

Suppose we want to know how educational expenditures rise as the number of students increases. The object of the analyses performed will be to determine how student-body size (the independent variable) affects expenditures (the dependent variable). We are attempting to explain how expenditures vary, from low to high, as influenced by the size of the student enrollment. Similarly, if the concern is with how program enrollments are affected by the sex of the participant, we want to relate sex (the independent variable) to enrollments (the dependent variable). If we're concerned with the effect of social class upon school participation, then the relationship between an independent variable called social class and the dependent variable (school participation) is of interest.

It should be clear that there are no hard and fast rules for classifying a variable as independent or dependent. It is situationally dependent, and the ultimate classification depends upon the question of interest. The same variable may be an independent variable in one instance and a dependent variable in another. We may, for example, wish in one case to relate educational attainment (the independent variable) to earnings (the dependent variable), and in another to relate social class (the independent variable) to educational attainment (the dependent variable). This is an example where

the variable for educational attainment is used as an independent variable in one instance and as a dependent variable in another.

Assessing relationships between variables is often done with the underlying assumption that understanding the relationship may permit some degree of control for changing things. This is often a fallacious notion. While this idea will be repeated several times in chapters to come, it warrants some mention here. A less ambitious reason for exploring relationship questions is that it permits increased description of the situation of interest to the planner. We may, for example, find that earnings increase with an increase in the educational attainment level of individuals. This may be a reasonable description of the relationship between those two variables in the sample or the universe of individuals for which data was collected. It is an entirely different thing to say, however, that if we increase the educational attainment levels of low-income peoples then their earnings will rise. This may be the unfortunate impression left by our findings, but one that may be quite unwarranted. Even though we have provided a description of the relationship between the two variables, our data do not permit us to determine the mechanism, or process, that gave rise to the data which we are describing.

Without knowing the mechanism, or underlying process, it is almost fruitless to suppose that policy recommendations can be made from the description alone. To influence events with policy may require a great deal of additional information about the situations involved, the culture, the history, the peculiarities of place or time, and a variety of other things not obtained with the quantitative data.[4] For many educational situations this can only be supplied, if at all, by individuals who know about the problem being studied. Policy recommendations must occur only after a synthesis of the descriptive results and the theories, propositions, beliefs, hypotheses, or presumed understandings of the mechanisms involved, and these must be supplied by people familiar with the problem. This is often a social process occurring in a political context. It may be that the planner analyst will participate in this synthesis process, not only as the producer of the analysis but as someone familiar with the situation of concern. Certainly, familiarity with the situation increases the likelihood that the analyst will recognize the limitations of the data.

3.5. *Determination of techniques* (Boxes 13 through 17 in Figure 1.1)

Just as the level of measurement determines the statistical procedures to be used with descriptive analysis, the kind of data determines the procedures to use when assessing relationships between variables. It is somewhat more complicated here because two or more variables must be considered, and the proper choice of technique depends on the kind of data for each variable. In Figure 1.1, Box 13, the various possibilities are arrayed.[5] Independent variables, classified as categorical or continuous, are related through analysis to dependent variables, also classified as categorical or continuous. The entries in Box 13 of Figure 1.1 indicate the appropriate techniques to choose for a given situation. For example, if both the independent and dependent

variables are categorical, then contingency table analysis is indicated. This is a fairly common kind of analysis problem dealt with by educational planners, with all tables of counts falling in this category. If, for example, enrollments in various programs are cross-classified by the sex of the participant, this information is easily presented in the form of a column of male enrollments in programs X, Y, and Z, as well as a column of female enrollments in the same programs (see Table 1.1). In this case we have what is called a *contingency table*, or a table of counts. Various statistics can be calculated to assess whether the variables are related and the strength of the association. Analysis of contingency tables is discussed in Chapter 3.

Table 1.1:

An Example of a Table of Counts Showing Program Enrollments by Sex

Program	Sex		Totals
	Males	Females	
Program X	10	40	50
Program Y	60	40	100
Program Z	30	5	35
Totals	100	85	185

Another form of tabled data occurs frequently in planning situations. Shown in Table 1.2 are data indicating that the average salary of teachers in urban areas of a national school system is nearly double that in rural areas, for both males and females. It also shows that females' salaries are generally less than salaries for males, irrespective of the region of the country. Many of the important features of these tables can be observed by mere inspection. But there are some simple analysis techniques which can be employed to obtain additional information from such tables, and which may in the course of the analysis uncover results undetected by mere inspection. The technique discussed in this text, called exploratory two-way analysis, is described in Chapter 4.

Table 1.2:

An Example of a Table of Means, Showing Average Teacher Salaries by Sex and School Location

Sex	Location	
	Urban	Rural
Male	1254	685
Female	812	396

As indicated in Figure 1.1 (Box 17), regression techniques are used for the analysis of data when both the dependent and independent variables are continuous. Regression would be the appropriate technique to use, for example, if we wished to relate educational attainment (in years of schooling) to earnings (in actual amount earned). Regression analysis has been described by some as the most powerful statistical tool available to the researcher or planner. But, as always, where there is potentially great utility for a technology, there is also much room for abuse. In Chapter 5 we deal with some of the problems in the application of regression analysis to planning data, as well as with the basics of regression and the assumptions which underlie the technique.

4.0 *Summary*

In this chapter, a schema for categorizing analysis techniques has been presented as a means to introduce terminology that will be used throughout the text, and to describe the purposes of the chapters to follow. Figure 1.1 may also be useful as a preliminary guide in the selection of an appropriate technique. If the analyst knows what general type of question is to be addressed and the kind of data involved, the appropriate box in Figure 1.1 will indicate an analysis technique for the situation. If, for example, we have a relational question, with both variables being interval level, then the information in Figure 1.1 identifies regression analysis as an appropriate technique.

However, Figure 1.1 does not provide a complete listing of the conditions necessary for the proper use of the techniques. There are many assumptions that must be met, for example, before a regression analysis can be performed. Each of the chapters to follow describes the further conditions that must be met for the proper use of a given procedure. In actual practice, not all conditions are likely to be met. In such circumstances the analyst is faced with a choice: abandon the analysis, choose a different technique if one is available (it often is not), or use the procedure and interpret the results with caution, knowing full well that some of the results might appear only because of the method employed. It is this latter alternative that seems most reasonable, and only adds to the oft-mentioned advice that the analyst remain cautious of what the analysis seems to say.

5.0 *Terms for Review*

- universe
- sample
- level of measurement
  - interval
  - ordinal
  - nominal
- population
- parameter

statistic
variable
  independent
  dependent
  categorical
  continuous
simple random sample

6.0 *Practice Exercises*

To add a degree of reality to the set of problem exercises that will appear at the end of each chapter, a data set from an actual planning situation has been selected for use. While the situational context has been disguised, the data are all taken directly from the data set used by the planners in responding to a request for information made by decision-makers in the operating system.

The Setting: North Central is a region in the northern and central parts of a country that need not be identified. The population in North Central was approximately 4 million in 1975. It is governed by a semi-autonomous group of administrators located in the major city of the region, but is also subject to direction from the national level. There are 436 smaller districts in North Central that comprise local school districts. Each of these is run by local educational authorities, but is also subject to some control, direction, and planning initiated at the level of the regional government.

The Problem: Because of population migration to more attractive areas in other parts of the country, North Central has experienced a steady decline in the number of students attending the publicly controlled schools in the region. Demographers estimate that, between 1975 and 1985, the following decline in the school-age population will occur:

Primary	*Secondary*
-48,460	-160,876

This decline of over 200,000 students in the next ten years is a problem of concern to educators in North Central. The magnitude of this enrollment decline would seem to indicate a reduction in staff of several thousand teachers and administrators.

To complicate matters, however, the population was not declining uniformly throughout the region. Within the region the two large cities were both experiencing growth. People were moving from the more remote areas into urban areas, attracted by hopes for increased opportunities. Migration also occurred between the rural areas and some of the larger towns in the region. The general outmigration and the internal migration patterns acted together to create fluctuating enrollments in the region. Some areas, particularly the remote rural areas, experienced a more rapid rate of enrollment decline than the average in the region. On the other hand, even though the overall trend was one of decline, there were some areas in and around the two major cities and some of the larger towns where there was only a moderate

enrollment decline or even an enrollment increase.

Educational authorities at the regional level were concerned that these fluctuating enrollment trends would have adverse effects on both the quality and the cost of education. They listed as likely consequences of fluctuating enrollments, the following:

Effects on Quality:

a. Staff reduction
   - increased pupil staff ratio
   - need for alternative staffing patterns
   - fewer new teachers, with an increase in age and experience of teachers

b. Reduction and elimination of programs

c. Closing schools

Effects on Cost:

a. Increasing percentage of teachers at top of salary schedule

b. Economies of education cannot be made proportional to declining enrollments

The Mandate: Planners in the regional education office were requested by educational authorities to "examine the impact of fluctuating school enrollments and their consequent effects on the quality and cost of education."

The Data: For our purposes, we will be concerned only with the quantitative data that were collected as part of the planning process. However, the planners viewed the analysis of this data as only a *part* of their entire study. They also conducted interviews with educational leaders at the regional and local levels; with members of government having special interests in education; with planners and demographers in other parts of the regional government; and with representatives of local organizations concerned with the operation of schools. They conducted public hearings so that people in the communities could attend and express opinions on how their schools or their children were being affected by fluctuating enrollments. The planners also conducted special studies into problems they felt to be related to fluctuating enrollments, such as transportation of children to schools, the extent and forms of current cooperation between neighboring school districts, and the extent of teacher mobility. Finally, they chose twelve school districts to study in more detail, and conducted extended interviews with administrators, teachers, students, and parents in these communities.

The planners also compiled a quantitative data base, to be used to address the problems identified earlier. It is this data base that will now be described and which will form the basis of the problems appearing at the end of each chapter. Variables for which information was collected, or could be collected given the circumstances, included the following:

Variable Name	Description	Categories
Enrollment	School district enrollment in 1975	no

Variable Name	Description	Categories
FTE Staff	Full-time equivalent teachers and staff	no
Salary	Average monthly salary of teachers and staff	no
Staff age	Age of staff	no
Staff experience	Years of teaching or administrative experience of staff	no
Staff training	Education of staff	1 = no degree 2 = university degree 3 = higher degree
Valuation	A measure of school district property values	1 = low 2 = medium 3 = high
Enrollment trend	Enrollment trend, 1970-1974	1 = 10% or more growth 2 = 6-9% growth 3 = 5% growth to 5% decline 4 = 6-9% decline 5 = 10% or more decline

Information from the school districts was obtained for all 436 districts, while information from staff was obtained in a sample of staff from throughout the region.

Notice that for most of the variables, the data were quite specific. For example the figures for enrollment in 1975 were recorded as the number of pupils enrolled that year. But notice also that in other cases the data represented less specific units. For example, salaries for each teacher were not available. Rather, average figures for all the teachers and staff in a given district were supplied. Similarly, information on the enrollment trend over the period from 1970 to 1974 was grouped into categories. In this case there were five categories. A school district with a 7% decline in enrollment over the period would, for example, fall into the 4th category.

This, then, is the data base that the planners used to address the questions given them by the regional education authorities. It is not a particularly rich data base, but this is often the case in planning. The important thing is to use what is available, subject to the constraints on time and resources involved in

collecting better data, to reflect upon the problem at hand.

With this as an introduction to the North Central data base, the following questions and problems can be answered using material from Chapter 1.

— There are several different universes, or populations, represented in the quantitative data collected by the planners. Name two such possible universes. Did the planners attempt to sample from these universes?

— Assign a level of measurement for each of the variables for which data were collected. Which of the variables are categorical? Which continuous?

— What kind of technique might be useful for analyzing features of school enrollment in 1975? What measures could be computed to provide summaries of this analysis?

— What kind of technique might be useful for analyzing features of staff training levels?

— One of the questions to be addressed by the planners concerned the effect of fluctuating enrollments on costs of education. Suppose the planners used average monthly salary as a measure of the cost of education offered in a school district. Further suppose that they wished to see how salaries varied with the size of the enrollment in 1975. What technique might be used to do this analysis? What is the dependent variable in the analysis? What is the independent variable?

— Using North Central variables, give an example that would require an analysis of a table of counts; that would require two-way exploratory analysis.

— Why do you think the planners conducted so many interviews, public meetings, and special studies? Why couldn't they just solve the problem with the data?

Footnotes

[1]See Hubert M. Blalock, Jr. *Social Statistics* (2nd ed.). New York: McGraw-Hill, 1972, p. 18.

[2]These definitions are slightly different from those that might be found elsewhere, the reason being that the author wants to avoid a problem with semantics. Technically, a variable is said to be continuous when it is possible to get ever more refined measurements between two points on the scale. Age qualifies as such a true continuous variable: we can measure the age as 8 years crudely; as 8 years, 3 months somewhat more accurately; as 8 years 3 months 5 days still more accurately; as 8 years, 3 months 5 days, and 7 hours to an even greater degree of accuracy, and so on until until the absurd is reached at seconds, milliseconds, or even beyond. Income, on the other hand, is strictly not a continuous variable. It is possible to say that in the United States a person's monthly income is $1,123.28; but beyond this there are no further dimensions to the measurement scale. It is impossible to have a monthly income of $1,123.28659, for example, because the smallest monetary unit in the United States is the penny ($.01), which means that income can be measured to only the nearest hundredth of a dollar. By using the definition given in the text, this problem of semantics is eliminated, and both age and income are considered continuous variables. It should be obvious that both variables utilize interval levels of measurement.

[3]The term *independent* is something of a misnomer, although it is commonly used and will be used throughout this text. By implication it seems to indicate that the variable may be independent and hence controllable by the planner. Such may be the case in an experimental situation when the experimenter can control the amount of treatment (i.e., the independent variable). But such is usually not the case for the educational planner, who usually deals with non-experimental data. In such circumstances the "independent" variables may be subject to some influence by the planners, but by so influencing the system of concern they are likely to affect not only the dependent variable of interest but other related variables as well. In this sense the *independent* variable is anything but independent. Some authors (see Mosteller and Tukey, 1977) have taken to calling independent variables by a different term, "carriers," in response to this problem. Since *independent* is widely used, it will be retained in this text, but with the warning that for most educational planning situations such variables are anything but independent.

[4]It is in this sense that our *independent* variable (in the example, educational attainment) may be anything but independent. Intervention on the part of the planner or decision-maker to raise general levels of education is not likely to leave the rest of the relevant system (culture, peculiarities of time and place, etc.) unchanged. In such circumstances intervention to change an *independent* variable may actually result in changes quite different from those anticipated on the basis of the data analysis.

[5]Strictly speaking, the table in Box 13 of Figure 1.1 is not complete. There should also be a row and a column devoted to *Mixed* variables. For example, what kind of analysis is indicated if the dependent variable is continuous and there are two independent variables, one categorical and the other continuous? This is an example of mixed independent variables. These situations are

not included, because in many cases suitable procedures for handling such data are the subject of current research. Note, too, that analysis techniques where the dependent variable is categorical and the independent variable is continuous are not dealt with in this text (Box 15). Such situations occur infrequently enough with planning data to make their inclusion here unwarranted.

Bibliography

Andrews, Frank M. et. al. "A Guide for Selecting Statistical Techniques for Analyzing Social Science Data." Ann Arbor: Institute for Social Research, University of Michigan, 1974. Uses a decision tree to help the planner select analysis techniques appropriate for use with a given kind of data. Quite detailed, but many situations included occur infrequently in practice.

Blalock, Jr., Hubert M. *Social Statistics*. 2nd ed. New York: McGraw-Hill, 1972. A general statistical text for social scientists.

Loether, Herman J., and McTavish, Donald G. *Descriptive Statistics for Sociologists*. Boston: Allyn and Bacon, 1974. An introductory text of statistical techniques, with the emphasis on their use in sociology. Includes many simple techniques of practical use for planners.

Mosteller, Frederick, and Tukey, John W. *Data Analysis and Regression*. Reading, Mass.: Addison-Wesley, 1977. This is a very useful general text, with an introduction into exploratory data analysis and a thorough discussion of regression techniques.

CHAPTER Chapter 2

Visual Data Displays

Gary Lewis

1.0 *What's Ahead?*

This chapter will explore some useful and simple data analysis techniques, each having visual components that offer effective and quick summaries of interesting features of the data. There are four principal objectives for the chapter.

(a) Techniques that are appropriate for use in different situations will be described and illustrated with examples. This will include stem-and-leaf diagrams for use with interval data; histograms for use with either ordinal or interval data; bar graphs for use with nominal data; statistical maps where the data are categorized by geographic area; and time series graphs for plotting trends.

(b) Where appropriate, statistics that summarize important features of analysis will be introduced and computational formulas provided.

(c) Using interval level data, an approach to comparisons among groups will be discussed, indicating how such comparisons can be made, and the degree to which we might believe there are significant differences between groups. In this regard the usefulness of data transformations is discussed, and examples of such transformations worked out.

(d) Finally, we return to the North Central data set to permit application of the techniques discussed in this chapter, and to concentrate on interpretation of the analyses.

2.0 *Displays for Interval Data*

2.1. *The Stem-and-Leaf Diagram*

The stem-and-leaf technique is useful for displaying a group of numbers in such a way that a visual impression of several features of the numbers is apparent. The easiest way to understand the stem-and-leaf diagram and some of the data features uncovered is to work through an example and then talk about some extensions and complications.

The following numbers are the ratios of pupils to teachers in school systems in selected cities in the United States: 24.0, 29.1, 32.3, 20.2, 24.0, 25.8, 26.5, 28.2, 18.7, 22.1, 23.6, 23.1, 24.0, 24.1, 25.4, 27.6, 27.5, 22.0, 23.8, 22.0, 23.1, 27.3, 28.8, 25.7, 27.2, 25.5, 22.9. By themselves these numbers are not very informative. It would be better if there were some way of displaying data so that the numbers take on additional meaning. Figure 2.1 is one such visual

presentation that is useful. This type of diagram is called a *stem-and-leaf display*. The numbers to the left of the vertical line are called *stems*; each digit to the right of the line is called a *leaf*. Together they represent one number. For example, the smallest number is 18.7, represented in the diagram as an 18 to the left of the vertical line and a 7 to the right of the line, or 187. Note that the stem-and-leaf form one whole number and must be multiplied by the unit scale to yield the original data. For example, 18.7 = 187(0.1).

Figure 2.1:

A Stem-and-Leaf Display of Pupil-Teacher Ratios in California, for School Systems with Enrollments Larger Than 25,000: 1971 (Unit = 0.1)

18	7
19	
20	2
21	
22	0019
23	1168
24	0001
25	04578
26	5
27	2356
28	28
29	1
30	
31	
32	3
33	

N= 28			
M 14h	24.6		
H 7h	23.1	27.3	4.2
1	18.7	32.3	13.6

Key

N= Number of items in diagram
M= Median
H= Hinges
1= Extremes

Source: National Center for Education Statistics, Statistics of Local Public School Systems: Pupils and Staff, Fall 1971, Washington, D.C.: National Center for Education Statistics, 1975, p. 63.

Having drawn the stem-and-leaf diagram, there are some simple things that can be observed:

— Pupil-teacher ratios in these school systems range from a low of 18.7 to a high of 32.3.

— The mid-range appears to be between 24 and 25.

— Most of the values are clustered somewhere near the midpoint, with fewer toward the extremes.

These statements summarize the concepts of *range*, or *spread*, *central tendency*, and *shape* of the distribution. A stem-and-leaf diagram is useful to make statements concerning these features of a group of numbers. Calculations discussed shortly permit these concepts to be succinctly summarized by single statistics. These statistics are shown in the open-faced rectangle directly below the stem-and-leaf display in Figure 2.1, and will be discussed in the next section. But for the present let's consider in general terms what the features of the data mean and why they're important.

A group of numbers representing data for a single variable and displayed in a diagram such as a stem-and-leaf display is said to illustrate the *distribution* of the numbers. The display shows how the numbers are distributed over the categories (i.e., stems) selected for use in the diagram. As already indicated, there are several features of such distributions that may be of interest. First, we would like to have some single measure which can be used to describe the entire group of numbers. Intuitively, this must be some central number which is close in value to as many of the other numbers as possible (see Figure 2.2 (a)). These values are useful for answering questions about "How much?" for groups, and for comparing groups when asking "Is it different from?" For example, we might categorize a group of school children taking an examination by stating that the *average* value for the group was 67%. *Or* we might compare two such groups of children, say boys and girls, by stating that one group averaged 58%, while the other averaged 75%. The concepts of *average* and other measures of central tendency are very useful and are treated in the following section.

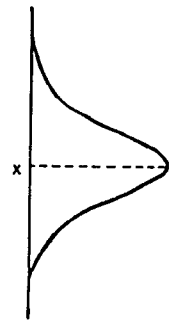
The second feature that is important to describe about a stem-and-leaf diagram is the spread among the data points. Is there, for example, a large or small range in the values (see Figure 2.2(b))? If we use a measure of central tendency to describe the group of numbers, are the other values clustered closely around this number or are they spread out? As we shall see in a later section, these questions about the spread of the numbers are important because they can help us answer the "Can I believe it?" concerns that are always present when you ask questions about group differences. If one group of students scored 58% on an examination and the second group scored 75%, how much confidence can the decision maker or the analyst have that there are significant differences between the universes from which the groups were selected?

Finally, the shape of the distribution of the numbers is of concern. Are there, for example, several numbers which are either large or small compared to the majority of numbers? Are most of the numbers located near one extreme, with a tail toward the other extreme (see Figure 2.2 (c))? These questions are important because they help us assess how good a description our measure of central tendency is. If values are distributed closely and symmetrically about some central measure, then we are likely to believe that

Figure 2.2:

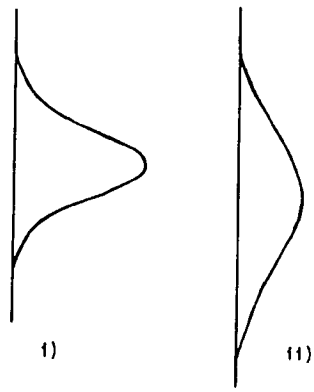
Features of Interest in Stem-and-Leaf Diagrams

a) The Concept of a Central Value



x is a measure of the central value for this group of numbers

b) The Concept of Spread



The values in ii) are spread out more than the values in i)

c) The Shape of the Distribution

The values in this distribution are more concentrated near one extreme than the other extreme.

this single measure provides a relatively good description of the entire group of numbers. However, if there is a lack of symmetry, with the distribution skewed, then our single measure is suspect. In such cases it is more representative of one end of the distribution than the other end where the tail lies. To help assess the shape of the distribution, another type of display called the *box* or *schematic plot* is described in a section to follow.

Before proceeding to more complicated applications of stem-and-leaf diagrams, let's reiterate several points about the simple case just discussed. First, a stem-and-leaf diagram is useful for displaying interval level data *on a single variable*. Secondly, the reason we want to do this is because the numbers by themselves often are not very illuminating unless we do something to them so that their meaning becomes apparent. Thirdly, we hope that by using such a display, we will be able to find out additional information about features of the data. In particular, we should be able to comment on the extreme values, what the range between the extremes is, what single value best describes the group, and to give a qualitative comment on the shape of the distribution. We are, in fact, responding to the need to answer several different "how much?" questions. For the data in Figure 2.1, we found that pupil/teacher ratios ranged from approximately 19 to a maximum of 32, with the most commonly occurring value somewhere around 24 or 25. This is simple, but potentially useful information to an educational decision maker.

The example given above was particularly nice for a stem-and-leaf display, but many numbers do not so readily lend themselves to the technique, and we have to make adjustments for that. One complication often encountered, and one readily dealt with, involves the appropriate number of stems to use. It is not useful to have either too large nor too small a number of stems. In the former case the numbers get spread out over too broad a range to be visually useful; in the latter case they are too condensed to show much detail. Something between 5 and 20 stems is usually sufficient. If the numbers cannot conveniently be placed in this many stems, it may be necessary to create special categories.

Figure 2.3 shows an example in which such special categories have been created to conveniently display data on educational attainment. Values of educational attainment range from a minimum of 0 to a maximum of 18 in Figure 2.3, which would require 19 stems if we were to display each year as a separate stem. This number of stems is bordering on the unwieldy, and one way to reduce the display to more manageable proportions is to include a broader range of values in each stem. So, for example, instead of separate stems for 12 and 13, we might create a single stem which then would display all units with educational attainments of 12 and 13. Just such broader ranges in stems are shown in Figure 2.3. Even broader ranges are possible, and one commonly sees numbers grouped into fives. For this the following symbols are often used: (• 0-4; and · 5-9). If this convention is used, any number between 15 and 19 could be represented as a leaf on the single stem labeled 1*.

Figure 2.3:

A Stem-and-Leaf Display of Years of Educational Attainment, for a Sample of Economically Active Persons in El Salvador: 1975 (Unit = 1.0)

a) Frequency Distribution

Years of education	0	1	2	3	4	5	6	7	8	9	10	11	12	...	18
Numbers of people with indicated educational attainment	34	3	14	10	6	5	21	1	0	4	0	1	5		1

b) Stem-and-Leaf Diagram

```
0·  | 00000000000000000000000000000000000111
0t  | 22222222222222333333333
0f  | 44444455555
0s  | 66666666666666666667
0*  | 9999
1·  | 1
1t  | 22222
1f  |
1s  |
1*  | 8
```

Key

· = 0 or 1
t = 2 or 3
f = 4 or 5
s = 6 or 7
* = 8 or 9

```
N= 105

M 53 |        3        |
H 27 | 0          6    | 6
1    | 0         18    | 18
```

Examples

"0*" includes all units with 8 or 9 years of education

"1·" includes all units with 10 or 11 years of education

Source: Center for Studies in Education and Development, "An Assessment of the Relevance of Education in El Salvador," (Harvard University, 1977), unpublished documentation.

Exercise 2.1: What features about the years of educational attainment in the El Salvador sample are evident from the stem-and-leaf diagram in Figure 2.3?

It should also be obvious that the stem-and-leaf technique is not useful when there are either too few or too many numbers to display. It would not be acceptable for displaying the achievement scores on a standardized test taken by 1000 eighth grade students. If the data set is large, a smaller group can be selected, as with a simple random sample (see Appendix) and then the sample units displayed in a stem-and-leaf diagram. This helps if there are too many numbers, but a stem-and-leaf diagram is never useful with a small number of values. For example, the salaries of teachers in three first grades in a school could not be usefully displayed with a stem-and-leaf diagram. There is simply no distribution to display.

2.1.1 Measures of Central Tendency. In this section we will consider two measures of central tendency, both easily calculated and occurring often in practice. The arithmetic *mean* or *average*, is one such measure. The formula for computing the mean is shown in equation (2.1).

$$\text{Mean} = \frac{\sum_{i=1}^{N} X_i}{N} \tag{2.1}$$

where X_i ($i = 1, \ldots, N$) are the N data points, and the Σ sign indicates that one must sum all of the data points.

The mean of the five numbers, 10, 12, 18, 20, 10 is $(10 + 12 + 18 + 20 + 10)/5 = 14$.

While the mean is a commonly used measure of central tendency, it does have one disadvantage. It is quite sensitive to values that are considerably different from the other numbers being used to compute the mean. Such values are called outliers, because they are extreme in the sense that they are much smaller or larger than the other numbers. Suppose that in our previous calculation of the mean, the value 12 had been 1000. The mean then is $(10 + 1000 + 18 + 20 + 10)/5 = 211.6$. In other words, the mean has changed from 14 to 211.6 because of the presence of one outlier value.

There are other summaries which are more resistant to the effect of outlier values. One such measure is the *median*. Suppose we use the same group of numbers as previously: (10, 12, 18, 20, 10). The *extremes* are the highest and the lowest values: 10 and 20. If the numbers are rank-ordered (from low to high, for example), the *median* is the single value in the middle between the two extremes. Since there are 5 numbers total in our example, the median will be the third number from either extreme since there would then be two numbers lower and two numbers higher than this central value. In this example the rank order is: (10, 10, 12, 18, 20), and the median is 12.

In a case where there is an even number of data points, there is no middle value. When this happens, the median is found by averaging the two middle values. Suppose that the above data set had another number, 8. There would be six numbers total, so no single number would have an equal number of data points smaller or larger. In this case we average the two middle numbers, 10 and 12 to get: $(10 + 12)/2 = 11$. The following formula can be used for determining which rank-ordered value will be the median:

$$\text{Rank Order of Median} = \tfrac{1}{2}(1 + \text{total count of data points}) \tag{2.2}$$

For example, if there are 27 numbers in the batch, the median value is 14th from the end ($14 = \frac{1}{2}(1 + 27)$). If there are 36 numbers in the batch, then $\frac{1}{2}(1 + 36) = 18\frac{1}{2}$ and the median is found by averaging the values which are 18th and 19th from either end in the ranked list of data values.

It is easy to show that the median has the attractive feature that it is resistant to outliers. In our previous example, the original numbers were 10, 12, 18, 20, 10 and the new set was 10, 1000, 18, 20, 10. For the first set of numbers, the median value is 12; in the second it is 18. This is a much less dramatic change than that for the mean which went from 14 to 211.6.

2.1.2 Measures of Spread. The amount of *spread* in a distribution is a second feature of concern and for which fairly simple summaries are available. Is the distribution fairly narrow, centered in a small band of values, or is it spread out, ranging from quite low to quite high? The *variance* is one measure commonly used to assess such questions. The formula used to calcu-

late the variance in a set of values is shown in equation (2.3).

$$\sigma^2 = \text{Variance} = \frac{1}{N} \sum_{i=1}^{N} X_i^2 - \overline{X}^2 \qquad (2.3)$$

where, in addition to the definitions for equation (2-1), $\overline{X}$ is the mean of the data values.

If the data set is that previously used (10, 12, 18, 20, 10), the variance in the set of numbers is:

$$\frac{1}{5}(100 + 144 + 324 + 400 + 100) - 14^2$$

$$= 213.6 - 196 = 17.6$$

The *standard deviation* (σ), which is the square root of the variance, is another commonly used measure of spread. For the above example, the standard deviation is $\sqrt{17.6} = 4.2$. Two symmetrical distributions having the same mean but different variances, or standard deviations, will appear to have different shapes. The distribution with the smaller variance will be more narrowly distributed about the mean, while the distribution with the larger variance will be spread out more around the mean. (See Figure 2.2(b)).

While the variance is a commonly used measure of spread, it too is not very resistant to outliers. Were the data set to be (1000, 10, 18, 20, 10) instead of (10, 12, 18, 20, 10), the variance would change from 17.6 to 155,410 — a considerable transformation! For this reason, as well as because of ease in computation, measures such as the interquartile range are also used.

The *interquartile range*, or *H-spread*, is the spread between *hinges* (abbreviated with the letter H). A *hinge* is the middle value between one extreme and the median. There is a simple formula for determining the number of values to count in from the extreme to locate a hinge. It is given as:

$$\text{Rank order of Hinge} = \tfrac{1}{2}(1 + \text{rank order of median}) \qquad (2.4)$$

For example, if there are 21 values, the median is the 11th from either extreme. The hinges are the 6th numbers from either extreme (i.e., $6 = \frac{1}{2}(1 + 11)$), when the numbers have been rank-ordered, as in a stem-and-leaf diagram. The five numbers represented by the median, the two hinges, and the two extreme points constitute what is called *a five-number summary* of the distribution. These summaries were attached to the stem-and-leaf display in Figure 2.1. Examples of the calculation of median and hinge values in Figure 2.1 will now be shown.

Median. In Figure 2.1 there are a total of 28 numbers. The median value is $\frac{1}{2}(1 + 28) = 14\frac{1}{2}$ (written as 14h in the figure). It is the average of the 14th and 15th number from either extreme. Thus the median is the average of the numbers 24.1 and 25.0, which is 24.55, (recall the unit in Figure 2.1 is 0.1).

Hinge. The rank of the hinge value is $\frac{1}{2}(1 + 14) = 7\frac{1}{2}$, and is the average of the 7th and 8th number from either extreme. The two hinges are therefore 23.1 and 27.3. Note in the formula for calculating the hinge value that if the median is a half-value (as above, where the median is $14\frac{1}{2}$), the half is dropped before using it in the equation.

Within the range from the lower hinge to the upper hinge (the H-spread),

50% of the values will be found. This is a useful and easily calculated measure of spread which is resistant to extremes. The H-spread on the data set (1000, 10, 18, 20, 10) is 10, not the large value found for the variance of this set of numbers.

The advantage of working with the median and H-spread, aside from resistance, is that for the type of work dealing with stem-and-leaf displays, these summaries are easily found with paper and pencil and involve essentially no computation, only counting from extreme positions.

2.1.3 The Shape of the Distribution In many cases a simple visual impression of a stem-and-leaf diagram is sufficient for deciding if the distribution is symmetrical or skewed toward one extreme or another. The distribution may also have other peculiarities, such as two or more modes where many values are concentrated. It is, however, possible to refine the description of the shape of a distribution by combining features from both the stem-and-leaf diagram and its five number summary.

A *box plot*, or a *schematic plot* as it is sometimes called, is really just a visual representation of the five-number summary table used to describe a stem-and-leaf diagram. Shown in Figure 2.4 is a stem-and-leaf diagram showing how pupil-teacher ratios in the primary education level are distributed for 117 countries of the world. Note that the five number summary shows that the median value is 33, while the interquartile spread is 16. The extremes are 14 and 77, and the visual impression from the diagram is that values tail out toward the higher end of the scale.

Shown in Figure 2.5 is a box plot of this same information. To construct a box plot the following steps are necessary:

1) Complete a stem-and-leaf diagram and the five number summary.
2) Draw a convenient vertical scale which will include the range of numbers in the stem-and-leaf diagram.
3) Draw a horizontal line at the level of the median.
4) To complete the box part of the diagram, draw horizontal lines at each hinge.
5) On either side of the hinge, draw a dotted line the length of the interquartile spread. Place an X to indicate the value which is nearest but does not exceed the end of the dotted line, and stop the dotted line at the X. It may be that one dotted line will be shorter than the other. Such a situation could occur if the extreme value at one end of the distribution were close to the median while values tailed out at the other end. Just such a situation occurs in Figure 2.5.
6) If any values in the stem-and-leaf diagram are outside of the X marks, indicate these outliers with circles.

Outliers, or those values far removed from other values, are potentially interesting and are specially designated in the box plot as a signal to their presence and to indicate that one might wish to investigate what makes them different from the other data points.

Figure 2.4:

A Stem-and-Leaf Display of First Level Pupil Teacher Ratios for 117 Countries, Latest Figures as of 1974 (Unit = 1.0)

```
1·  |
1t  |
1f  | 45
1s  | 66677
1*  | 9
2·  | 0001111
2t  | 2222233
2f  | 4444555555
2s  | 6666677
2*  | 999
3·  | 0001111111
3t  | 222233333
3f  | 445555555
3s  | 6666677
3*  | 8888999
4·  | 011
4t  | 333
4f  | 44444555
4s  | 7777
4*  | 899
5·  | 0
5t  | 33
5f  | 5
5s  | 7
5*  |
6·  | 0
6t  | 3
6f  | 55
6s  | 6
6*  |
    :
7s  | 7

N= 117

M 59 |      33      |
H 30 | 25       41  | 1f
1    | 14       77  | f3
```

Source: UNESCO, Statistical Yearbook. New York: United Nations, 1973 and 1974 .

From Figure 2.5 it is very clear that the distribution of pupil-teacher ratios is skewed toward the higher values. There are six such numbers which appear outside the X marks, and they are all on the high extreme of the distribution.

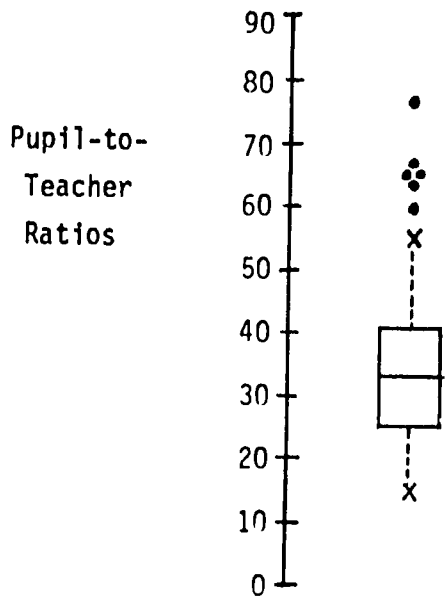
Such diagrams add yet another visual dimension to assessments of shape, but we will see in a later section that they are also useful when making comparisons between distributions.

2.2 *Data Transformations*

It was earlier remarked that measures such as the mean, or median, were

Figure 2.5:

A Box of Pupil-Teacher Ratios in Figure 2.4



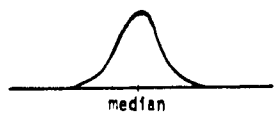
more descriptive when the distribution is symmetrical. Shown in Figure 2.6 are hypothetical distributions, one which is symmetrical about the measure of central tendency, and the other which is not symmetrical. It should be clear that the measure of central tendency for the asymmetrical distribution is not as good a description of some of the extreme values in the tail as it is of the values in the other extreme.

An analyst may sometimes want to describe a distribution by a single number, even though the distribution is not symmetrical. In such situations steps can be taken to yield better descriptive measures of central tendency, by transforming the original data, using some functional expression. For example, we might take the square root of each number, or the logarithm of each number, and then display these to see the effect upon the shape of the distribution. This may yield a measure of central tendency which is more descriptive of the entire distribution. This is so because it is often possible, by choosing the right transformation, to make an asymmetrical distribution into a roughly symmetrical one.

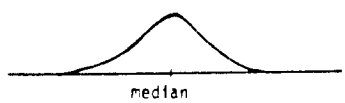
Figure 2.6:

Illustration of the Descriptive Quality of a Measure of Central Tendency, in Various Distributions

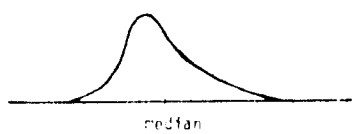
a) Symmetrical distribution, small variation: median provides a good description of values in the distribution



b) Symmetrical distribution, large variation: median provides a good description of values in the center of the distribution, but is less adequate for values in the tails of the distribution



c) Asymmetrical distribution: median provides a better description of values in the central portion and one tail of the distribution than in the other tail



A disadvantage of transforming the data is that the original units are not used in describing the measures of central tendency. Instead of saying that "the average pupil-teacher ratio is . . ." we end up saying that "the average of the logarithms of pupil-teacher ratios is . . ." Many decision makers find it easier to use units which have operational meaning, irrespective of whether the measure of central tendency is a good description of the entire distribution or not.

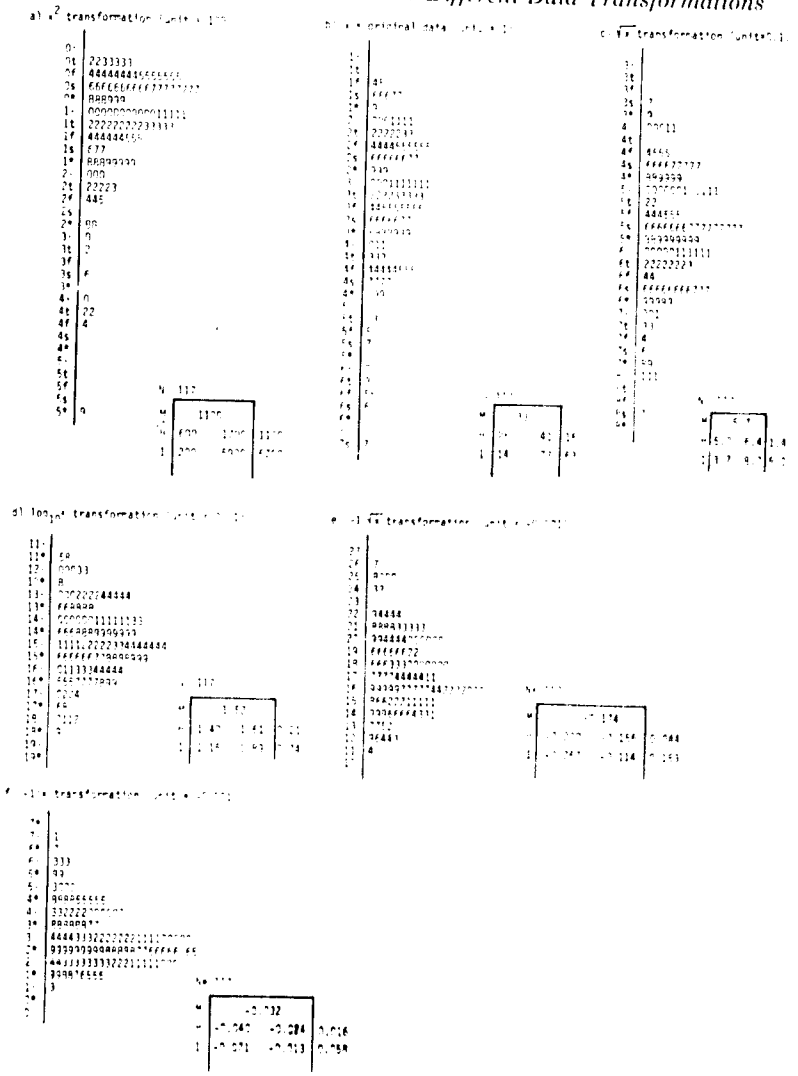
However, data transformations can be very useful when comparing groups. Transformations are also used to advantage with several of the analysis techniques discussed in later chapters. It often happens that the initial analysis of a given set of data is not very satisfactory. By transforming the data, the analyst may gain new insights into the interpretation of results. For these reasons this section will explore the influence of data transformations upon the shape of a distribution. Simple rules for deciding upon which transformation to use will be provided. In the following section we will come back to the data in Figure 2.4, and ask how pupil-teacher ratios compare between groups of nations. To answer this question, we will have to make use

of transformed data when addressing the "Can I believe it?" portion of the question.

Shown in Figure 2.7 are a series of stem-and-leaf diagrams for the pupil-teacher ratios in nations of the world, with each diagram showing the effect of a data transformation. The original data is marked x. The data marked x^2 is transformed by squaring each of the original data items and then plotting this new set. Similarly, the diagram marked $-1/x$ uses the negative reciprocals of the original data.

Figure 2.7:

Stem-and-Leaf Diagrams of the Pupil/Teacher Data for 117 Countries, as Shown in Figure 2.4, but under Different Data Transformations



The important thing to note in Figure 2.7 is the effect upon the shape of the original distribution. The original distribution tails out toward higher numbers. It appears that the x^2 transformation accentuates this tailing. The square-root transformation alleviates the tailing to some extent, as does the logarithmic transformation. The other two transformations, the negative of the reciprocal of the square root and the negative of the reciprocal, transform the data in such a way that there are now outlier values at the smaller extreme.

If we are looking for a transformation that will make the distribution roughly symmetrical, then either the square root or the logarithmic transformations are good bets. The x^2 transformation goes in the opposite direction from what we want, accentuating a feature we wish to get rid of. On the other hand, the final two transformations go too far and create a tail of values at the smaller extreme.

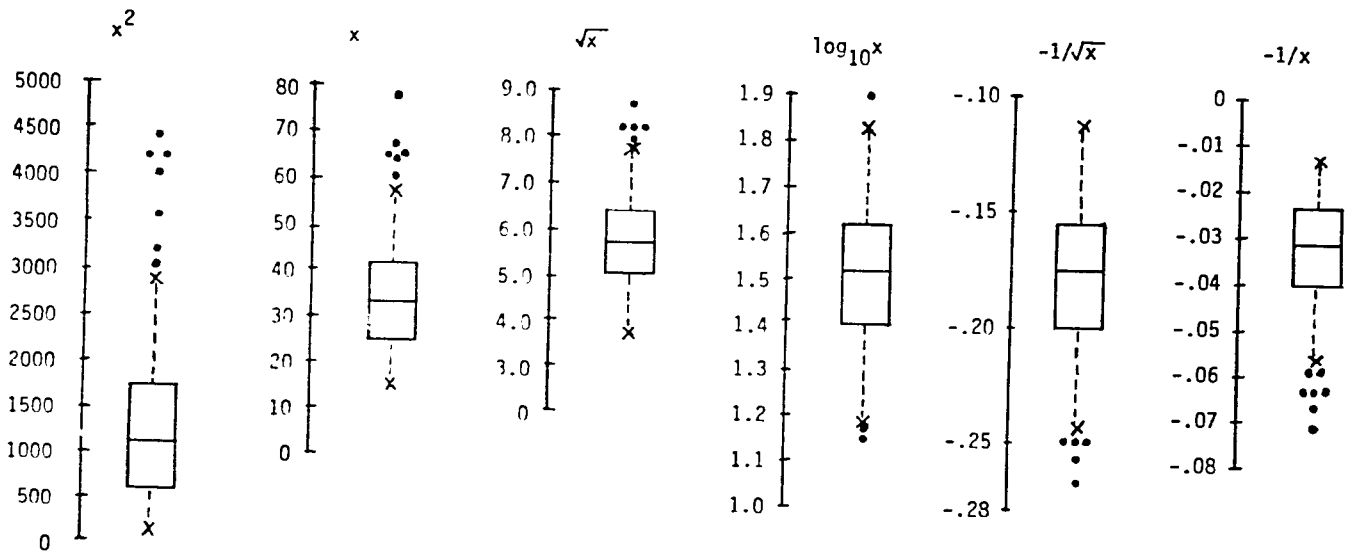
These trends are shown quite clearly in Figure 2.8, where the box plots for all these data transformations are shown. Looking particularly at the outlier values shown in the box plots, note the steady progression as you go from left to right in the figure. The x^2 transformation spreads the outlier values out even more than in the original data. The square-root transformation has succeeded in eliminating some of the tailing, although there are still 4 outlier values, while the logarithmic transformation seems to have done about the best. Note that here there are three outliers, one at the high extreme and two at the low extreme. By the time we get to the last two transformations there are an increasing number of outliers, and they are all at the low extreme of the distribution.

Is there any way that one could choose a transformation without going through the trouble of computing all the transformations and then plotting and comparing the distributions? Fortunately, the answer is a qualified yes. There is no rule that will tell you exactly which transformation to choose for a given distribution so as to transform it into a symmetrical one. But there is a rule that will tell you roughly how to proceed in selecting a transformation to try. Furthermore, with a few simple computations, a quite satisfactory transformation can often be selected.

The rule to be applied in selecting the proper transformation makes use of the steady progressions evident in Figures 2.7 and 2.8. Note that there is a scale, from x^2 to x to $\sqrt{x}$ to $\log x$ to $-1/\sqrt{x}$ to $-1/x$. Moving from the original data (x) toward the x^2 end of the scale will emphasize the differences among the larger values in the distribution. Moving toward the $-1/x$ end of the scale will emphasize differences among the smaller values in the distribution. In thinking about how to use this rule with the pupil-teacher data, we note that differences among large values are already apparent in the original data. In fact, that is precisely the problem — the distribution tails out toward higher values. It doesn't help much to move from x to x^2. Rather, we want to emphasize differences among small values, in essence spreading them out more than they are in the original data. Of course, if we move too far toward the $-1/x$ end of the scale we will overcorrect and end up with a tail at the small numbers.

Figure 2.8:

Box Plots of the Transformed Pupil/Teacher Data Shown in Figure 2.7



As shown in Figures 2.7 and 2.8, there is a scale to the transformations in the sense that they have different effects upon the original data. Some transformations emphasize differences in large numbers, while others emphasize differences in small numbers (i.e., spread out the small numbers). This fact permits us to make use of a general rule when selecting a data transformation. The rule is this: when selecting a data transformation, move along the scale of transformations in the direction that will alleviate the tailing problem. For example, with the pupil/teacher data in Figure 2.4, the numbers tailed out at the high end. This distribution could be made more symmetrical if we emphasized the differences and spread out the numbers at the low end of the distribution. To do this we must move toward the $-1/x$ end of the scale to select a transformation.

Now that the general direction of movement along the scale of transformations has been decided upon, is there any way to select one transformation out of the many possibilities still remaining? In fact there is. It requires the use of five-number summary figures from the original data. These numbers are transformed with the various transformations until one is found that appears symmetrical, based upon extremes, hinges, and medians. Shown in Figure 2.9 is a series of five-number summaries for the pupil/teacher example, using the various transformations that initially appear reasonable. Using the transformed five-number summaries, two ratios are computed for each transformation. The first is the ratio of the median-to-hinge distances. It is found as:

$$\text{Ratio of median-to-hinge distances} = \frac{\text{Median} - \text{Lower Hinge}}{\text{Upper Hinge} - \text{Median}}$$

For example, with the original data, this ratio is given as:

$$\text{Ratio of median-to-hinge distances} = \frac{33 - 25}{41 - 33} = \frac{8}{8} = 1.00$$

Similarly, the ratio of the median-to-extreme distances is found as:

$$\text{Ratio of median-to-extreme distances} = \frac{\text{Median} - \text{Lower Extreme}}{\text{Upper Extreme} - \text{Median}}$$

These ratios give some indication of the symmetry under the transformations and save the analyst from the chore of first transforming all the data points and then plotting them in a stem-and-leaf diagram to determine symmetry.

If the ratios are 1.0, it indicates that a particular transformation has succeeded in placing the median equal distance between the two hinges, or between the two extremes, depending upon which ratio is in question. By using these ratios it is possible to select an appropriate data transformation, because we look for those transformations which yield ratios of about 1.0. From Figure 2.9 we see that the square-root transformation succeeds in producing symmetry between median and hinges but is not so successful with the extremes, which still tail slightly to the higher numbers. The logarithmic transformation succeeds in producing a median value squarely between the

Figure 2.9:

Illustration of a Procedure for Selection of an Appropriate Data Transformation, Based Upon Five-Number Summaries

Transformation	Five Number Summary	Ratio of Median-to-Hinge Distances	Ratio of Median to Extreme Distances
x	M 33 H 25 41 1 14 77	8/8 = 1.0	19/44 = .43
$\sqrt{x}$	M 5.7 H 5.0 6.4 1 3.7 8.7	.7/.7 = 1.0	2.0/3.0 = .67
$\log x$	M 1.52 H 1.40 1.61 1 1.15 1.89	.12/.09 = 1.33	.37/.37 = 1.0
$-1/\sqrt{x}$	M -.174 H -.200 -.156 1 -.267 -.114	.026/.018 = 1.44	93/60 = 1.55
$-1/x$	M -.032 H -.040 -.024 1 -.071 -.013	.008/.008 = 1.00	.039/.019 = 2.05

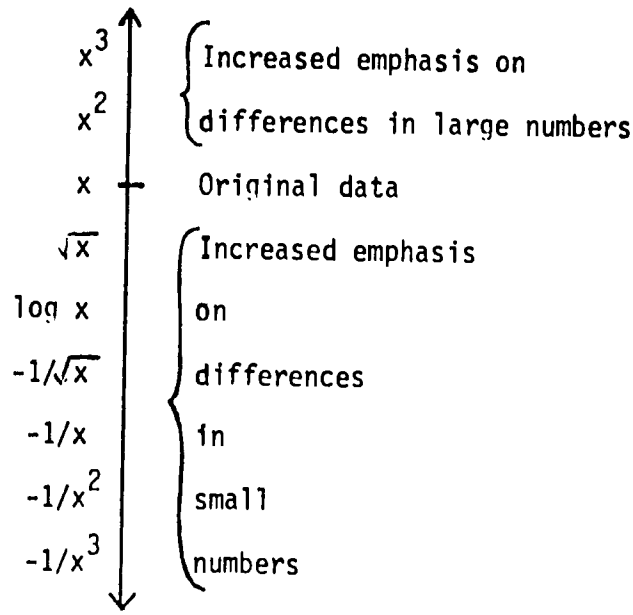
two extremes, but causes an overcorrection in the hinges. So in essence we've narrowed the selection process down to two possibilities, either of which may be satisfactory. These same two transformations were also seen in Figures 2.8 and 2.9 to do the best job of producing symmetry when the entire set of data was used. The process illustrated in Figure 2.9 is convenient because it saves a good deal of time and extra effort.

Shown in Figure 2.10 is a somewhat more extensive scale for making decisions about transformations of data. In principle it is exactly similar to the scale just discussed, except that more transformations are included. We will return to this figure in later chapters when transformations are discussed in connection with other procedures.

2.3 *Comparisons Among Groups*

In this section we will explore one approach to comparing groups. In Figure 2.4, pupil-teacher ratios for nations of the world were displayed. But we

Figure 2.10:
A Scale for Data Transformations



might expect that this single distribution contains richer information if only we had more detail about the countries. For example, it is not unreasonable to assume that wealthier countries would be better able to employ greater numbers of teachers, teacher aides, and administrators per pupil than some of the poorer countries. We might test out this idea by classifying nations by some indicator of wealth — say, per capita GNP — and constructing separate stem-and-leaf diagrams for each classification. This has been done, and the results are shown in Figure 2.11.

Several things are obvious from Figure 2.11. It does appear that the wealthier countries have lower pupil-teacher ratios. The median ratio for countries with per capita GNP in excess of $2000 is 22, while the median for countries with per capita GNP less than $500 is 38. Countries with a mid-range per capita GNP, from $500 to $2000, also have a median pupil-teacher ratio which is between the values for the wealthier and poorer countries. But we can also note from the stem-and-leaf diagrams that the distributions overlap considerably. Not all wealthy nations have pupil-teacher ratios lower

than the poor countries. And the distribution for the mid-range countries overlaps both the distributions of the wealthy and the poorer countries. The question then arises: *Can we believe* our results which seem to indicate that pupil/teacher ratios differ, depending on the general level of wealth in a nation?

Figure 2.11:

Stem-and-Leaf Diagrams of First Level Pupil-Teacher Ratios for 117 Countries, by Per Capita Gross National Product, Latest Figures as of 1974 (Unit = 1.0)

a) GNP < $500

1·	
1t	
1f	
1s	
1*	
2·	
2t	2
2f	44455
2s	6
2*	9
3·	0111
3t	222233
3f	455555
3s	66667
3*	888899
4·	011
4t	33
4f	44444555
4s	777
4*	899
5·	0
5t	33
5f	
5s	
5*	
6·	0
6t	3
6f	55
6s	6
6*	
⋮	
7s	7

b) GNP $500-2000

1·	
1t	
1f	4
1s	6
1*	9
2·	01
2t	33
2f	4
2s	6667
2*	9
3·	00111
3t	333
3f	455
3s	67
3*	9
4·	
4t	3
4f	
4s	7
4*	
5·	
5t	
5f	5
5s	7
5*	
6·	
6t	
6f	
6s	
6*	

c) GNP $2000+

1·	
1+	
1f	5
1s	6677
1*	
2·	00111
2t	2222
2f	5555
2s	667
2*	9
3·	1
3t	
3f	
3s	
3*	
4·	
4t	
4f	
4s	
4*	
5·	
5t	
5f	
5s	
5*	
6·	
6t	
6f	
6s	
6*	

N= 63			
M 32	38		
H 16h	32.5	45	12.5
1	22	77	55

N= 31			
M 16	31		
H 8h	25	35	10
1	14	57	43

N= 23			
M 12	22		
H 6h	20	25	5
1	15	31	16

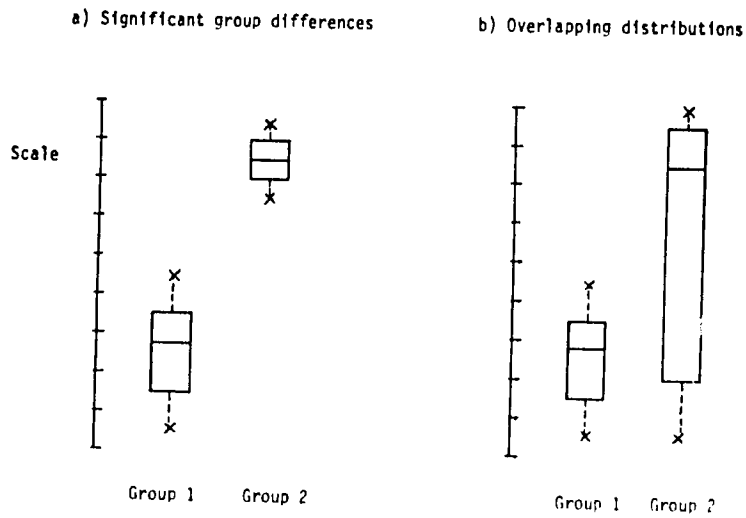
Source: UNESCO. Statistical Yearbook. New York: United Nations, 1973 and 1974; International Bank for Reconstruction and Development. World Bank Atlas. Washington, D.C.: IBRD, 1975.

Overlapping distributions should give us pause when comparing groups. Consider, for example, the hypothetical box plots shown in Figure 2.12. In part (a) of that figure we would probably be quite confident in saying that

there was a significant and considerable difference between the units in group 1 and the units in group 2. There is no overlap in the two distributions. But what about the situation in part (b) of Figure 2.12? The medians are identical to those in part (a), but one significant change has been made — the spread in the distribution of group 2 is considerably different than in part (a). How does that alter our interpretation of the differences between groups 1 and 2? Can we believe that there are differences between the two groups?

Figure 2.12:

Box Plots for Hypothetical Distributions for Two Groups



The spread around the median in Group 1 is relatively small, and thus the median is a pretty good measure of the values in that group. But the problem occurs because the spread in Group 2 is quite large. This means that our median is not very typical of many of the values in that distribution. It is far better to make comparisons of group values like means or medians when the spreads of the distributions are about the same. If we have two groups and the spreads about the median are both small, then a fairly accurate statement can be made comparing the two groups, since each median is a fairly good representation of the values in its distribution. If both spreads are large, then neither median is a good representation of the values in the distribution, and we may simply be faced with the situation that nothing meaningful can be said by way of comparison unless there is little or no overlap in the values of the distributions. It is when there are differences in the spreads that ambiguity most often arises, as in part (b) of Figure 2.12. Can we believe a comparison of Groups 1 and 2, when the median for Group 1 is a fairly good representation of its group, but the median for Group 2 is not typical for many

in its group? Without really answering this question, let's just say that there may be ways of getting around the problem. The method for dealing with differences in spreads again requires data transformation.

Before discussing how data transformations can sometimes help when making comparisons among groups, let's specifically identify some features of group comparison that have been evident in the previous statements:

— A single number is used to represent each group. This is usually the mean, or the median, since we want the single number to be as representative of the entire group as possible.

— Groups are compared using the means or medians representing the groups.

— Differences in group means or medians do not necessarily indicate significant differences between groups. We also must consider the spread of the groups.

— Particular difficulty arises when the spreads are large and result in distributions that overlap to a considerable extent. Depending upon the degree of spread and overlap between groups, we can qualitatively assess the believability of the group comparisons.

— Data transformation can sometimes help even out spreads between groups and make for more meaningful group comparisons.

In the previous section, we saw that data transformations could help eliminate asymmetry from distributions. When comparing several distributions, however, it is more important to find a transformation that will even out the spreads than it is to find a transformation that will cause all the distributions to be symmetrical. In Figure 2.13 box plots are shown of the data in Figure 2.11, where pupil-teacher ratios were categorized by the level of wealth of a country. Note that the spreads steadily decrease as wealth increases. The interquartile spread for the distribution of countries with GNP's of more than $2000 is only 5, compared to the value of 12.5 for countries where the GNP is less than $500. This is not a severe case of uneven distributional spreads, but for instructional purposes the corrective steps will be illustrated.

Actually, there is an easy method of selecting the proper transformation to equalize spreads, once the original stem-and-leaf displays have been prepared for each of the groups of interest. Shown in Table 2.1 are a series of transformations on the hinge values taken from the five number summary tables for each of the stem-and-leaf diagrams in Figure 2.11. We end up choosing that transformation which most nearly equalizes the interquartile spreads. As can be seen from Table 2.1, both the $-1/\sqrt{x}$ and $-1/x$ transformations do a pretty good job (note again the progression in effects as you move from left to right in the figure, or down the scale presented in Figure 2.10). If we use the $-1/x$ transformation, on the basis that the percentage differences between the smallest spread and the largest spread is minimized with this transformation, then it is a simple matter to construct the five-number summary table and the box plots for this transformation, without having to go through the process of transforming all the data items.

Figure 2.13:

Box Plots of Pupil-Teacher Ratios for Countries, by Per Capita Gross National Product, Data from Figure 2.10

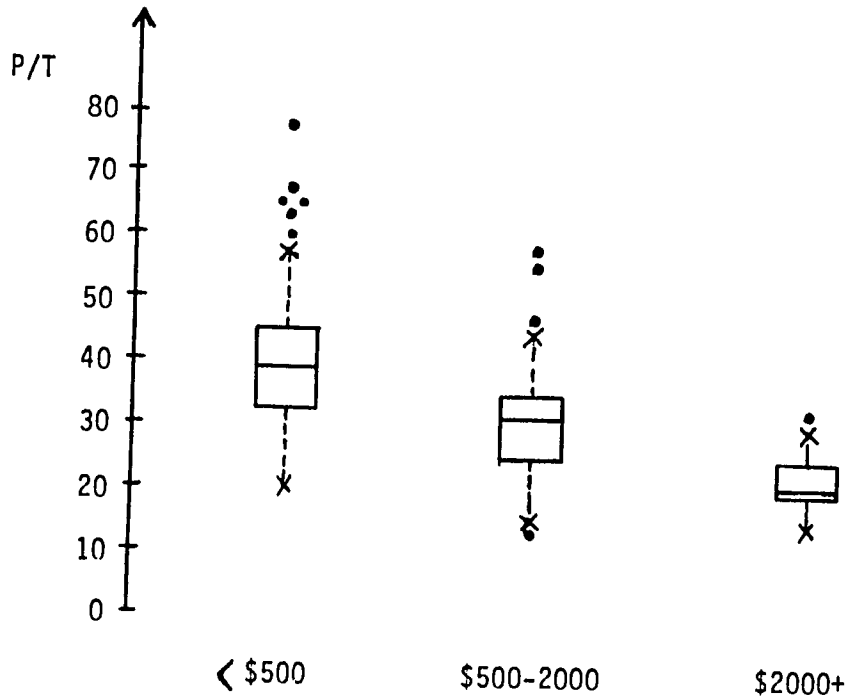


Table 2.1:

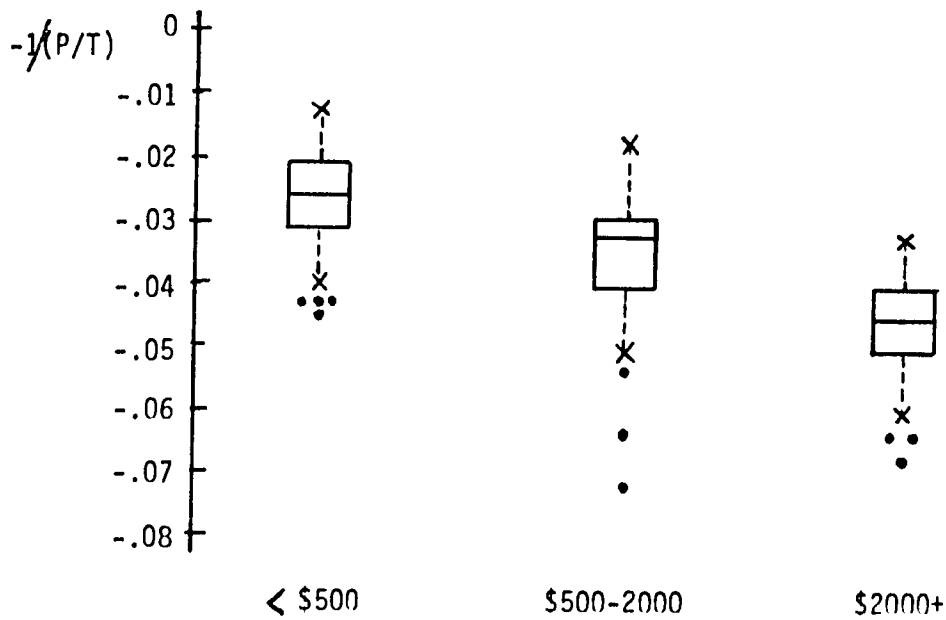
Transformations of Hinge Values and H-Spread for the Pupil Teacher Ratios by Country, for Data in Figure 2.11.

GNP	Hinges	Transformation				
		x(original)	log x	$-1/\sqrt{x}$	$-1/x$	$-1/x^2$
<$500	Lower H	32.5	1.51	-0.175	-0.031	-0.0009
	Upper H	45.0	1.65	-0.149	-0.022	-0.0005
	H-spread	12.5	0.14	0.26	.09	0.0004
$500-2000	Lower H	25.0	1.40	-0.200	-0.040	-0.0016
	Upper H	35.0	1.54	-0.169	-0.029	-0.0008
	H-spread	10.0	0.14	0.31	0.11	0.0008
$2000+	Lower H	20.0	1.30	-0.224	-0.050	-0.0025
	Upper H	25.0	1.40	-0.200	-0.040	-0.0016
	H-spread	5.0	0.10	0.24	0.10	0.000[illegible]

Shown in Figure 2.14 are the box plots for the -1/*x* transformation. It is quite evident from this figure that the median values for the three groups are considerably different, and that there is a steady progression, with poorer countries generally having higher pupil-teacher ratios than the more wealthy countries.

Figure 2.14:

Box Plots of Pupil-Teacher Ratios for Countries, by Per Capita Gross National Product, Under a Negative Reciprocal Transformation, Data from Figure 2.11



3.0 *Displays for Ordinal Data*

3.1 *Histograms with Interval Data*

The histogram is a technique for displaying both interval and ordinal data. In this section, we will describe the basic procedure for constructing histograms using interval data. Its use for ordinal data will then appear quite straightforward, and is treated in a section to follow.

Shown in Figure 2.15 are examples of two kinds of histograms, both showing the distribution of the average number of school days attended, for states in the United States. We use such a display almost like the stem-and-leaf diagram. A histogram allows the analyst to make some judgements

about central tendency, to gauge the spread of the numbers involved, and to assess visually the shape of the distribution. With ordinal data the use is more limited, because of the nature of the data, but it can still be used visually to illustrate the shape of the distribution. This will become clear momentarily.

For interval data, the basic ideas of the histogram are quite simple. They can be summarized in the following steps:

a) The entire *area* inside the rectangular segments of the histogram represents the total number of units. In Figure 2.15, there are 50 units (i.e., 50 states), and we want to represent these 50 units as the total area inside the figure.

b) The first step in constructing a histogram is to establish *class intervals*. In Figure 2.15 the numbers for average days attended run from a low of 145 to a high of 172. First, we divide up this range into equal-sized intervals. How this is done is somewhat arbitrary, but it should result in neither too few nor too many intervals. Something between 5 and 10 class intervals is usually sufficient. In Figure 2.15 there are six class intervals, each with a width of 4 units.

c) After establishing class intervals, the number of units falling in each category is counted. This is the *frequency* for that interval. From the table in Figure 2.15 it is possible to see that there are 9 states with 164 to 168 average number of days attended per pupil. These frequencies can be converted to *percentages* by dividing the total frequency for each interval by the total number of units being considered and then multiplying by 100. Thus the percentage for the class interval 164–168 is (9/50)100 = 18%, (e.g., 18% of the units fall in the range from 164 to 168).

d) Either the frequency or the percentage distributions can be plotted. Along the vertical axis are the frequency or percentage values. The horizontal axis represents the range of values, with class intervals marked off. Typically the midpoint of each interval is identified, as they are in Figure 2.15. The height of each rectangle corresponds to the frequency or the percentage of units in that class interval. Thus the rectangle over the class interval from 164 to 168 is 9 units high on the frequency histogram (Part a) and is 18 units high on the percentage histogram (Part b). For purposes of interpretation, it does not matter whether a frequency or a percentage histogram is plotted. The scale on the vertical axis of the percentage diagram can only go from 0 to 100, which may make for some degree of simplicity in plotting, but either may be plotted.

As always, however, there may be complications. One common complication occurs when the original data are not available to the analyst. If the data have already been classified into class intervals, and the original data are not available, two potential problems could occur. First, the class intervals could be of unequal sizes. Secondly, one or both extreme intervals could be open-ended, meaning that all units less than (or more than) a given value have been grouped into one category. Both of these problems occur frequently.

In Figure 2.16, an example of a histogram with unequal class intervals and one open-ended interval is displayed. These data were from a large sample

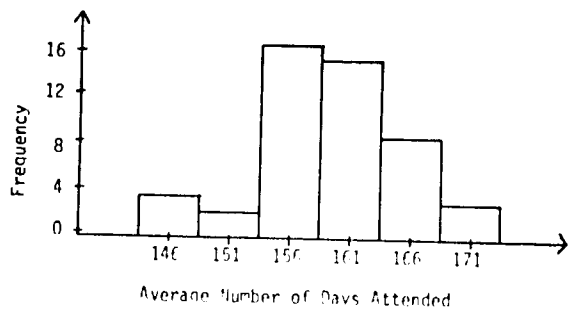
Figure 2.15:

Frequency and Percentage Histograms of the Average Number of Days Attended Per Pupil Enrolled, by State for the United States: 1973-74.

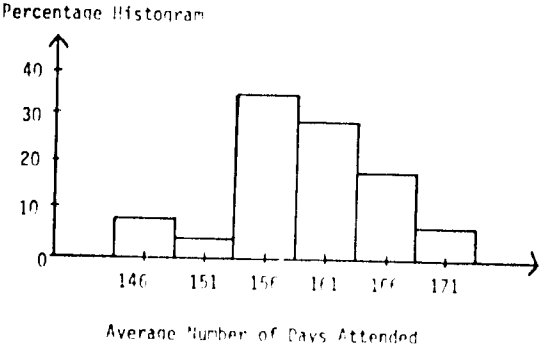
a) Data

Class interval	Frequency	Percentage
144-148	4	8
149-153	2	4
154-158	17	34
159-163	15	30
164-168	9	18
169-173	3	6

b) Frequency Histogram



c) Percentage Histogram

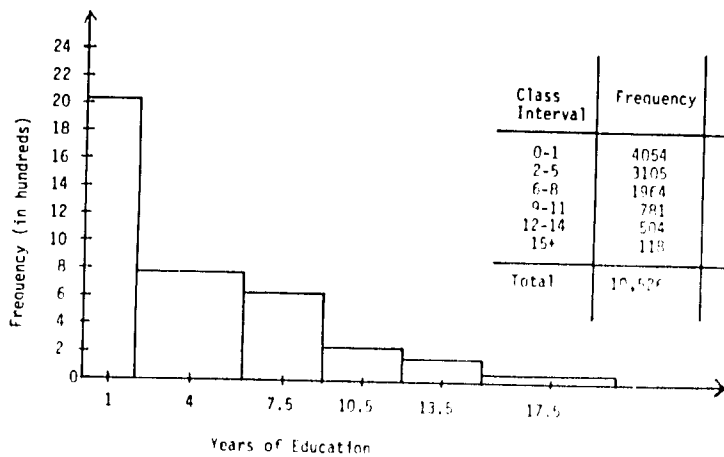


Source: National Center for Education Statistics, 1977 Digest of Educational Statistics. Washington, D.C.: U.S. Government Printing Office, 1977, Table 34.

survey of the workforce in El Salvador. Note that the intervals in the table vary from size two (0–1 years) to size four (2–5 years). Also, the highest interval is open (15 years and above). In Figure 2.16, the open-ended interval has arbitrarily been closed by assuming that nobody has more than 20 years of education. When reasonable bounds can be determined, the diagram can be completed, but this is not always possible.

Figure 2.16:

Years of Education for the Economically Active Population in El Salvador, 1975



Class Interval	Frequency	Percentage
0-1	4054	39
2-5	3105	29
6-8	1964	19
9-11	781	7
12-14	504	5
15+	118	1
Total	10,526	100

Source: Center for Studies in Education and Development. "An Assessment of the Relevance of Education in El Salvador." Harvard University, 1977, p. IV-14.

It is also possible to deal with the problem of unequal class intervals, as long as we recall that it is the total *area* in the histogram that represents the total frequency. If the frequencies were plotted as expressed in the table, the visual impression would be wrong. Note that the 4054 individuals with 0 to 1 years of education have been plotted with a bar two units wide and height of 2027 (i.e., the total area is 2 × 2027 = 4054). Individuals with 2–5 years of education must be represented as a bar 4 units wide and height 776, so that the total area is 4 × 776 = 3104, and the bar will be in the correct proportion to the bar from the 0–1 group. Similarly, the 1964 people with 6–8 years of education would be represented in a bar 3 units wide with height of 1964/3 = 655. This bar contains 19% of the area in the whole figure (see Figure 2.16). The disadvantage of using unequal class intervals becomes clear, because to make sense of the histogram one must recall that it is the *area* of the bars one is comparing, not just the height.

Figures 2.17 and 2.18 show two additional histograms. In Figure 2.17, initial incomes for graduates of postsecondary schools in the Sudan are shown. It is similar in construction to the other histograms just discussed, but not all histograms need follow this same pattern. Figure 2.18 shows one interesting possibility. Essentially the histogram has been turned on its side and a second one added. The right-hand side shows the age structure in the city of Newton, Massachusetts, and the left a comparison age structure for the United States. In addition to providing a visual description of the distribution, as all histograms do, Figure 2.18 uses a second histogram for comparative purposes.

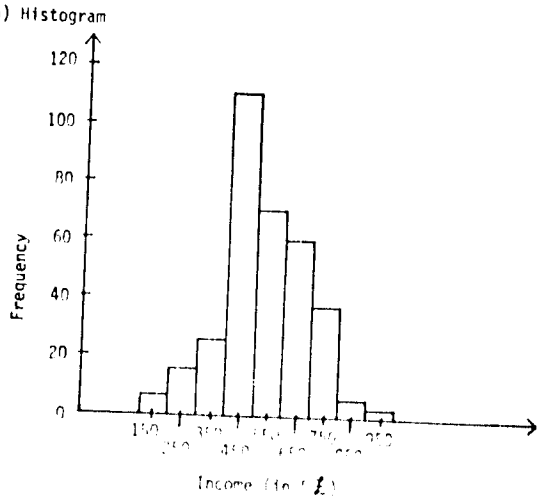
Figure 2.17:

Initial Incomes of Graduates of Post-Secondary Institutions, for the Sudan: 1975 (Survey Data)

a) Data

Class Interval	Frequency
100-199	6
200-299	16
300-399	25
400-499	109
500-599	71
600-699	59
700-799	37
800-899	6
900-999	3

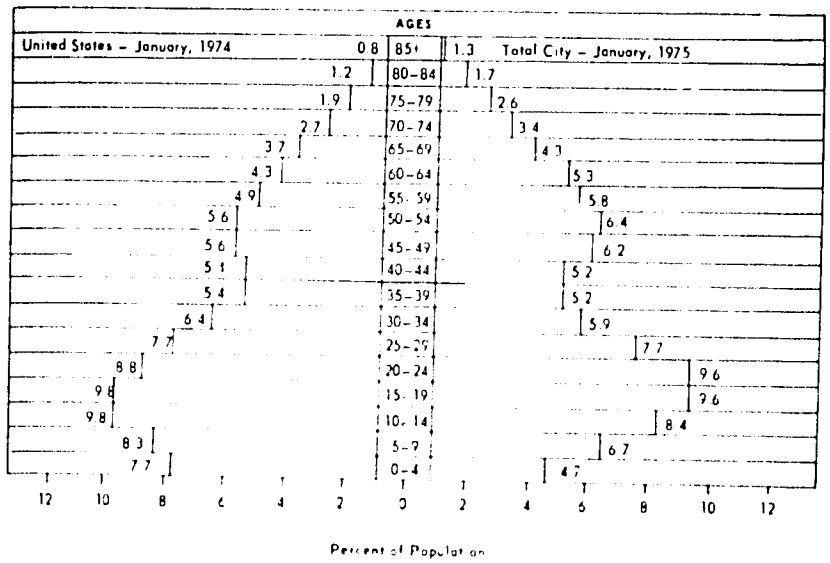
b) Histogram



Source: Bikas C. Sanyal and Jan Mersluis, "Higher Education, Human Capital and Labour Market Segmentation in the Sudan," International Labour Office, World Employment Programme Working Paper No. 10, March 1976.

Figure 2.18:

Age Structure of the Newton, Massachusetts, Population Compared with the Population of the United States



Source Vincent S. Russo, *A Study of Declining Enrollments and Long Range Alternatives for Organization of the Newton Public Schools*, Department of Research and Planning, Newton Public Schools, 1975, p. 17.

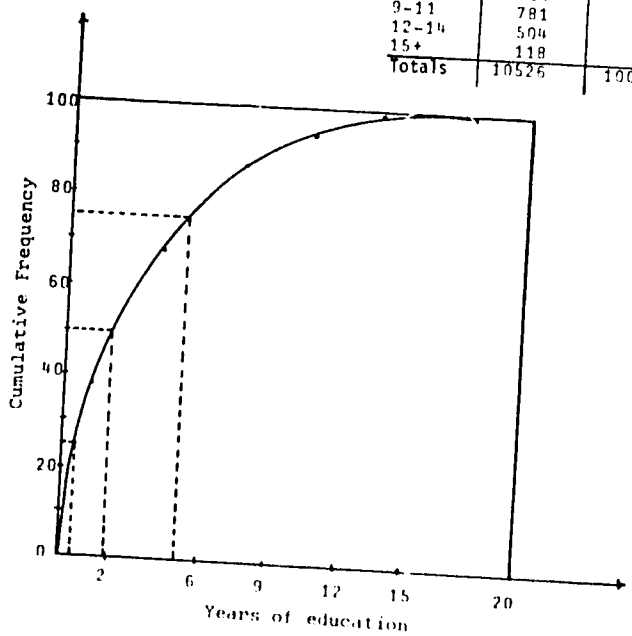
Exercise 2.2 What features of interest are evident in the histograms shown in Figures 2.15 through 2.18?

3.2 *Cumulative Frequency Graphs*

By constructing something called a cumulative frequency graph, it is possible to find values for the median and the interquartile range of data categorized in class intervals. Shown in Figure 2.19 is a cumulative frequency graph for the data used to construct the histogram of the distribution of educational attainment in El Salvador, shown previously in Figure 2.16. To construct a cumulative frequency graph, the frequency counts in each class interval must first be converted into a relative frequency (percentages). When these percentages are cumulated, the result is a cumulative frequency. For example, in Figure 2.16 the percentage of all individuals with 5 or fewer years of education is 68 = 39 + 29. When these cumulative frequencies are plotted, the diagram shown in Figure 2.19 results.

Figure 2.19:
A Cumulative Frequency Graph of the Distribution of Years of Education, El Salvador: 1975

Class Interval	Frequency	Relative Frequency	Cumulative Frequency
0-1	4054	39%	39%
2-5	3105	29	68
6-8	1964	19	87
9-11	781	7	94
12-14	504	5	99
15+	118	1	100
Totals	10526	100.0	



Source: Center for Studies in Education and Development. "An Assessment of the Relevance of Education in El Salvador." Harvard University, 1977, p. IV-14.

With this diagram it is possible to find a median value and an interquartile range. Recall that the median is that point where 50% of the values are higher and 50% of the values lower. The hinge values occur at cumulative frequencies of 25% and 75%. The interquartile range, a measure of spread, is just the difference between these two hinges. All three values are easily read off the graph. The median is determined in the following manner: find the 50% mark on the vertical axis (the cumulative frequency axis) and follow this line over to the graph; then drop a vertical line to the horizontal axis. The median value can be read off directly. In this case it is about 2 years. Similarly, the two hinges are found to be about 0.5 years and 5 years, making an estimate of the H-spread of 4.5 years.

Exercise 2.3: Construct a cumulative frequency for the histogram in Figure 2.17. Find the median and interquartile range, and comment.

3.3 *Histograms with Ordinal Data*

Histograms for ordinal data are constructed much as for those interval data, but no class intervals can be established because of the nature of ordinal data. Recall that such data are ordered in the sense that one category can be placed lower or higher on some scale than a second category, but that the distance between items on the scale is not possible to determine.

To construct a histogram for ordinal data, one again computes either the frequency or the percentage distribution for a given category. The diagram is similar to that for interval data, with the vertical axis showing either the frequency counts or the percentage values. The horizontal axis is used for the ordered categories, which may or may not be placed next to each other. Since it is not possible to establish distance between points on an ordinal scale, all the intervals are of *uniform width*. Shown in Figure 2.20 is a percentage histogram showing enrollments, by level of education, for the USSR in 1975. Note that in part (b) of the figure, the rectangles have been placed slightly apart, so as to accentuate the point that these are ordinal and not interval data.

4.0 *Displays for Nominal Data*

To this point we have discussed a variety of techniques for displaying interval and ordinal data. It is now time to turn our attention to nominal data. Recall that with nominal data there is no underlying order to the categories into which data are classified. Consequently, we cannot look for features like central tendency or the range in the distribution, because these features have no meaning with nominal data. Displays of nominal data typically summarize features such as absolute frequency (i.e., counts of things), relative frequencies (i.e. percentages), or ratios (comparisons of absolute or relative frequencies).

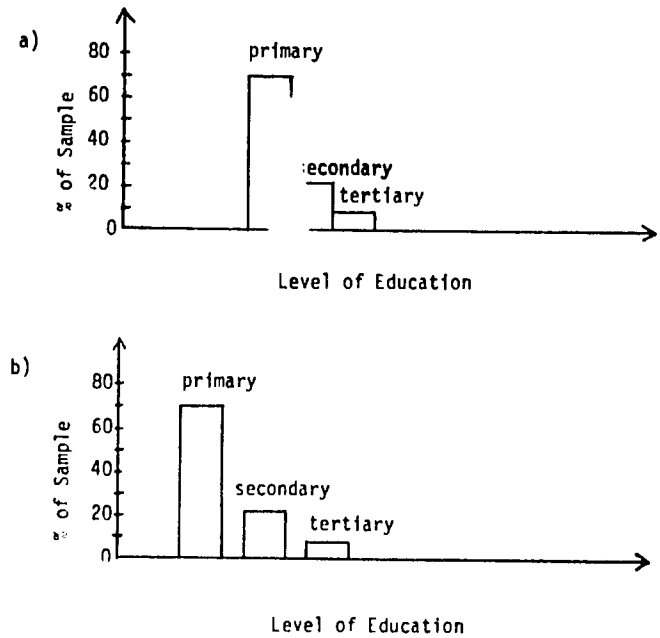
In this section we will describe the use of bar graphs to display nominal data, and will also show how various summary statistics can be computed to describe features of the display. This is a short section, based on the belief that a general explanation supplemented by several examples, will provide sufficient guidance for the use of such graphs for other purposes.

The bar graph is very similar in appearance to a histogram. It consists of a series of bars, separated by spaces, with the height of each bar proportional to the quantity being plotted. This quantity can be either a frequency count or a relative frequency, as was true with histograms.

In constructing such a bar graph two rules of classification must be consid-

Figure 2.20:

A Percentage Histogram of Enrollment by Level of Education, for USSR in 1975.



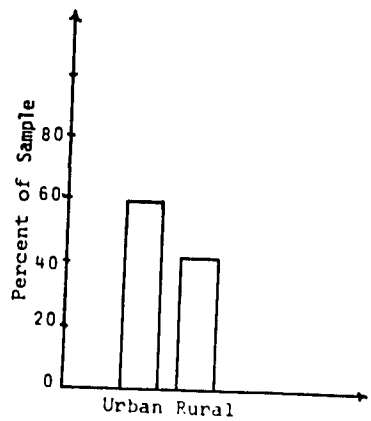
Source: Japanese National Commission for UNESCO and the National Institute for Educational Research (Japan). "Educational Developments in Asia and Oceania," Chart XVI, p. 21.

ered. First, the classification must be comprehensive enough to cover all observations involved. For example, there is no point in omitting Catholics from a graph showing frequencies of students of various religious backgrounds, unless none of the students is Catholic. The second rule is that the classification must be made according to a single criterion, that no mixed and non-exclusive categories are involved. In other words, it should not be possible to classify a single observation into more than one category.

In Figure 2.21 a simple bar graph shows the urban-rural distribution of the economically active population in El Salvador. All bars are *a unit wide*, the height in this case showing the relative frequency of those surveyed in each of the two mutually exclusive categories.

Figure 2.21:

Distribution of the Economically Active Population in El Salvador, by Urban and Rural Region: Survey Data, 1975



Classi-fication	Frequency	Relative Frequency
Urban	6213	59.0%
Rural	4313	41.0
Total	10526	100.0

Source: Center for Studies in Education and Development, "An Assessment of the Relevance of Education in El Salvador." Harvard University, 1977, p. IV-13.

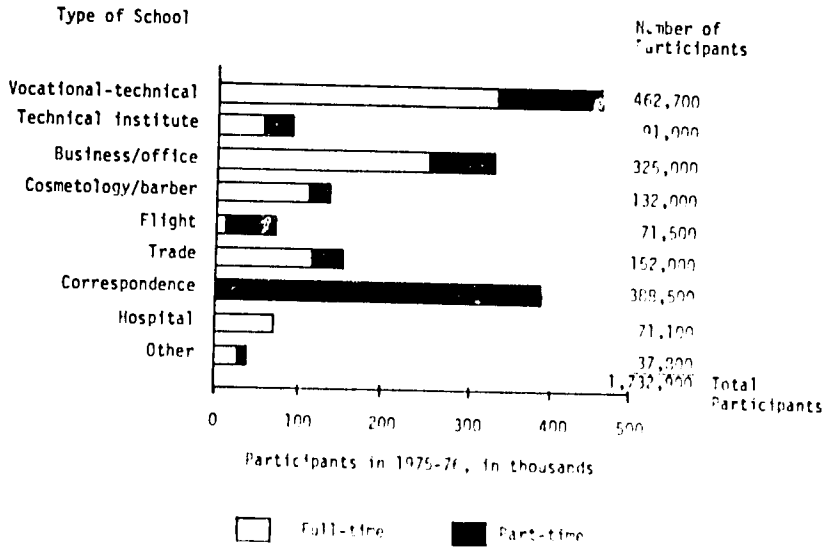
In Figure 2.22 there is another example of a bar graph used to display frequency data. In this case, enrollment in various types of postsecondary schools that offer occupational programs is displayed. Note that by using shading, the typical bar chart can be extended to include another piece of information: each bar in this figure distinguishes between full-time and part-time status of the students.

Bar graphs are typically used to display in a visual fashion information that the analyst wishes to emphasize. From the display, one can convey an indication of the total counts in each of several categories. Usually this is done for comparative purposes. One purpose is to impart visually a feeling for the absolute numbers in each category; a second purpose is to convey how these numbers compare across categories. For example, Figure 2.21 shows that most of the economically active population in El Salvador live in urban areas. This is simple information, but it is quite effectively presented in a display such as a bar graph, particularly where differences in categories are quite striking and the analyst wishes to make sure the differences are recognized by the decision makers reading the report of the analysis.

Graphs resembling frequency bar graphs are also commonly used with interval data. In these a bar is used to represent some summary number. For example, in Figure 2.23 the average charges in a variety of occupational programs are shown. It is clear from this figure that for each program, private

Figure 2.22:

Enrollments in Non-Collegiate Post-secondary Schools Offering Occupational Programs: United States, 1975-76



Source: National Center for Education Statistics. The Conditions of Education 1977. Washington, D.C.: U.S. Government Printing Office, 1977, p. 78.

schools charge more than public ones, and that for both public and private schools the range in average charges is substantial — the most expensive costing 2 to 3 times the least expensive.

Note that graphs like those in Figure 2.23 are comparable to the box plot described in an earlier section. The main intent is to illustrate a summary number describing the underlying distribution of interval values. In Figure 2.23, average charges are shown by program type. For each program type there are a number of specific programs that were used in computing the average. But it is only the average that is displayed. Missing is information on the shape of the distribution and on measures of spread. This additional information, if available, can help when we want to decide how much importance to place on apparent differences between averages or medians. If the original data are not accessible, however, the bar graph can be effectively used to illustrate such differences in central tendencies.

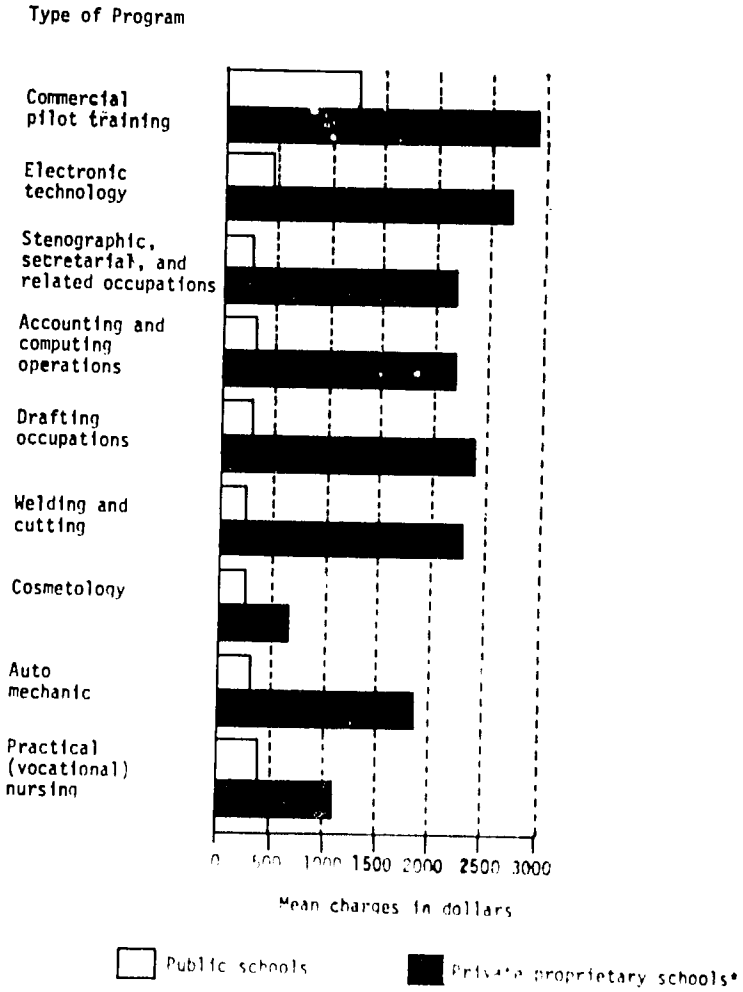
5.0 *Displays of Data Categorized in Other Ways*

5.1 *Statistical Maps for Use with Geographical Categories*

Geographical variation in teacher-pupil ratios, per pupil expenditures, achievement, or other educational inputs or outcomes are of concern to

Figure 2.23:

Charges for Occupational Programs Offered by Non-Collegiate Postsecondary Schools in the United States



*excludes independent nonprofit schools and church-affiliated schools

Source: National Center for Education Statistics. The Condition of Education 1977. Washington, D.C.: U.S. Government Printing Office, 1977, p. 41.

educators. The statistical map offers one very concise method of displaying such variation by geographical units. The main idea is to symbolize varying frequencies or levels of activity by various shadings (from light to dark, for example). Each geographic unit on a map is shaded a particular way to

represent the level or frequency of occurence in that particular area.

An important part in developing a statistical map is the definition of the geographical area. Educators are mainly interested in geographic areas that also coincide with administrative units. School systems or regional educational districts might be appropriate in some countries. For small maps, the actual districts from which single schools draw students might also serve as the geographic unit.

While there is a good deal of theoretical work developed around statistical mapping, particularly as performed with computers, the rudiments are nonetheless simple:

— A geographic area of interest is mapped with internal boundaries representing the units for which data are available for display.

— Several class intervals are established (usually four or five are sufficient for statistical mapping), so that each unit for which data exist can be classified into only a single class interval.

— Each class interval is represented by a different form of shading or design.

— The map is then filled unit by unit, each unit according to the shading or design representing the class interval in which the data for the unit fall.

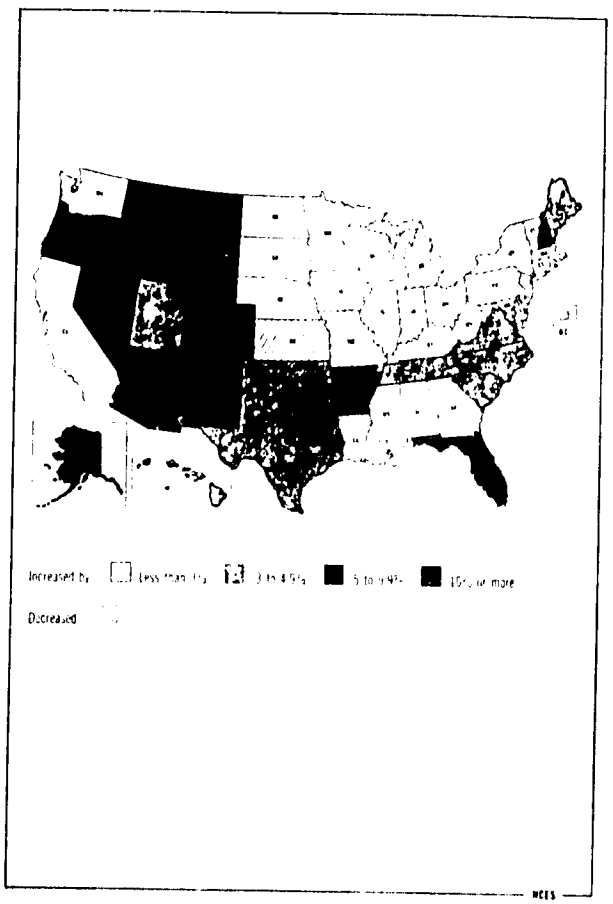
— A key is included which indicates what class interval corresponds to a particular pattern of shading.

Suppose, for example, we wished to represent, on a statistical map, the average levels of educational attainment in African countries. The geographic area of interest is Africa. The internal boundaries coincide with country boundaries, so this type of map is readily available. Our data in this example represent average educational attainment in each country. We would likely plot these figures in a stem-and-leaf diagram and then decide on appropriate class intervals. After this is done, each interval is represented by a design or shading. In the final step, each country on the map is filled in with the appropriate shading. Suppose the average level of educational attainment in country X is 6.3 years and this falls in the class interval 6.0 - 6.9, represented by a design of the following kind: This design would be used to fill in the area within country X's boundaries.

Shown in Figure 2.24 is a statistical map of the United States, showing the changes in school-age population due to migration. The advantage of statistical mappings is that geographical patterns become readily apparent. In this example, states with the largest percentage of school-age population increase caused by migration tend to be in the western part of the country. States losing school-age children because of migration are found in the middle and eastern parts of the country. The usefulness of statistical maps occurs because geographical variation in the variable of interest is made readily evident. Statistical maps could find wider use in educational planning but have not yet been fully utilized. There are several computer routines being developed which plot such maps, and it seems certain they will receive increased use.[1]

Figure 2.24:

Change in 5- to 17-Year-Old Population Due to Migration, for the United States: 1970 to 1975.



National Center for Education Statistics. The Condition of Education 1978. Washington, D.C.: U.S. Government Printing Office, 1978, p. 33.

5.2 *Displays of Time Series Data*

When data are collected on a single variable, but for different points in time, the set of data is commonly called a time series. Enrollments in a school,

collected for the past ten years, is just one example of data in a time series. Short time series, of two or three different points in time, are conveniently displayed with a histogram or bar graph. But longer time series are effectively presented in graphical form. Graphs are particularly convenient for portraying trends and, of course, can lead to questions about the source of the trends or its impact. Graphs can vary enormously in design, and no attempt will be made to discuss all varieties that might be appropriate for educational planners. Rather, examples are given which demonstrate possible variations. The utility of such graphs, in any event, is similar to that described for other displays earlier in this chapter. The intention is to display data in such a way that patterns or other features of importance are made apparent.

The mechanics of constructing a time series graph are fairly simple:

- It is assumed that data have been collected on a variable of interest for a number of time periods (i.e., years, months, days, etc.);
- The data are plotted on a coordinate system much like that used for constructing histograms. The vertical axis, also commonly called the Y-axis, is used to represent measurement levels appropriate to the variable of interest. The horizontal axis, also commonly called the X-axis, is used to mark off the time periods.
- The origin, or zero point for both axes, occurs where the horizontal and vertical axes cross.
- Each combination of data and the corresponding time period is located on the graph and indicated by a point.
- After plotting the data points, the points may be connected with straight line segments or left to stand alone, if that seems preferable in the situation.

Shown in Figure 2.25(a) is a display of the yearly births in the United States, from 1940 to 1975. With nearly universal attendance in elementary school, one would expect to see enrollments in primary grade 1 reflect birth patterns quite closely, and total elementary school enrollments to reflect birth patterns to some extent. Shown in Figure 2.25(b) are yearly enrollments in the first grade of primary school. Note that the enrollment pattern is quite similar to the trends in births, but lagged by 5 to 7 years. It is not hard to predict the trend in enrollments beyond 1976, barring unusual circumstances. Following the birth pattern, first-grade enrollments will rise for about two years and then fall again for 4 to 5 years.

Figures 2.26 through 2.27 show additional uses of time series educational data, displayed in graphical form. Note that it is possible to introduce more than one variable on a graph. For example, Figure 2.26 shows education expenditures per student by institutions of higher education. If the data are plotted for all institutions, the heavy line results. If the data are available according to the public private control of the institution, then two additional lines result. From this plot it is obvious that over the time period involved, private institutions have spent more per student than public institutions, and that the gap has widened slightly.

Figure 2.25:

Births, 1940-1975, and First Grade Enrollments, 1952-1976, for the United States

a) Births

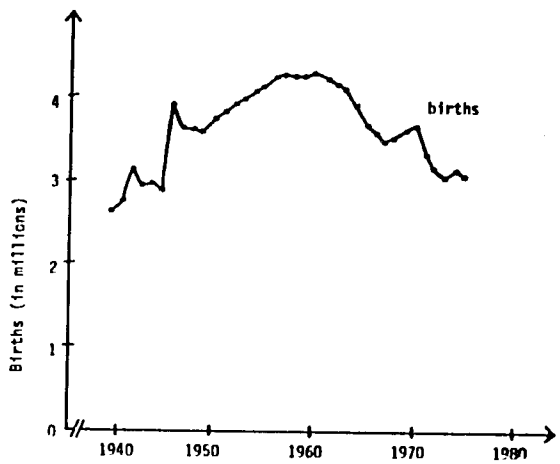
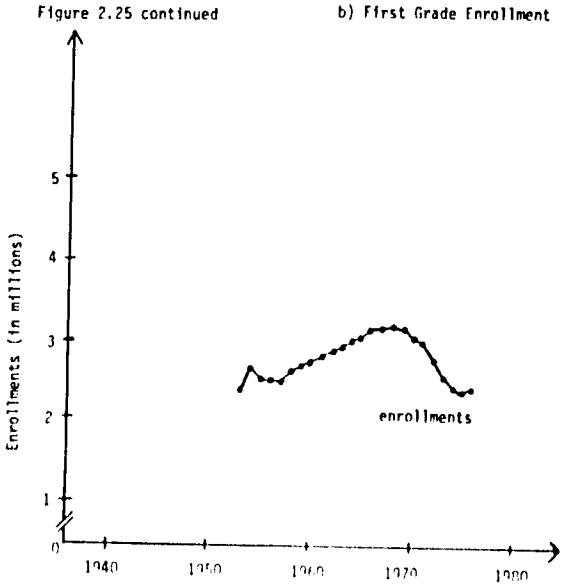


Figure 2.25 continued

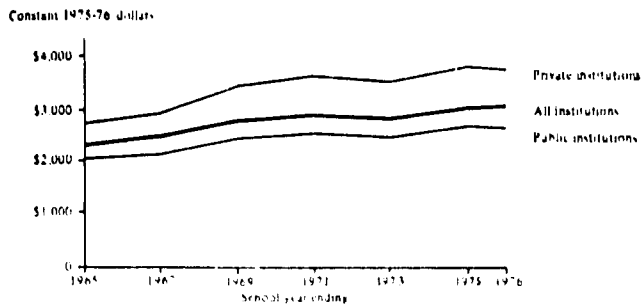
b) First Grade Enrollment



Source: U.S. Bureau of the Census, Current Population Reports, Series P-25, No. 601, "Projections of the Population of the United States, 1975 to 2050." Washington, D.C.: U.S. Government Printing Office, 1975, Table 1; National Center for Education Statistics, Digest of Educational Statistics. Washington, D.C.: U.S. Government Printing Office, various years.

Figure 2.26:

Education Expenditures per Student, by Institutions of Higher Education, for the United States: 1965–1976.



Source: National Center for Education Statistics, The Condition of Education 1977. Washington, D.C.: U.S. Government Printing Office, 1977, p. 59.

Three variable graphs can also be effectively displayed using a cumulative pattern if frequency counts are of concern. This is made clear in Figure 2.27, where Master's degrees conferred from 1900 to 1976 are also shown by sex of the recipient of the degree. Since the total number of recipients is just the sum of the males and females receiving degrees, the area between the "total" line and the "male" line represents degrees awarded to females. Clearly there has been remarkable growth in the number of degrees received recently, this being true for both sexes.

Exercise 2.4: What does the graph in Figure 2.27 tell us about the relative distribution of Master's degrees between males and females?

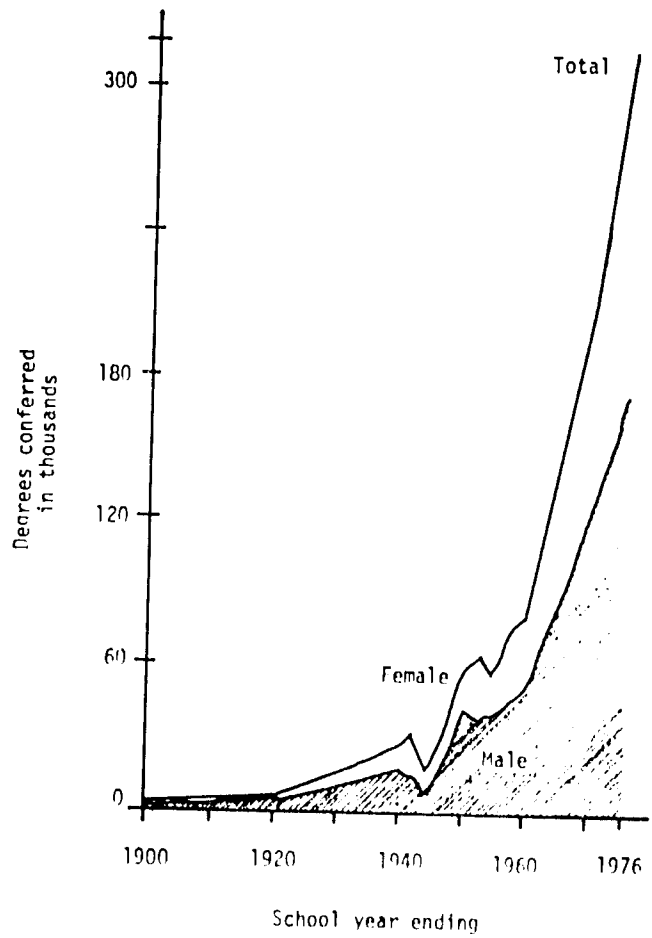
A final, and an important, use of graphs should be mentioned. The interrupted time series design with comparison series is a fairly powerful research design (see Volume III), and these kinds of data are very effectively presented in the form of a graph. Figure 2.28 shows a graph which illustrates the apparent impact of the introduction of a particular strain of wheat into Mexico in 1961. The vertical line indicates the time of the introduction of the innovation. Data for Argentina and for Chile are included as *controls*, to indicate that there was rough comparability between the three countries prior to the time of the introduction of the Semi-dwarf wheat in Mexico in 1961, but a widening gap after the introduction. Not all of this may be attributable to the particular strain of wheat, however. Note that even prior to 1961 the wheat yields in Mexico seemed to be increasing more substantially than in the other two control countries. So some of the apparent effect of the Semi-dwarf wheat may actually be attributable to a continuation of other

factors that would have continued to increase wheat yields in Mexico even without the new form of wheat.

The rationale for developing such graphs is this: controls are used for comparison purposes; the innovation is located at a particular point in time; and the trends before and after introduction of the innovation form the basis for assessment of the impact of the innovation. Without the use of the control groups the technique is less indicative, and arguments have to be based on changes in the slope and level of the single trend line involved. This is always risky.

Figure 2.27:

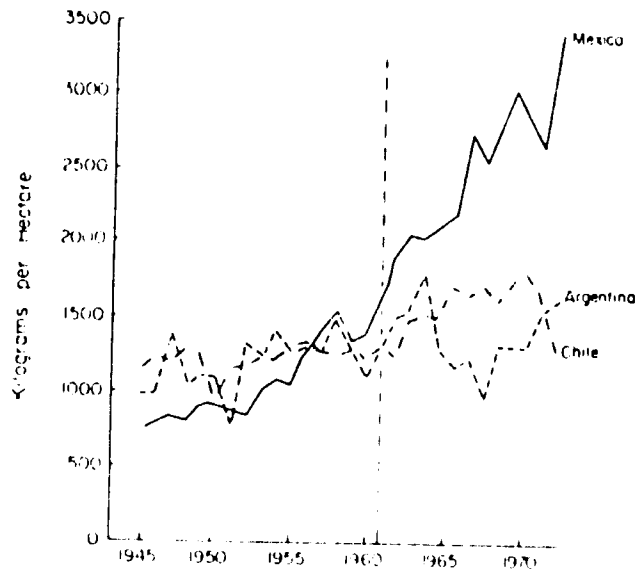
Master's Degrees Conferred, for the United States: 1900–1976



Source: National Center for Education Statistics. The Conditions of Education 1977. Washington, D.C.: U.S. Government Printing Office, 1977. p. 65.

Figure 2.28:

Introduction of Semi-Dwarf HYV Wheat in Mexico in 1961: An Illustration of the Interrupted Time Series Design With Comparison Series



Source: Donald T. Campbell, "Focal Local Indicators for Social Program Evaluation." In Marcia Guttentag, ed. Evaluation Studies Review Annual, Vol. 2 (1977), p. 134. Reprinted with permission.

6.0 *Summary*

In this chapter, a number of analysis techniques, each having a visual component, were described and illustrated with examples. The choice of technique is determined by a wide variety of factors, including the level of measurement for the variables involved, the quantity of data being used; whether the analyst has access to the original data; how the variables are categorized, with geographical and time series indicating special types of analysis; and, finally, the intention of the analyst. In all cases, however, the objective of using visual data display techniques is to make clearer the important features of the data, whether these features are measures of central tendency, the spread of the distribution of numbers, the general shape of the distribution, the pattern of geographical variation, the trend over time, or some additional feature.

In this chapter the reader was also introduced to the concept of transforming the original data, for purposes of achieving better analysis. For example, when comparing two or more groups, we found it useful to transform the original data so that the spreads (interquartile range or variance) of the data in each group was roughly comparable. Making valid comparisons, as we saw, is sometimes difficult, and requires work in addition to the computation of measures of central tendency for each group.

7.0 *Terms for Review*

stem
leaf
mean
extremes
hinges
median
variance
interquartile range
outliers
class interval
histogram
bar graph
cumulative frequency graph
statistical map
box plot
central tendency
spread
shape of a distribution
data transformation
percentage frequency
frequency count
time series

8.0 *Practice Exercises*

1. Assessing the impact of fluctuating enrollments on the cost of education was one of the principal mandates given to a small group of analysts by educational authorities in North Central. From the data available to the planners, they selected average teacher salaries as one measure of the cost of education. Shown in Figure 2.29 are average teacher salaries for 100 randomly selected school districts in the region. Construct a stem-and-leaf diagram of these values and a five-number summary table. Comment on some of the important features of this analysis. If the data were to be transformed so as to make a more symmetrical distribution, what transformations might be tried? Why would an analyst want to make such a transformation?

2. Shown in Figure 2.30 are five-number summaries of average salaries, by the size of the school district, for 100 randomly selected school districts. Make box plots of this information and compare how average salaries differ by size of the school district. Would it be useful, in making this comparison, to transform the data? If so, suggest a transformation that might work. Why is it that such transformations of data are useful?

3. Shown below are frequency counts for ordinal variables of interest to North Central analysts. Construct histograms based on this information. For the variable District Size, construct a cumulative frequency graph and roughly determine the median and H-spread. Comment on the distributions of District Size and Enrollment Trend.

Variable	Categories	Absolute frequency
District Size		
	0-299	49
	300-799	166
	800-1799	124
	1800+	97
Enrollment (% change)		
	10% or more	59
	6-9%	33
	+5% to -5%	165
	-6% to -9%	76
	-10% or below	103

4. How might a statistical map be used to advantage to illustrate average teacher salaries in the school districts of North Central?

Figure 2.29.

Average Teacher Salaries in 100 Randomly Selected School Districts in North Central

249	278	300	262	280	[illegible]	271	285	259	236
294	260	250	291	257	271	301	338	282	293
309	297	301	291	306	295	281	236	285	284
279	319	324	309	295	250	350	321	334	296
307	262	[illegible]	[illegible]	346	339	359	301	241	[illegible]
260	294	[illegible]	[illegible]	[illegible]	292	[illegible]	[illegible]	338	302
297	321	277	[illegible]	345	[illegible]	234	383	296	239
314	249	261	[illegible]	291	229	[illegible]	247	283	303
262	316	[illegible]	298	399	434	235	385	369	283
294	277	291	274	229	289	289	259	236	243

Figure 2.30:

Five-Number Summaries of Average Teacher Salaries for 100 Randomly Selected School Districts in North Central, by Enrollment Size

Size

Summaries	District Enrollment			
	0-299	300-799	800-1799	1800+
N	10	46	29	15
M	270.5	278.5	289	294
H	260 281	266 298	273.5 313	282 308.5
I	220 294	230 371	258 352	262 346

Footnotes

[1]See, for example, the computer routines offered by the Harvard Laboratory for Computer Graphics and Spatial Analysis, *Lab-Log*, Cambridge, Mass., 1978.

Bibliography

Loether, Herman J., and McTavish, Donald G. *Descriptive Statistics for Sociologists*. Boston: Allyn and Bacon, 1974. This text contains a useful chapter on visual displays, including histograms, bar graphs, statistical maps, and other displays.

McNeil, Donald R. *Interactive Data Analysis*. New York: John Wiley, 1977. There is a good discussion on the rationale behind data transformation and its role in comparison between groups. In addition, there are computer routines in FORTRAN and APL, for those with access to computers. Routines are included for stem-and-leaf diagrams and box plots.

Tukey, John W. *Exploratory Data Analysis*. Reading, Mass.: Addison-Wesley, 1977. Tukey provides an extensive discussion of stem-and-leaf diagrams, box plots, and data transformations.

CHAPTER 3

Tables of Counts

Gary Lewis

1.0 *What's Ahead?*

Table 3.1 presents information about second and third-year students in the Botswana Brigades, which is an education program designed in part to provide basic vocational and technical training. In the table there is information on the age and prior work experience of the Brigade students surveyed. The table shows counts of students with certain characteristics. We find, for example, that there are 72 students who are younger than 20 and who have had no work experience prior to the Brigades. Tables of counts, such as Table 3.1, are also commonly called *contingency tables*, and are frequently and easily constructed using data common in educational planning situations.

Table 3.1

A Table of Counts, Showing Age and Prior Work Experience for a Sample of Students in the Botswana Brigades, 1978

Prior Work Experience	Age		Totals
	<20	20+	
Experience	12	46	58
No experience	72	69	141
Totals	84	115	199

Source: Herbert Mucherwa Murerwa, "The Planning of Non-Formal Education for Rural Youth Employment: An Analysis of Training in the Botswana Brigades." USAID/Harvard Planning Project Report; later Ed.D. Thesis, Harvard Graduate School of Education, [illegible], p. [illegible].

Using Table 3.1, we can ask if there is an association between the age of the student and work experience before entering the Brigades. We expect that older students would have more work experience than younger students.

In asking such questions, we often differentiate between independent and dependent variables (review these terms in Chapter 1 if necessary). In the present case, age is the independent variable and work experience the dependent variable, because we choose to think that the age of individuals will partially determine the amount of work experience, rather than the reverse.

A table by itself may be sufficient to answer questions in obvious cases, but often other things must be done to assist in the analysis. The primary intention of this chapter is to describe and illustrate techniques that can be used to increase the amount of information available from tables of counts like that shown in Table 3.1. In the process we will be interested in the following things:

— Defining in more precise terms what the association of variables means;

— Determining whether two variables, such as age and work experience, are associated.

— If there is an association between two variables, can we assess whether it is weak or strong?

— Is it possible to obtain additional information about the kind of relationship existing, in terms of the way in which one variable is related to a second variable?

As with past chapters, practice problems use the North Central data set to provide some familiarity with the techniques and approaches introduced.

2.0 *Independence and Association*

2.1 *Independence*

Using Table 3.1, we might want to know if the age of students and their prior work experience were related. It was speculated that older students would generally have more work experience than younger students, that in fact the two variables would be associated. What precisely does it mean to say that two variables are associated?

To answer this, we first must understand what is meant by the *independence* of two variables, because association turns out to be defined as the lack of independence. The statistical notion of independence is close to that in common language usage. One might say, for example, that the height of an individual is independent of the month in which he or she was born. Stated another way, this means that knowing the birth month of an individual will not change the likelihood that one can predict the individual's height. In general, independence between two variables implies that having information about one variable, for a number of people or other units, would not be of any assistance in predicting values for the second variable.

In terms of tabled values, what does independence mean? Table 3.2 shows some hypothetical data on height and month of birth for a group of 20 people. But before addressing the question of independence, some terminology should be introduced because it is commonly used with contingency tables.

Note that each variable in Table 3.2 has two categories. Thus the table turns out to be a 2 × 2 or "2 by 2" table, and is read as the number of rows by the number of columns. The numbers around the table perimeter give the counts of people or things in the various categories of each variable. For example, there are 8 people who are short and 12 who are tall; there are 5 people born between January and June, and 15 born between July and December. The total number of people, 20, is called the *total N*. The values around the table perimeter are called the *margin als*, because they are found at the margins of the table. The numbers inside the table are each said to be in one of the *cells* of the table. For example, there are 9 people in the lower right-hand cell, indicating there are 9 people who are both tall and born between July and December.

Table 3.2:

A Table of Counts Illustrating the Independence of Height and Month of Birth, for a Hypothetical Sample of People

a) Table of counts

Height	Birth Month		Totals
	Jan-June	July-Dec	
Tall	3	9	12
Short	2	6	8
Totals	5	15	20

b) Percentage table

Height	Birth Month		Totals
	Jan-June	July-Dec	
Tall	60%	60%	60%
Short	40%	40%	40%
Totals	100%	100%	100%

Height and birth month are independent if having information about the month of birth does not assist us in predicting height (or vice versa). From Table 3.2(a) we know that 8 out of the 20 people, or 40%, are short. The question of independence then comes down to this: does this percentage change if we know something about the birth month of the individual.

Table 3.2(b) shows the data converted to percentages, where the percentages total to 100% along the columns. Note that the columns are identical to each other and to the marginal column. Forty percent of people born in January to June are short, just as 40% born in July to December are short, and just as 40% of all 20 people are short. In other words, knowing the birth month of the 20 individuals would not help us predict their height (i.e., whether they were short or tall).

If we had the original data, we could construct a game in which each person's height (as short or tall) was written on a piece of paper and put into a hat. Then we could draw out these pieces of paper, one at a time without looking at them, and guess whether the person was short or tall. Presuming we went with the odds each time, we would always guess they were tall, and would end up being right 60% of the time and wrong 40% of the time. We would make this same percentage of errors, even if we were first told whether the person was born in January to June or July to December. This is what is meant when two variables are described as being independent.

One very simple method, then, for determining whether two variables are independent, as represented in the table, is to compute percentages. These are normally computed so that the percentages sum to 100% for the categories of the independent variable. In the present example, with birth month and height, we thought of birth month as the independent variable that might help us predict the dependent variable, height. Birth month was tabled as the columns, and percentages were calculated so that they would sum to 100% for each category of birth month. This permits us to make comparisons across categories of the dependent variable. In Table 3.2 these comparisons showed that the percentages were identical, and hence the variables are independent.

2.2 *Association*

Two variables are said to be *associated* if they are not independent. What would a table look like if the two variables were associated? Tables 3.3(a) and (b) show additional hypothetical data for height and birth month, in which there is an association between the variables. The marginals are the same as in the previous data in Table 3.2. For example, 40% of the people are short, as before. But the cell values in Table 3.3 differ from those in Table 3.2, meaning that the cell percentages will differ as well. The percentaged table is shown in Table 3.3(b). Note what would happen now if we played the game with the names in the hat. If we knew nothing about the birth month, we would still guess tall for everybody, and end up making mistakes 40% of the time (this assumes rational behavior, betting on the odds and not guessing with hunches). On the other hand, if we were first told the birth month of the

individual, we could do considerably better than this. If we are told that the person is born between January and June, then we would guess that the person is short. Since 80% of these people are short, we would make mistakes only 20% of the time. Similarly, if we were told that the month of birth was between July and December, our guess would be tall, and we would be wrong only 26.7% of the time. In other words, knowing something about the month of birth is a considerable help in predicting the height of the individuals in Table 3.3. In such cases, the variables are not independent, but rather are said to be associated.

Table 3.3:

A Table of Counts Illustrating Association of Height and Month of Birth, for a Hypothetical Sample of People

a) Table of counts

Height	Birth Month		Totals
	Jan-June	July-Dec	
Tall	1	11	12
Short	4	4	8
Totals	5	15	20

b) Percentage table

Height	Birth Month		Totals
	Jan-June	July-Dec	
Tall	20%	73.3%	60%
Short	80%	26.7%	40%
Totals	100%	100%	100%

Table 3.4 illustrates an additional situation where association exists. In Table 3.4, however, the association is said to be *perfect* because all observations occur only in selected cells. In this case, all boys are enrolled in the

program A and all girls in the program B. There are no boys in program B and no girls in program A, for reasons unknown.

Table 3.4:

A Table of Counts Illustrating Perfect Association Between Type of Program and Sex of the Individual, for a Hypothetical Sample of People

a) Table of counts

Sex	Program Type		Totals
	A	B	
Male	10	0	10
Female	0	20	20
Totals	10	20	30

b) Percentage table

Sex	Program Type		Totals
	A	B	
Male	100%	0%	33%
Female	0%	100%	67%
Totals	100%	100%	100%

In Table 3.5 another important aspect of association is demonstrated. When the variables are both ordinal in nature, we can speak of the *direction* of the association. Comparing across the percentaged table, for Low Earnings, we see that low levels of educational attainment tend to go with low levels of earnings. Similarly, high levels of educational attainment tend to go with high levels of earnings (62.5% of those with high educational attainment have high earnings, while only 16.7% of those with low educational attainment have high earnings). When low values of one variable generally go with low values of the second variable, and high values of the first

generally go with high values of the second, the association is said to be in a *positive* direction. When low values of one variable generally go with high values of the second variable, the association is in the *negative* direction.

Table 3.5:

A Table of Counts Illustrating the Direction of the Association Between Educational Attainment and Earnings, for a Hypothetical Sample of People

a) Table of counts, showing low educational attainments generally occurring with low earnings, and high educational attainments generally occurring with high earnings.

Earnings	Educational Attainment		Totals
	Low	High	
High	10	25	35
Low	50	15	65
Totals	60	40	100

b) Percentage table

Earnings	Educational Attainment		Totals
	Low	High	
High	16.7%	62.5%	35%
Low	83.3%	37.5%	65%
Totals	100%	100%	100%

In summary, then, there is a very simple procedure for determining if two variables are independent or associated, and for determining the direction of the association if the variables are ordinal. All one needs to do is compute percentages in the direction of the independent variable. Shown in Table 3.6 is the information for the Botswana Brigades again, but this time also tabled by percentages. It is easy to see that age and experience are not independent, that in fact they are associated. Furthermore, we see that a younger age tends to go with no experience. This is seen by comparing across the columns, along the row for "No Experience": 85.7% of those younger than 20 have had no prior work experience, while only 60% of those older than 20 have had no prior experience.

One final note should perhaps be sounded. Note that association does not mean that there is necessarily a causal relationship between the variables, although this is certainly one circumstance yielding association. Two variables can be associated, with no causality involved, if there is a third variable related to each. If, for example, a person's age affects both his income (through seniority) and his physical strength, then strength and income will be associated even though there is no causal mechanism operating to pay strong people differently from weaker people.

Table 3.6:

A Percentaged Table Showing Age and Prior Work Experience for a Sample of Students in the Botswana Brigades. Data from Table 3.1.

Prior Work Experience	Age		Totals
	<20	20+	
Experience	14.3%	40%	29.1%
No experience	85.7%	60%	70.9%
Totals	100%	100%	100%

2.3 *The Chi square Measure of Independence*

Computing and comparing cell percentages is a simple method of determining if two variables, as represented by the tabled data, are independent. In this section we will discuss one additional method for determining independence, involving the computation of a number called the Chi-square statistic.

The Chi-square number is useful for two reasons. First, it can be used to assess the likelihood of independence in a population when in fact the data are only available for a sample. In the previous examples in this chapter we have implicitly assumed that we were dealing with populations, that the data represented the information on all relevant units. When this is the case, or when the analyst is only interested in results for the sample, the previous method of computing percentages is excellent for determining independence or lack of it. However, where sample data are being used to make inferences about the entire population, things are more complicated. It may be, for example, that a random sample of units has been drawn for analysis, but that decisions will be made which pertain to the entire group of units from which the sample was drawn. In such cases we want to have some assurance that what seems to be true for the sample is also true for the entire population.

Unfortunately, there is no way to deal with this problem in any absolute sense. It is not possible to say definitely that the results from the sample data will also hold for the population data. We can only speak in terms of the likelihood of this occurring. If our sample data meet certain assumptions, to be discussed shortly, then it is possible to indicate the likelihood that we could obtain certain results with a given sample when in fact the population data would give much different results. Recall from Chapter 1 that this problem arises because, when drawing a sample, there is a possibility that the units sampled will not be representative of all the units in the population. So, in short, the Chi-square statistic can serve as an aid for the planner when dealing with sampled data.

Chi-square is also of help in computing other measures which will indicate the strength of an association between the two variables. If our tests for independence indicate that two variables are associated, we would like to have some indication of how strongly associated. Is it a very mild association, almost to the point of indicating independence, or is it in fact a moderate or strong association? The Chi-square statistic cannot by itself be used to indicate the strength of association, for reasons to be discussed shortly, but it can be used in other formulas to compute indicators of the strength of association.

In computing the Chi-square measure, a comparison is made between the cell values expected if the variables are independent and the actual cell values observed. The comparisons for all cells are added together to get a single number called the *Chi-square statistic*. The calculation is done in the following way: For each cell we calculate the following number: (actual − expected)²/expected. Once the number has been calculated for each cell, all these numbers are added up to give the Chi-square statistic. The entire formula for computing Chi-square is thus:

$$\text{Chi-square} = \sum_{i=1}^{K} \frac{(\text{actual} - \text{expected})^2}{\text{expected}}, \qquad (3.1)$$

where the sigma (Σ) indicates summation over the K cells in the table.

How is it possible to know what values to expect if the two variables are independent? If we are given the marginal values in a table such as that

shown in Figure 3.1(a), what cell values would make these variables independent? This is easily answered if we return to the earlier example in Table 3.2(a) and (b). There we saw that independence occurred when the percentaged cell values were identical across any given row (for a table with percentages computed so they sum to 100% in the columns). So it is possible to determine cell values under an assumption of independence by first computing the percentaged figures for the marginals, as in Figure 3.1(b). We know that if the variables are to be independent, the number in cells marked *a* and *b* in Figure 3.1(b) must be 40% of their respective column totals. Similarly, the cells marked *c* and *d* must be 60% of their respective column totals. So if the two variables are independent, the value in cell *a* must be 40% of 50, or 20.

It is possible to compute cell values under the assumption of independence for each of the other cells in exactly the sample manner. For cell *d*, for example, we would have (.60)(30) = 18. Note, however, that there is a faster method for filling in the cell values, as shown in Figure 3-1(d). Here we make use of the fact that cell values must add together to give the appropriate marginal numbers. So if we know that cell *a* is 20, with independence, then cell *b* must be 12 because cell *a* and cell *b* must add together to give the marginal of 32. Similarly, cell *c* must be 30 and cell *d* must be 18. Once any cell value is obtained for a 2 × 2 table, it is possible to fill in all the rest of the cell values simply by using the marginals and the single-cell value to determine what the others must be.

Figure 3.1:

Computation of Cell Values for a Table of Counts, under the Assumption of Independence

a) Marginal values

Earnings	Educational Attainment		Totals
	Low	High	
High	a	b	32
Low	c	d	48
Totals	50	30	80

Figure 3.1:
continued

b) Percentaged marginal values

Earnings	Educational Attainment		Totals
	Low	High	
High	a	b	40%
Low	c	d	60%
Totals	100%	100%	100%

c) Computation of first cell value

Earnings	Educational Attainment		Totals
	Low	High	
High	.40(50)=20		32
Low			48
Totals	50	30	80

d) Other cells filled in

Earnings	Educational Attainment		Totals
	Low	High	
High	20	12	32
Low	30	18	48
Totals	50	30	80

If the table is larger than a 2 × 2, then more than a single cell value is required before the rest can be easily determined. In general, if there are R rows and C columns, it will take (R-1) (C-1) cell values before the rest can be automatically determined. So if we have a table with 2 rows and 3 columns, a 2 × 3 table, we would need a minimum of (2-1) (3-1) = (1) (2) = 2 cell values before the rest can be determined.

Exercise 3.1: Shown below is a table for some hypothetical data. Compute the cell values under the assumption of independence.

Variable B	Variable B			Totals
	Low	Medium	High	
High				100
Low				50
Totals	48	72	30	150

Once the expected cell values are computed, it is a simple matter to calculate Chi-square. In Figure 3.1 we computed expected cell values for a table of Earnings and Educational Attainment, under the assumption of independence. But suppose that the values that we actually found with our data were those in Figure 3.2(a). We see that the expected values and the actual values are different. Hence the variables as represented in Figure 3.2 are associated. This will also be indicated by a Chi-square value which is positive, as shown in Figure 3.2(c).

Exercise 3.2: Compute the Chi-square value for the data on the Botswana Brigades in Table 3.1.

Chi-square values that are close to zero indicate that there is near independence. But unfortunately Chi-square itself cannot be used to assess the degree of association between variables. This is true because the statistic has no fixed bounds. The magnitude of Chi-square depends partly on the total N (the total number of items in the table) and partly on the number of rows and columns in the table. Because of this, Chi-square can only be used to assess whether two variables are independent. If they are not, additional measures described in the next section must be used to determine the strength of the association.

Figure 3.2:

Computation of Chi-Square for Some Hypothetical Data on Educational Attainment and Earnings

a) Actual data

Earnings	Educational Attainment		Totals
	Low	High	
High	10	22	32
Low	40	8	48
Totals	50	30	80

b) Expected data from Figure 3 1(d)

Earnings	Educational Attainment		Totals
	Low	High	
High	20	12	32
Low	30	18	48
Totals	50	30	80

c) Computation of Chi-square

$$\text{Chi-square} = \frac{(40-30)^2}{30} + \frac{(8-18)^2}{18} + \frac{(10-20)^2}{20} + \frac{(22-12)^2}{12}$$

$$= 3.33 + 5.56 + 5.00 + 8.33$$

$$= 22.22$$

Note, too, that care has been taken in the wording of the uses of Chi-square. If we are using population data, there is no hesitation in interpreting Chi-square, because all units are tabled. If Chi-square is not zero, the variables are associated. However, if the data are a sample drawn from some relevant population, it might be possible to draw a sample in such a way that the Chi-square value computed from the sampled data would be different than the value computed from the entire data set. For example, we might find a small positive Chi-square when in fact there was no association between the variables for the entire population. Determining whether or not a particular Chi-square value is sufficiently large for us to believe that an association exists is one of the concerns of inferential statistics. In such cases the analyst wishes to make inferences about the population from the sample data and is concerned about what degree of confidence to place in the sampled results.

In some educational decisions, knowledge of the situation and an assessment of the closeness of actual and expected cell values may be sufficient to determine whether the variables are associated. However, the values in Table 3.7 can provide some guidance. *If the sample has been randomly selected and if the total N is large*, a Chi-square value in excess of the tabled values can generally be considered large enough to conclude that there is association between the variables, even at the population level. What is meant by a "large total N" is open to some conjecture, but a rule of thumb is sufficient: a sample is considered large for 2×2 tables when the expected values in each cell exceeds 10. For larger tables the expected values in each cell should exceed 5. If, in fact, the assumptions about the sample are not met, particularly if the sample was not randomly selected, then little confidence can be placed in the Chi-square value.

In Table 3.7, R stands for the number of rows, C for the number of columns. To use the table, one must first compute (R-1) (C-1), and then find the appropriate tabled value. For a 2×3 table (R-1) (C-1) = 2, and we find that any Chi-square value of 6 or higher must be considered large. In such cases it is likely that even for the population the two variables in question are associated.

Exercise 3.3: Assume that the data in Table 3.1 for the Botswana Brigades are for a randomly drawn sample of all students in the Brigades. Is the value for Chi-square computed in Exercise 3.2 sufficiently large so that we could assume with a fair degree of confidence that age and prior work experience are associated for the entire population of Brigade students?

It was said previously that we could never be sure that the results for our sampled data would also hold for the population data. If we use Table 3.7 to determine whether or not we have a large Chi-square value, just how confident can we be that our results are not some chance occurrence and quite different from the population results? To answer this question, assume that we had a population for which we had complete information on two variables

Table 3.7:

Chi-square Values Large Enough to Assume an Association, by Table Size

Table Size (R-1)(C-1)[a]	Values Considered Large Enough to Assume Association
1	4.0
2	6.0
3	8.0
4	9.5
5	11.0
6	12.5
7	14.0
8	15.5
9	17.0
10	18.5

[a] R = number of rows; C = number of columns.

for all the units. Further assume that for the population the two variables are independent. We can randomly select a sample of this population, table the data, and compute a Chi-square value. If, in fact, we repeated this experiment many times, randomly drawing a sample and computing a Chi-square value, we would find that in only 5% of the samples would there be a Chi-square value larger than those shown in Table 3.7. If we have a 2 × 2 table, for example, where (R-1)(C-1) = 1, only 5% of our samples would have a Chi-square value larger than 4.0, when in fact the Chi-square for the entire population was 0.0. In other words, *if* the assumptions necessary to use Table 3.7 are met, and we compute a Chi-square value larger than the appropriate value in Table 3.7, we will make the wrong inference about the population only 5% of the time. About 5 times out of 100 we will assume there is association between the variables when in fact there is not

3.0 *Measures of Association*

Several measures have been developed to assess the strength of the association between two variables. All the measures developed have restrictions to their usefulness, and hence may be appropriate in some circumstances but not in others. Four measures of general use will be discussed in this section, with an indication of the restrictions that apply to each.

3.1 *Cramer's V.*

The Cramer's V statistic is based directly on Chi-square. Recall that in the previous section it was stated that Chi-square was not appropriate for measuring the degree of association because it was not bounded, but rather depended on the total N and the number of rows and columns in the table. Hence comparison of two Chi-square values based on different total N's would not be meaningful. Cramer's V alleviates this drawback in Chi-square by taking account of both N and the table size. The formula for Cramer's V is given as:

$$\text{Cramer's V} \quad \left[\frac{\text{Chi-square}}{\text{N} \cdot (\text{R-1 or C-1, whichever is smaller})} \right]^{\frac{1}{2}} \qquad (3.2)$$

Cramer's V runs from a minimum of zero when the variables are independent, to a maximum of 1 when there is perfect association, as in Table 3.4. This statistic, with bounds of zero and one, gives a feeling for the degree of association between the variables, but not for the direction of the association. For ordinal data we want to know if high levels of one variable tend to go with high or low levels of the second variable. Thus Cramer's V is typically used for tables where the variables are nominal level.

Exercise 3.4: Use the Chi-square value computed in Exercise 3.2 to compute Cramer's V for the Botswana Brigades data.

3.2 *Yule's Q*

If a table is 2 · 2, a very simple measure of association called Yule's Q can be computed. Shown in F ure 3.3 is a general 2 · 2 table, where the cell values are represented i To compute Yule's Q, the following formula is used, where the syn hose given in Figure 3.3.

$$\text{Yule's Q} = \frac{bc - ad}{bc + ad} \qquad (3.3)$$

Figure 3.3:

An Illustration of the Computational Formula for Yule's Q[a]

Variable Y	Variable X		Totals
	Low	High	
High	a	b	a + b
Low	c	d	c + d
Totals	a + c	b + d	a+b+c+d

$$^{a}\text{Yule's } Q = \frac{bc - ad}{bc + ad}$$

Values for Yule's Q range from − 1.0, when there is perfect negative association, to + 1.0, when there is perfect positive association. When Yule's Q is zero, the variables are independent. Note, however, that both Tables 3.8(a) and (b) would produce Q values of + 1.0, indicating perfect association, even though the association in Table 3.8(a) is stronger than that in (b) in the sense that there are no off-diagonal elements in (a).

One nice advantage of Yule's Q is that it does indicate the direction of the association. If tables are set up as in Table 3.8, with low to high values extending from left to right across columns, and from bottom to top across rows, then a positive Q value means that low values of one variable tend to go with low values of the second variable, and similarly for high values. A negative Q value indicates that low values of one variable go with the high values of the second variable. But it must be emphasized that the meaning of the sign, either positive or negative, can only be interpreted from the table and how the rows and columns are laid out. The layout in Tables 3.8 is similar to the graphical techniques discussed in Chapter 2, where low to high values extend from left to right along the horizontal axis and from bottom to top along the vertical axis. Standardizing on this layout for tables will insure that interpretation of the sign on Yule's Q and other directional measures of association does not become confused.

Exercise 3.5: Compute Yule's Q for the Botswana Brigade data in Table 3.1. What does it indicate about the strength and direction of the relationship between age and prior work experience?

Table 3.8:

Examples of Perfect Association, as Measured by a Value of Yule's Q Equal to +1.0, for Hypothetical Data

a) No off-diagonal elements

Variable Y	Variable X		Totals
	Low	High	
High	0	40	40
Low	30	0	30
Totals	30	40	70

$$\text{Yule's } Q = \frac{(30)(40) - (0)(0)}{(30)(40) + (0)(0)} = +1.0$$

b) One off-diagonal element

Variable Y	Variable X		Totals
	Low	High	
High	0	40	40
Low	30	10	40
Totals	30	50	80

$$\text{Yule's } Q = \frac{(30)(40) - (0)(10)}{(30)(40) + (0)(10)} = +1.0$$

3.3 *Gamma*

While Yule's Q statistic is only useful for 2×2 tables, the Gamma statistic can be used for larger tables. In fact, it can be shown that for 2×2 tables, Gamma and Yule's Q are identical, so one may consider Gamma as an extension of Q to tables larger than 2×2. However, one must make the additional assumption that the variables are ordinal in nature. For purposes of interpretation of Gamma, however, any variable with only two categories can be considered ordinal. For example, sex is not an ordinal variable. But it is possible and meaningful to compute Gamma for tables where sex, or any other two-category nominal variable is present, because directionality in this case indicates which group tends to have the characteristics represented by the second variable in the table.

As already seen, hypotheses with ordinal data normally take the form, "the more of this, the more of that." The hypothesis for the Botswana data was that the older the students, the more likely it was that they would have work experience. As with Yule's Q, the sign of Gamma will indicate directionality. The magnitude of Gamma will indicate the strength of the relationship. In this section the computation of Gamma will be described in some detail, because it is somewhat more complex a calculation than the previous measures of association discussed.

Suppose, for simplicity, that we have only 4 individuals, and that we give these individuals the names a, b, c, and d. For each of the 4 individuals suppose we have observations on two characteristics of the individuals. Call the characteristics variables X and Y, and let each characteristic be composed of two ordered categories (low and high). The hypothetical data collected are given in the contingency table in Figure 3.4(a).

Conventional measures of association for ordinal data are not calculated directly from the data as collected in a contingency table, but rather from a table constructed from this contingency table. The constructed table is based upon pairs of observations, so that instead of the data for the 4 people, we will be calculating a measure based upon all possible pairs of observations. The principal idea is to form pairs of individuals and then compare them on the two characteristics, ranking one individual "less than," "more than," or "the same as" the second individual on each of the characteristics. The rationale for doing this is to use the ordered state of the data to make assessments of statements such as "the more of X, the more of Y." If we allow each individual to form a pair with himself (for completeness) there will be a total of $4 \times 4 = 16$ pairs. See Figure 3.4(b) for a listing of these pairs.

The manner in which ordinal measures of association are calculated is quite straightforward, but tedious for anything but small tables. Consequently, this simple example will be worked out in its entirety and then some general rules will be presented that can be applied to any size table involving any level of counts. Normally, these computations would be done by a computer, but it is a simple 10 minute problem with a hand calculator.

The data in Figure 3.4(a) indicate that two individuals (a and b) are low on both characteristics X and Y. Individual c is low on X and high on Y, while individual d is high on both X and Y. In Figure 3.4(b), the characteristics are

Figure 3.4:

Computation of the Paired Comparison Form for Gamma

Suppose we are given the following 2x2 contingency table (the letters in the cells represent names for the items counted in those cells)

Variable Y	Variable X		Totals
	Low	High	
High	1(c)	1(d)	2
Low	2(a,b)	0	2
Totals	3	1	4

b) Pairs and comparison

Pairs	Comparisons on:	
	X	Y
a,a	Same	Same
a,b	Same	Same
a,c	Same	Less
a,d	Less	Less
b,a	Same	Same
b,b	Same	Same
b,c	Same	Less
b,d	Less	Less
c,a	Same	More
c,b	Same	More
c,c	Same	Same
c,d	Less	Same
d,a	More	More
d,b	More	More
d,c	More	Same
d,d	Same	Same

Figure 3.4:
continued

c) **Paired form**

Variable Y	Variable X			Totals
	Less	Same	More	
More	0	2	2	4
Same	1	6	1	8
Less	2	2	0	4
Totals	3	10	3	16

d) Calculations of Gamma

$$\text{Gamma} = \frac{P-Q}{P+Q}$$

where P = number of pairs in which the order on one variable is the same as the order on the other, i. e., concordant pairs[a]

Q = number of pairs in which the order on one variable is the opposite of the order on the other, i.e., discordant pairs[a]

Above, P = 2 + 2 = 4

Q = 0 + 0 = 0

$$\text{Gamma} = \frac{4-0}{4+0} = 1.0$$

[a] In calculating Gamma, ties (anything with a "Same" in it) is not counted. Other measures of association treat ties differently. For a discussion, see Hildebrand, Laing and Rosenthal (1977).

compared for each of the possible 16 pairs. Take, for example, the pair listed as *c, b*. We must ask ourselves how individual *c* compares with individual *b* on the two characteristics X and Y. If they are in identical categories, we call it a tie and mark it as the "same." If individual *c* has a higher value than individual *b* for the variable being considered, we write *more*. Similarly, if individual *c* has a lower value than *b*, we write *less*. In Figure 3.4(b) it is shown that individuals *c* and *b* have the same value for variable X (they are

both low), but that individual *c* has a higher value on variable Y. Note that the order of comparison is important. Here we are comparing *c* with *b*, not the other way around. In fact, there is a (*b*, *c*) comparison too, and one can see from the Figure 3.4(b) that it reads (same, less), just as it should.

Once this list of comparisons has been completed it is possible to construct a 3 × 3 table, as in Figure 3.4(c). The cell values for this table are obtained directly from the list of comparisons in Figure 3.4(b). For example, there are two comparisons on the list which read (less, less), these being for (*a*, *d*) and (*b*, *d*). There is one (less, same), for paired comparison (*c*, *d*); and so on until all cells are filled.

Once this 3 × 3 table has been constructed, it is possible to calculate a variety of statistics that measure the degree of association. Differences among the measures revolve around how they treat ties (i.e., what they do with the categories called *same*). One of the simplest measures, and one often used, neglects ties altogether. This measure is called *Gamma*, and it is given by the following equation:

$$\text{Gamma} = \frac{P - Q}{P + Q}, \tag{3.3}$$

where P is the number of pairs for which there is agreement (i.e., it is the number of less, less or more, more pairs), while Q is the number of pairs for which there is disagreement (i.e., the number of pairs for either more, less, or less, more). For our simple example, we get a Gamma value of 1.0 (see Figure 3.4(d)). This is the maximum value that could be obtained. Values of Gamma run from − 1.0 through 0 to + 1.0, with maximum association at + 1.0 or − 1.0 (the sign merely indicating direction of the association). There is no association at 0.

While there is disagreement among statisticians about the way to treat ties, and about the desirability of one measure over others, Gamma has some nice features. First, it is limited, running only between − 1.0 and + 1.0, and is symmetrical about 0 in the degree of association. Secondly, it is fairly easy to interpret. The interpretation goes something like this. If we discard all instances in which ties occur (even on only one of the variables), then Gamma is the probability of correctly predicting the order of a pair of cases on one variable once the ordering on the other variable is known. What does this mean? Go back to our example in Figure 3.4, where there were 16 pairs. Of these 16 pairs, there were 12 that had a tie in at least one variable (1 + 2 + 1 + 2 + 6). Of the remaining 4 pairs, 2 were (less, less) and 2 were (more, more). In other words, of the 4 pairs upon which Gamma is to be based, we have none that are discordant such as (more, less) or (less, more). So if somebody were to tell you that the order for a particular pair of individuals was less on variable X, you would guess that the order on variable Y was also less — and you would always be right, since there were no (less, more) pairs. That is why Gamma is 1.0 in this case; the probability of correctly guessing the order on one variable given the order on the other is unity. There are no instances in which discordance appears.

This example also points up one of the disadvantages of using Gamma. We have calculated a measure of association for pairs of observations, but it is based upon only 4 out of the 16 possible pairs. It may be that we would be unwilling to eliminate that much information. The next section will discuss an association measure which uses the same paired form observations as for Gamma, but which treats ties in a different manner.

It is obvious that for anything but the smallest of tables, creating a list of all possible pairs will be too time-consuming to be worthwhile. However, it is a fairly easy matter to calculate the necessary 3 × 3 table of paired comparisons for contingency tables of modest size. A step-by-step procedure is given in Figure 3.5 for a hypothetical contingency table that is 2 × 3. The description of the steps apply to tables of any size and is general. In Figure 3.5(b), the simple example in Figure 3.4 is reworked, this time using the general equations, so that a comparison can be made to establish the identity of the two different procedures for obtaining the 3 × 3 table of pairs.

Figure 3.5:

A Simplified Procedure for Computing the Paired Comparison Form

a) The original table[a]

Variable Y	Variable X			Totals
	Low	Med	High	
High	n_1	n_2	n_3	N_4
Low	n_4	n_5	n_6	N_5
Totals	N_1	N_2	N_3	N_{total}

b) The paired comparison form

Variable Y	Variable X			Totals
	Less	Same	More	
More	r_1	r_2	r_3	R_4
Same	r_4	r_5	r_6	R_5
Less	r_7	r_8	r_9	R_6
Total	R_1	R_2	R_3	R_T

Figure 3.5:

continued

c) Computation procedure for finding paired form cell values

$R_T = (N_{Total})^2$

$R_2 = N_1^2 + N_2^2 + N_3^2$ (square each of the column marginals and add)

$R_5 = N_4^2 + N_5^2$ (square each of the row marginals and add)

$R_1 = R_3 = (R_T - R_2)/2$

$R_4 = R_6 = (R_T - R_5)/2$

$r_5 = (n_1^2 + n_2^2 + n_3^2 + n_4^2 + n_5^2 + n_6^2)$ (square each of the cell values in the original table and add)

[a] This example uses a 2x3 table, but can be extended to any size

$r_2 = r_8 = (R_2 - r_5)/2$

$r_4 = r_6 = (R_5 - r_5)/2$

$r_3 = n_3 \cdot (n_4 + n_5) + n_2(n_4)$

For r_3: Start in the High-High corner of the original table (upper right-hand corner). Multiply this number (n_3) times the values to the left and in all lower rows ($n_4 + n_5$ in the example). Move to the next most right-hand value in row 1 (n_2) and do the same. Here the only value to the left and in a lower row is n_4. Continue in this manner across the first row until the first column is reached and then proceed to the second row, far right-hand columns. Proceed across this row in the same manner unless it is the last row. When the last row is reached stop.

r_1, r_7 and r_9 can be filled in because the marginals are known and enough cell values are known to allow these terms to be calculated.

In these terms, Gamma = $\dfrac{r_3 - r_1}{r_3 + r_1}$

Figure 3.5:

continued

d) Example from Figure 3.4

Original form:

Variable Y	Variable X		Totals
	Low	High	
High	1	1	2
Low	2	0	2
Totals	3	1	4

Paired form:

$$R_T = 4^2 = 16$$

$$R_2 = 3^2 + 1^2 = 10$$

$$R_5 = 2^2 + 2^2 = 8$$

$$R_1 = R_3 = (16 - 10)/2 = 3$$

$$R_4 = R_6 = (16 - 8)/2 = 4$$

$$r_5 = 1^2 + 1^2 + 2^2 + 0^2 = 6$$

$$r_2 = (10-6)/2 = 2$$

$$r_4 = (8-6)/2 = 1$$

$$r_3 = 1(2) = 2$$

At this point we have:

Variable Y	Variable X			Totals
	Less	Same	More	
More		2	2	4
Same	1	6	1	8
Less		2		4
Totals	3	10	3	16

Figure 3.5: continued

From this we see that $r_1 = 0$, $r_7 = 2$, and $r_9 = 0$.

This table can be compared to the paired table in Figure 3.4(c) to see that it is identical.

Exercise 3.6: Table 3.9 shows the plans of students in the last year of secondary school for attending university or vocational schooling. The students have been classified according to whether they reside in rural or urban areas. In part (b) of Table 3.9, the paired-form contingency table has been constructed. Use the formulas provided in Figure 3.5 to verify the cell values in the paired form shown in Table 3.9. Compute Gamma and comment on the strength and the direction of the relationship between residence and schooling plans. Assume that the data in Table 3.9 are for a total population of interest.

Table 3.9.

Plans for Further Schooling among Urban and Rural Secondary School Students, for Hypothetical Data

a) Table of counts

School Plans	Residence		Totals
	Urban	Rural	
Plan to attend university	2072	286	2358
Plan to attend vocational school	284	61	345
Do not plan to attend any school	574	75	649
Totals	2930	422	3352

b) Paired comparison form (See Figures 3.4 and 3.5 for an explanation of terms)

School Plans	Residence			Totals
	Less	Same	More	
More	303,092	1,984,263	280,402	2,567,757
Same	652,966	4,794,458	652,966	6,100,390
Less	280,402	1,984,263	303,092	2,567,757
Totals	1,236,460	8,762,984	1,236,460	11,235,904

3.4 *Somer's D*

If we can identify one of the two variables in a table as the independent variable, then it is possible to compute a measure of association, called Somer's D, which uses more of the information in the paired form table than does Gamma. Recall that all ties on either variable were excluded from the computation for Gamma. But with Somer's D all those cases for which there are ties on the dependent variable but no ties on the independent variable are also included. The absence of ties on the independent variable is necessary if meaningful interpretation is to be made, as explained below.

The formula for Somer's D is given below, where P and Q are defined as previously, and where T is all pairs which are different on the independent variable but are tied on the dependent variable.

$$\text{Somer's D} = \frac{P - Q}{P + Q + T} \qquad (3.4)$$

This is made more meaningful if we return to our simple example in Figure 3.4. Assume in this figure that variable X is the independent variable. In Figure 3.4(c) we see that there are two pairs which are tied on the dependent variable Y, but not on the independent variable X (they are the one pair for less, same, and one pair for more, same). Thus Somer's D for this simple example is:

$$\text{Somer's D} = \frac{4 - 0}{4 + 0 + 2} = .67$$

How are we to interpret Somer's D? Basically, it has the same kind of interpretation as Gamma, but one must recall that for Somer's D we have explicitly made one of the two variables the independent variable. We are thus interested in the prediction of the dependent variable by the indepen-

dent variable, and have explicitly identified which variable is which. For the simple example above, Somer's D is .67, indicating that this is the probability of predicting the order of a pair of cases on the dependent variable, once the ordering of the independent variable is known. For example, if someone tells you that the order for a particular pair was less on X, you would go with the odds and predict that the order on Y was also less. And you would be right 67% of the time, since there are two (less, less) pairs and only one (less, same) pair.

As with Gamma, Somer's D runs from − 1.0 for a perfect negative association, through 0.0 for no association, to + 1.0 for a perfect positive association. Note, too, that Somer's D is always less than Gamma, because the numerator in the formula is the same, but the denominator is always larger for Somer's D than for Gamma. Because Somer's D utilizes more information from the paired-form table, it is generally the more suitable indicator of the strength of an association whenever it can be stated which of the two variables is independent and which is dependent. Note, too, that if we compute Somer's D for the case when X is the independent variable and Y is the dependent variable, it would differ from the Somer's D computed when the role of the variables was reversed — that is, where Y was the independent variable and X the dependent variable.

Exercise 3.7: If we identify Residence as the independent variable in Table 3.9, where Residence and Schooling Plans are shown, verify that Somer's D for the table is 0.018. Interpret this finding.

4.0 *The Nature of an Association*

In previous sections the direction of an association between ordinal level variables was discussed. In determining directionality we are asking "the more of this, the more of that" questions. For example, we might find that as educational attainment rises, so too do earnings. Measures of association assess the strength of such relationships.

But it sometimes is the case that there is not a steady or uniform pattern to an association. Instead of a smooth transition in the values of one variable as the second variable goes from low to high, the pattern may be quite different. For example, it may rise at first and then level off; or rise to a point and then begin to fall.

This pattern of change in one variable as the second variable goes from its low to higher values is referred to as the *nature of an association*. There are several simple things that can be done to assess the nature of an association, and the purpose of this section is to explore two such procedures.

4.1 *Percentaged Tables*

Shown in Table 3.10 are some data for a sample of economically active heads of households in El Salvador. For our present purposes we will treat

this sample as if it were population data, and not worry about inferences from the sample to the population. The percentaged table, shown in Table 3.10(b), can be used to get a visual impression of the nature of an association. Looking along any given row (i.e., a category of the dependent variable), identify the largest percentage by underlining or circling the figure. Follow this procedure for all rows in the table. If two or more percentages are nearly the same, mark both of them. The resulting pattern of marked percentages gives a feeling for how the variables are associated. Table 3.10(b) shows that the association between age and earnings does not increase steadily with age. Instead, we see a rise in earnings through the 35-54 age category, but a rather sharp drop after that time.

Table 3.10:

Age and Earnings for a Sample of Economically Active Heads of Households, El Salvador: 1975

a) Table of counts

Earnings	Age				
	15-24	25-34	35-54	55+	Totals
High	15	100	195	68	378
Medium	121	462	728	271	1582
Low	187	452	1015	511	2165
Totals	323	1014	1938	850	4125

b) Percentage table

Earnings	Age				
	15-24	25-34	35-54	55+	Totals
High	4.6	9.9	10.0	8.0	9.2
Medium	37.5	45.5	37.6	31.9	38.3
Low	57.9	44.6	52.4	60.1	52.5
Totals	100.0	100.0	100.0	100.0	100.0

Source: Center for Studies in Education and Development, "An Assessment of the Relevance of Education in El Salvador," Harvard University, 1977, unpublished documentation.

4.2 *Prediction Analysis*

Constructing percentage tables is a convenient and simple method of learning something about the nature of the association between two variables. However, if one wants to explore a little farther, there is another technique which may be of help, called prediction analysis. This technique permits the comparison of several predictions about tabled data. As the name implies, certain predictions, or hypotheses, about the data must be stated explicitly. With the age and earnings data in Table 3.10, for example, we could make several predictions.

Shown in Table 3.11 are two predictions that have some appeal. The first is a generally linear hypothesis, where the prediction asserts that young people will have low earnings; those 25 to 54 years will have moderate earnings; and people older than 55 years will have high earnings. The second prediction proposes an inverted U-shaped relationship between age and earnings: young people are again expected to earn little, while those in the prime working ages are expected to have high earnings, and the more elderly are expected to have low earnings.

Table 3.11:

Two Predictions About the Nature of the Relationship Between Age and Earnings, for the Data in Table 3.10

a) Prediction 1: Earnings generally increase with age

Earnings	Age			
	15-24	25-34	35-54	55+
High				X
Med		X		
Low	X			

b) Prediction 2: Inverted U-shaped relationship

Earnings	Age			
	15-24	25-34	35-54	55+
High		X	X	
Med				
Low	X			X

These two predictions are obviously asserting very different things about the nature of the association between age and earnings. If we had a way to assess which of the predictions was better, we would also be learning something about the nature of the association between the variables.

In fact, it is possible to determine the quality of the predictions, but it is necessary to have some comparison model against which to judge each one. The comparison model used is the independence model described when discussing the Chi-square statistic. Basically, each prediction is judged on how well it predicts the actual cell values as compared to how well the model of independence predicts the cell values.

We can evaluate how good our prediction is in terms of the *prediction success* (called Delta) and the *precision* (simply called U). Formulas for each exist and are given as:

Delta = prediction success = 1 − (observed errors/expected errors) (3.5)
U = precision = expected error rate = expected errors/total N (3.6)

The higher the prediction success and the higher the precision, the better the prediction. Values for the prediction success, Delta, can range from a maximum of one to negative numbers, but zero is pretty low. The lower the number, the worse the prediction. The precision can range from zero to one, with high values indicating a difficult prediction and lower values indicating an easier prediction. One prediction is said to *dominate* another if it has a higher prediction success and has at least as high a precision. This allows one to rank-order predictions in a hierarchy, which may be useful for distinguishing the relative merits of various predictions.

Shown in Figure 3.6 are the cells (encircled) where prediction 1 in Table 3.11 is wrong. These are the error cells for this prediction. Also shown in the figure are the actual cell values found in our data, as reproduced from Table 3.10, plus the cell values expected under independence. Recall that each of these expected cell values is computed as the product of the appropriate row and column marginals divided by the total N. With the numbers in Figure 3.6, it is possible to compute the prediction success and the precision of our prediction. This is also shown in Figure 3.6, where we see that the prediction success is positive but small, 0.024. On the other hand, the precision, 0.67, is fairly high. For prediction 2 we find that the prediction success is 0.033 and that the precision is 0.79, indicating a somewhat better prediction than the first. Since prediction 2 has both a higher prediction success and a higher precision, it is said to dominate prediction 1. In other words, the nature of the association between age and earnings, for our tabled data, is more the inverted U-shape than it is linear.

Exercise 3.8: Calculate the prediction success and the precision for Prediction 2 (in Table 3.11). The information in Figure 3.6 will be of help in these computations.

Figure 3.6:

Computation of Prediction Success and Precision for Prediction 1 in Table 3.11, Showing Age and Earnings Among Economically Active Heads of Households in El Salvador

a) Table of observed values, with error cells for prediction 1 in Table 3.11 circled.

Earnings	Age				Totals
	15-24	25-34	35-54	55+	
High	15	100	195	68	378
Med	121	462	728	271	1582
Low	187	452	1015	511	2165
Totals	323	1014	1938	850	4125

b) Table of expected errors

Earnings	Age				Totals
	15-24	25-34	35-54	55+	
High	30	93	178	78	378
Med	124	389	743	326	1582
Low	170	532	1017	446	2165
Totals	323	1014	1938	850	4125

c) Computation of prediction success (V) and precision (U)

$$V = 1 - \frac{15+121+100+452+195+1015+271+511}{30+124+93+532+178+1017+326+446}$$

$$= 1 - \frac{2680}{2746}$$

$$= 0.024$$

$$U = \frac{2746}{4125} = 0.67$$

Shown in Figure 3.7 is a summary of a number of predictions for the data on age and earnings that we have been considering. Prediction success and precision are indicated in the figure, as is a visual representation of the pattern of association indicated by a particular prediction. Note that none of the predictions has very good success; the highest Delta is just 0.060, for prediction 3. Prediction 4, with a prediction success nearly equal to that of prediction 3, but a considerably higher precision, is probably the *best* prediction of the group. It indicates that earnings rise with age, until the over-55 age group, who have low earnings.

Note that our previous predictions, 1 and 2, are represented in Figure 3.7, but that each is surpassed in some respect by others of the predictions. While prediction 4, our *best* prediction, does not dominate prediction 2, (because prediction 2 has a higher precision than prediction 4), the prediction success is sufficiently higher for prediction 4 that we would likely trade off some precision for the added success in prediction. Unfortunately, there is no uniformly agreed upon method for selecting one best prediction when a single prediction does not dominate all others.

Figure 3.7:

A Summary of Several Predictions Concerning the Nature of the Relationship Between Age and Earnings Among Economically Active Heads of Households in El Salvador, as Shown in Table 3.10

Prediction	Prediction Success	Precision	Pattern
3	0.060	0.58	
4	0.059	0.71	
2	0.033	0.79	
5	0.028	0.65	
1	0.024	0.62	
6	0.016	0.74	

5.0 *Summary*

This chapter has introduced a variety of techniques for obtaining information from tables of counts. Initially we were interested in determining if there was indeed a relationship between two variables. Secondly, we wanted to determine the strength of the association and if the variables were ordinal, its direction. Thirdly, we wanted to explore the nature of the relationship.

In the process of addressing these questions, we saw how percentaged tables can be used to assess both independence between variables and to get a feeling for the nature of the relationship between variables. The Chi-square measure of independence was discussed, and its application with sample data indicated. Several measures of association were described, including Cramer's V, which uses Chi-square; Yule's Q for 2 · 2 tables; Gamma for ordinal variables, where a paired comparison form is computed; and Somer's D, which is similar to Gamma but treats the ties in the paired comparison form differently. Prediction analysis was also introduced as a technique that can be used to provide information on the nature of the association between variables.

Without explicitly mentioning it, this chapter also provided the first indication of a fairly general approach in data analysis, that of comparing the actual data with some hypothesized model. For example, we used the model of independence as a hypothesized model for computing Chi-square, and also for computing expected errors for the prediction analysis technique. This method of using a model to predict expected values, and then comparing the expected values with the actual values to see how closely they compare, is a procedure that we will find in the next two chapters.

6.0 *Terms for Review*

contingency tables
association
independence
total N
marginals
cell values
"A · B table"
positive association
negative association
Chi-square
Expected cell values under independence
Cramer's V
Yule's Q
Gamma
Somer's D
paired form contingency table
nature of association
prediction success

prediction precision
prediction dominance
error cells in prediction analysis

7.0 *Practice Exercises*

1. Shown in Table 3.12 are data on enrollment trends and school district size, for all 436 school districts in the North Central region. Construct an appropriate percentaged table to indicate whether district size and enrollment trend are associated. Use this table to roughly indicate the nature and direction of the association between the two variables.

Table 3.12

A Table of Counts Showing District Size and Enrollment Trends for the North Central Data Set

Enrollment Trend	District Size				Totals
	0-299	300-799	800-1799	1800+	
Fast growth	2	11	19	27	59
Slow growth	3	6	14	10	33
Stable	15	65	49	36	165
Slow decline	6	34	24	12	76
Fast decline	23	50	17	13	103
Totals	49	166	123	98	436

2. Suppose that instead of the population data found in the Table 3.12, the analysts had drawn a random sample of 100 school districts. Such a table, for school size and enrollment trend, is found in Table 3.13. Compute a Chi-square value for this table. Are the two variables, as represented in the table, independent? What would you conclude about independence between the variables if you had to use these sample data to make inferences about the entire population of school districts in North Central?

3. Using the formulas provided in this chapter, construct the paired form contingency table for the data found in question 1. Compute both Gamma and Somer's D. Comment on the direction and strength of association between district size and enrollment trend.

Table 3.13:

A Table of Counts Showing District Size and Enrollment Trends, for a Sample of School Districts in the North Central Data Set

Enrollment Trend	Sample Size				Totals
	0-299	300-799	800-1799	1800+	
Growth	2	1	8	7	18
Stable	3	13	8	5	29
Decline	5	32	13	3	53
Totals	10	46	29	15	100

4. Use prediction analysis to explore in more detail the nature of the association between district size and enrollment trend, using the data in question 1. In particular, compare the following three predictions:

Size	*Prediction 1*	*Prediction 2*	*Prediction 3*
0-299	fast decline	fast decline	stable, slow decline
300-799	fast decline	slow decline	fast decline
800-1799	slow decline	stable, slow growth	fast decline
1800+	fast growth, slow growth, stable	fast growth	stable, slow decline

Compute prediction success and the precision for each of the predictions. Does any of the three predictions dominate the others? Comment on the nature of the association between district size and the enrollment trend.

Bibliography

Henkel, Ramon E. "Tests of Significance." Beverly Hills: Sage, 1976. This text, one in a useful series of Sage Publications on Quantitative Applications in the Social Sciences, discusses both the pros and cons of tests of significance like Chi-square, where one is interested in inferences to populations from sample data.

Hildebrand, David K.; James D. Laing; and Howard Rosenthal. "Analysis of Ordinal Data." Beverly Hills: Sage, 1977. In this text, the authors discuss various measures of the strength of association for ordinal data, using the prediction analysis format. It provides a good introduction into the procedures for prediction analysis.

Hildebrand, David K.; James D. Laing; and Howard Rosenthal. *Prediction Analysis of Cross Classifications*. New York: Wiley, 1977. This text provides a thorough discussion of prediction analysis in more detail and complexity than the monograph by the same authors listed above.

Loether, Herman J., and Donald G. McTavish. *Descriptive Statistics for Sociologists*. Boston: Allyn and Bacon, 1974. There are two useful chapters in this book dealing with contingency tables and various measures of association for both nominal and ordinal variables.

Reynolds, H. T. "Analysis of Nominal Data." Beverly Hills: Sage, 1977. The author discusses and compares various measures of association for nominal variables. Also included is a good introductory discussion of techniques for use when there are three or more nominal variables of interest.

CHAPTER 4

Tables of Central Tendency

Gary Lewis

1.0 *What's Ahead?*

In Chapter 2 the problem of comparing measures of central tendency was discussed. In one of the examples, comparison of pupil-teacher ratios was made for various countries, grouped by the level of GNP in the country. We used techniques like the stem-and-leaf diagram to find median pupil-teacher ratios, and box plots to make the comparisons across groups of countries. In the present chapter we will again be interested in comparing measures of central tendencies, but in situations where the measures of central tendency can be tabled.

Shown in Table 4.1 are average earnings in the United States, cross-classified by two variables, educational attainment and age. We see, for example, that the average earnings in 1976, for people 25 to 34 years old with a secondary school education, was $8,283. By scanning the table it is fairly easy to see the most important features:

— Individuals with higher levels of education earn, on average, more than those with fewer years of education. This is evident by scanning down columns.

— Older individuals earn, on average, more than younger people, although there is a reversal of this trend after age 55.

Table 4.1:

Average Earnings, by Educational Attainment and Age, for Those Employed, 16 Years and Older: United States, 1976

Educational Attainment	Age					
	16-24	25-34	35-44	45-54	55-64	65+
University	[illegible]	[illegible]	16146	17442	[illegible]	[illegible]
Secondary	4622	8283	9630	10214	9799	5405
Elementary	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Source: U.S. Bureau of the Census, Current Population Reports, Series P-60, No. 114, "Money Income in 1976 of Families and Persons in the United States," Washington, D.C.: U.S. Government Printing Office, 1978, Table 43.

These are important observations, and ones that can be readily made without any extra effort. For some purposes this may be sufficient. However, it is possible to do some simple analysis with these data that will extract additional information from the table. One such technique is called *two-way analysis*, because the measures of central tendency are cross-classified by two independent variables. In Table 4.1 the independent variables are, of course, age and educational attainment. The dependent variable is earnings.

The material in this chapter will describe how to analyze two-way tables of central tendency, so as to obtain more information than is possible by simply looking at the table. In the process of describing techniques and procedures of two-way analyses, we will:

a. Further develop the basic strategy of summarization and exposure that was begun in the previous chapter. Recall that with tables of counts we calculated expected cell values under the assumption (or model) of independence. When computing expected values we were attempting to *summarize* important features in the table. These expected values were then compared to the actual cell values to determine if they were close. The difference between the expected and actual cell values was then used to indicate the likelihood that the variables were independent or associated. By comparing expected and actual values, we *exposed* or tested the goodness of fit of our model to the data, so as to indicate whether the model and its implications seemed reasonable.

In the present chapter we will use a similar strategy to summarize the cell values in tables such as Table 4.1, although with a model different from that used in the last chapter. These summaries, or calculated cell values, will be compared to the actual cell values to indicate goodness of fit between the model we proposed and the actual data.

b. Identify in a step-by-step approach the various procedures necessary to fit a model to the data, and to compute a measure of goodness of fit.

c. Indicate how a comparison of expected cell values and the actual cell values can help in reassessing the model fitted to the data. This can be used to decide what other model might better fit the data.

d. Use examples that stress interpretation, so that the utility of two-way analyses is clear.

e. And, finally, we will use the techniques in this chapter to further explore the relationship between fluctuating enrollments and the cost of education in North Central.

2.0 *The Additive Fit*

2.1 *The General Idea*

As with the Chi-square techniques of the previous chapter, the analysis of tables of means or medians uses a prediction and comparison strategy. First, the cell values are predicted using a model, or hypothesis, describing how the variables are related. Secondly, the predicted values are compared to the

actual values to determine how well our hypothesis has succeeded in fitting the data. If the fit is close, we then assume that the hypothesis tells us something of importance about the data.

With Chi-square, our basic hypothesis was that the two variables in the table were independent. This permitted simple calculations of expected cell values. With two-way analysis, the basic model is called the *simple additive fit*. Stated as a hypothesis, when we predict cell values with an additive fit, we are assuming two things:

— that the dependent variable (i.e., the tabled values of central tendency) varies as a result of both the independent variables in the tables;

— and, furthermore, that the effects of one independent variable may be *added* to the effects of the second independent variable to help predict the dependent variable.

Shown in Table 4.2 are some hypothetical figures for average-per-pupil school expenditures, by region and year. We see, for example, that in 1970, schools in rural regions had average expenses per pupil of $50, compared to average expenses in 1960 of $30. The $20 differential between years also holds for schools in urban areas, but we see that the averages in both years are higher for urban schools than for rural schools. In both 1960 and in 1970, average expenses for urban schools were $30 higher than for rural schools.

Table 4.2:

Average-Per-Pupil School Expenditures, for a Hypothetical Sample of Districts, by Region of the Country and Year

Region	Year	
	1960	1970
Rural	30	50
Urban	60	80

One reasonable hypothesis about these data, then, is that average expenses in 1970 are about $20 higher than in 1960, and that expenses in urban areas are about $30 higher than in rural areas. This hypothesis can be more

formally stated in the following manner:

$$\text{Predicted average expenses} = \text{common value} + \text{year effect} + \text{region effect} \tag{4.1}$$

The *common value* is a number that is typical of all the numbers in the table. In Table 4.2, we have numbers of 30, 50, 60, and 70. The median of these in 55, and could be considered to be a value typical of the four tabled values. We could then choose − 10 (for 1960) and + 10 (for 1970) as the yearly effects, and − 15 (for rural) and + 15 (for urban) for the regional effects. Note we said earlier that expenses in 1970 were about $20 higher than in 1960. This is represented by a spread in the effects of $20 (from − 10 to + 10). Similarly, the difference of $30 between urban and rural is represented as the spread from − 15 to + 15.

With these estimates of a common value and effects of both year and region it is possible to predict the cell values. For the cell (rural, 1960) we have:

$$\text{Estimate for (rural, 1960)} = \$55 + (-\$10) + (-\$15) = \$30.$$

In this case our estimate of the (rural, 1960) cell is precisely the actual cell value. By making estimates for the other cell values, the reader will find that for this simple case all predicted cell values equal the expected cell values.

Exercise 4.1: Find the predicted cell values for Table 4.2, using the common and effect estimates given in the text.

In other words, our hypothesis has done a very good job of predicting actual expenses, so good, in fact, that we have made no errors in prediction. This is not a usual result, but this example was discussed primarily to make clear what was meant by an additive fit. With an additive fit, we estimate each cell value as a *sum* of three components: the typical value describing all cells, and the effects of the two independent variables. Note, too, that we interpret the effects by comparing them to each other. For example, the effect of − 10 for 1960 is not meaningful by itself. It takes meaning when compared to the + 10 effect for 1970, indicating that there is a $20 difference in average expenses between 1970 and 1960. Similarly, the effects for region indicate there is a $30 difference in average expenses between urban and rural areas.

In most cases predicted cell values will not equal the actual cell values. The difference between the predicted and actual cell values is called the *residual*. Thus we could rewrite equation (4.1) as:

$$\begin{aligned}\text{actual cell value} &= \text{predicted cell value} + \text{residual} \\ &= \text{common value} + \text{column effect} + \text{row effect} + \text{residual}\end{aligned} \tag{4.2}$$

In equation (4.2), "column effect" has been used as a general term to replace the specific term in equation (4.1) that we called the "year effect". Similarly "row effect" is the more general term for what we called "region effect" in equation (4.1). Column and row effects are commonly used because they are

appropriate for any table.

The simple additive model, or hypothesis, applied to the earnings figures in Table 4.1 would assert that average earnings could be expressed as the following:

$$\text{average earnings} = \text{common value} + \text{age effect} + \text{education effect} + \text{residual} \tag{4.3}$$

With the simple example in Table 4.2, it was possible to determine a typical value and the various effects by inspection. Most tables do not lend themselves to analysis by inspection. However, there are a series of steps that have been worked out to permit estimation of the various terms in equation (4.2). It is these steps that we will take up now.

2.2 *The Mechanics of an Additive Fit*

The manner in which the common, row, and column effects and residuals are found is quite mechanical. The steps involved are the following:

1. Start with the data written in tabular form.
2. Calculate row medians (means could also be used, but medians are easy to compute).[1]
3. Compute the *common value*, as the median of the row medians.
4. Subtract the common value from the row medians to get the *row effects*.
5. Subtract the row medians from the cell values in step 1 to get row residuals.
6. Get the *column effect* for each column by determining the median value of the row residuals in step 5 for each column.
7. Get final *residuals* by subtracting the column effects from the row residuals.

Each of these steps is detailed in Figure 4.1. The figure shows the steps involved in the analysis of the average earnings data in Table 4.1. Note that each one of the steps has been numbered so that it can be compared to the numbered list of steps just given. Also note that there is an easy way of writing the steps down on paper with lines to divide the various steps. This is not necessary, but makes for somewhat easier presentation of results.

Just to be certain that it is clear what has been going on with this arithmetical manipulation, recall that what we are after is a more refined analysis of the earnings data presented in Table 4.1. We were able to make two important observations just by examining the table: first, there is a steady increase in earnings with increased education level, irrespective of age; and second, irrespective of education level, earnings increase with age until 55, thereafter decreasing. These are important conclusions, but the analysis we just completed is being used to dredge more information from the table. We have chosen a two-pronged strategy of summarization and exposure, summarizing by fitting the observed data and exposing by determining how well the fit approximates the observed data. Recall too that the fitted data are based

Figure 4.1:

A Two-Way Analysis of the Average Earnings Data Given in Table 4-1[a]

	16-24	25-34	35-44 (1)	45-54	55-64	65+	(2)	(5)					
University	4502	11023	16146	17442	16475	8758	13584	-9082	-2561	2562	3858	2891	-4825
Secondary	4522	8283	9630	10214	9798	5405	8956	-4434	-673	674	1258	842	-3551
Elementary	3727	6331	7445	7934	6997	3260	6664	-2937	-333	781	1270	333	-3404
					(3)		8956	-4434	-673	781	1270	842	-3551 (6)
							4628	-4648	-1888	1781	2588	2049	-1274
					(4)		0	0	0	-107	-12	0	0 (7)
							-2292	1497	340	0	0	-509	147

The final analysis is in the form:

Common Effect (3)	Column Effects (6)
Row Effects (4)	Residuals (7)

[a] The circled numbers refer to steps in the computation procedures detailed in the text.

explicitly upon a particular *model*. The model we have chosen to use insists that earnings are dependent upon the age and upon the education of the individual, and furthermore that the effects of both age and education can be considered additive.

To make the model explicit let us take one example from the analysis in Figure 4.1: elementary school graduates 25–34 years of age. Our model tries to approximate the actual values for this cell. From Figure 4.1 we see that the equation for the model (equation 4.3) gives the following:

$$6331 = 8956 + (-2292) + (-673) + 340,$$

where 8956 is the value common to all observations; −2292 is the effect of having an elementary level education; −673 is the effect of being 25–34 years old; and the residual is 340. The predicted cell value is just 6331 − 340 = 5991.

What do these and the other numbers in Figure 4.1 mean? There are several factors of importance to note, all of which add to a substantial refinement of the previous visual analysis.

From the summarization:

1. The common effect value of $8956 gives an indication of the mid-range of average earnings in all educational attainment and age categories.
2. The row effects indicate substantial differences in average earnings for people with various levels of educational attainment. People with elementary level education earn about $2300 less than those with a secondary education, while those with a university education earn nearly $4600 more than those with a secondary education. Recall that effect estimates are interpreted by comparing groups.
3. The column effects are interesting for two reasons: first, while the average earnings begin to decline after age 55, it is not until the category of people with ages over 65 that a sharp decline occurs; and second, two age groups, these being the old (65 +) and the young (16–24) have much lower than average earnings levels. Compared to those aged 35–44, youth 16–24 earn about $5200 less (i.e., 781 + 4434 = 5215), and the elderly 65 and over nearly $4300 less.

These findings indicate that certain situations exist, they do not provide reasons why such situations occur. For example, young people might earn less because employers put them in entry level jobs where no experience is necessary and wage levels are low. Or they may earn less because they are in school and only working part-time. Or it may be that they are less committed to working, finding it disagreeable, and thus enter and exit at will from the labor market. There are many competing explanations for this finding, or any other. Which of the competing explanations is the most reasonable cannot be assessed with the information here. It is worthwhile pointing out that an interpretation of these results would have to be done by someone

familiar with the situation, or would require further collection of information and a more detailed understanding of other factors.

From the exposure:

1. If the magnitude of the residuals is small, our model has done a fairly decent job. The residuals in Figure 4.1 show that the additive model closely predicts average earnings of those with an elementary or secondary education, in most age groups. But the model does not predict average earnings of university-educated individuals very well. In Figure 4.1, many of the residuals are small in the elementary and high school categories, indicating a good fit. On the other hand, the residuals for college-educated groups are large in magnitude compared to the college row effect (4628), indicating a poor fit. Shortly we will discuss a method for better assessing the goodness of fit of the model.
2. From the residuals it is also possible to isolate specific instances where the model is not very powerful. In particular, the model performs poorly for elementary school graduates 18–24 years old (where the residual is a large positive number), and university-educated 18–24-year-olds (where the residual is large and negative). Both are interesting in that they suggest something else, apart from the effects of age and education, is affecting the earnings of these individuals. For example, if could be expected that many university-educated 18–24-year-olds would still be in school, perhaps working part-time. Hence their average earnings might be less than expected. Similarly, young people with only an elementary school education might find themselves doubly disadvantaged in the labor market, where both their inexperience and their relative lack of education would make it difficult to find employment that paid well.

2.3 *Polishing an Additive Fit*

The procedure just described for determining the common and effect values can be refined by doing a second or even third iteration of the basic steps discussed. This results in a better fit of the model, meaning that the actual and predicted values will be closer and the residuals smaller. With enough iterations, a point is reached where going through the steps produces no additional change. At this point, if one is using a computer to do the steps, the analysis stops. However, when doing the analysis by hand, it consumes too much time and effort, and one would not often complete more than one or two iterations.

The steps for doing these "polishing" iterations are similar, with minor variations, to the steps for the basic fit. Below, they are numbered in sequence, beginning where the basic procedure leaves off:

Step 8. For each row of residuals, *and* the row of column effects, find the median.

Step 9. Subtract the medians found in step 8 from the appropriate residuals or column effects.

Step 10. *Add* the medians found in step 8 to the appropriate row effects (i.e., the row effects in step 3).

Step 11. For each column of new residuals from step 9, *and* for the column of new row effects from step 10, find the median.

Step 12. Subtract the medians in step 11 from the appropriate new residuals or new row effects found in step 9.

Step 13. *Add* the medians from step 11 to the new column effects from step 9.

Step 14. Repeat step 8 to verify that the new medians are small and converging toward zero, and then stop.

These steps seem confusing when written, but are quite simple in application. In Figure 4.2 the basic fit and the polished fit for the average earnings data of Table 4.1 are shown as one continuous process. The various parts of Figure 4.2 are numbered to correspond with the numbering systems used to describe the basic and polished fits. For example, find the section in Figure 4.2 marked "12." These numbers are the result of performing step 12 in the polished fit procedure. We see, for example, that the column (1601, − 33, 0) is formed by subtracting the row median (− 74) from the appropriate column of residuals from step 9 (1527, − 107, − 74). If we take the (University, 35–44) cell as an example, we see it is represented as:

$$16{,}146 = 9010 + 653 + 4882 + 1601$$

The common value is 9010, the column effect for those aged 35–44 is 653, the row effect for university education is 4882, and the residual is 1601.

The interpretation of the numbers in Figure 4.2 proceeds in the same way as with the numbers from the basic fit. For example, those with an elementary education earn about $2218 less than those with a secondary education, while those with a university degree earn about $4882 more. These numbers compare with the effect estimates from the basic fit of $2292 and $4628, respectively. By polishing the fit we have achieved a more refined estimate of the row, column, and common effects. As with the basic additive fit, however, we note the large residuals for the university-educated, indicating that polishing does not do much to improve the fit for this group.

If we compare the medians computed at each of the steps, we can see that the polishing technique results in a more refined analysis. For example, the largest medians decrease in the following pattern: 13,584 (step 2), 4434 (step 6), 254 (step 8), 74 (step 11), and 37 (step 14). If the polishing step were taken further, the medians used to compute new residuals would go toward zero, meaning that the computed effects would stabilize. Usually one iteration is sufficient polish when doing the analysis by hand, but one should complete step 14 to verify that convergence is occurring.

2.4 *Goodness of Fit*

In previous sections we said that small residuals indicated a good fit of the model to the data. But the term *small* is not very definitive. It is possible to compute a ratio which indicates how closely we have come to predicting the actual values. The ratio, shown below in equation (4.4), compares the interquartile distance (H-spread, see Chapter 2) of the original cell values with the interquartile distance of the residuals. If, for example, the residuals are

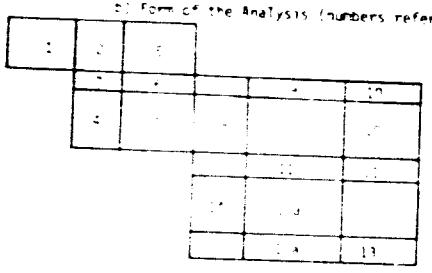
Figure 4.2:

A Two-Way Analysis and a Polished Fit of the Average Earnings Data in Table 4.1

a) Polished fit

Education	Age 16-24	25-34	35-44	45-54	55-64	65+														
University	4502	11023	16146	17442	16475	8758	13564	-9082	-2561	2562	3858	2891	-4825							
Secondary	4522	8083	9630	10214	9798	5405	8956	-4434	-673	674	1258	842	-3551							
Elementary	3727	6321	7445	7934	6997	3260	6664	-2937	-333	781	1270	333	-3404							
							8958	-4434	-673	791	1270	842	-3551	54	-4458	-727	727	1216	788	-3605	9010
							4628	-4646	-1588	1791	2588	2043	-1074	254	-4902	-2142	1527	2334	1795	-1528	4882
							0	0	0	-107	-12	0	0	0	0	0	-107	-12	0	0	0
							-2292	1497	340	0	0	-509	147	74	1423	266	-74	-74	-583	73	-2218
															0	0	-74	-12	0	0	0
														37	-4902	-2142	1601	2346	1795	-1528	4882
														0	0	0	-33	0	0	0	0
														37	1423	266	0	-62	-583	73	-2218
															-4458	-727	653	1204	788	-3605	9010

b) Form of the Analysis (numbers refer to steps listed in the text)



12 = Row effects
12a = Residuals
13 = Common effect
13a = Column effect

near zero, then the ratio in equation (4.4) will be near 1.0, indicating nearly complete reduction in the variation of the original data. In other words, with residuals near zero, nearly all of the variation in the data has been accounted for by the additive model.

$$\text{Goodness-of-fit ratio} = \frac{\text{H-spread original} - \text{H-spread residual}}{\text{H-spread original}} \quad (4.4)$$

Figure 4.3 shows stem-and-leaf diagrams and five-number summaries for both the original earnings data in Table 4.1 and for the residuals in the polished fit in Figure 4.2. The goodness-of-fit ratio can then be computed as:

$$\text{Goodness-of-fit ratio} = \frac{48 - 0.4}{48} = 0.92$$

In other words, the simple additive fit to the data has succeeded in reducing the variation in the original cell values by 92%.

Figure 4.3:

Stem-and-Leaf Diagrams for the Earnings Data in Table 4.1 and for Residuals from the Polished Additive Fit in Figure 4.2

a) Original Data (unit in hundreds)

```
 3 | 37
 4 | 75
 5 | 4
 6 | 3
 7 | 049
 8 | 38
 9 | 68
10 | 2
11 | 0
12 |
13 |
14 |
15 |
16 | 15
17 | 4

N = 18

M |      81      |
H | 54      102  | 48
1 | 33      174  | 141
```

b) Residuals (unit in hundreds)

```
-4* | 9
-4·
-3*
-3·
-2*
-2· | 1
-1* | 5
-1·
-0* | 6
-0· | 10
+0· | 31000000
+0*
 1· | 4
 1* | 68
 2· | 3
 2*

N = 18

M |       0       |
H | -1.0     3.0  | 4.0
1 | -49      23   | 72
```

c) Goodness-of-fit ratio

$$\text{Goodness-of-fit ratio} = \frac{\text{H-spread original} - \text{H-spread residual}}{\text{H-spread original}} = \frac{48 - 4}{48} = 0.92$$

Reduction in variation by 92% indicates a good fit, but there is one additional condition that we would like to have met. We would also like the fit of the data to leave *unstructured residuals*. Shown in Figure 4.4 is an example of two situations, one with relatively random residuals and the other with highly structured residuals. If the residuals are highly structured, as in Figure 4.4(b), we assume that our model is not cast in the correct form, that perhaps the simple additive model is not a good hypothesis in this situation. When we do find structure in the residuals, there are procedures for fitting more complex models, and some of these will be discussed shortly.

Figure 4.4:

An Illustration of Random and Structured Residuals

```
a) No structure                b) Structure

  - . . . +                      + + . - -
  . + . . -                      + + . - -
  . - . + .                      - - . + +
  + . . - .                      - - . + +
```

Key: + = residuals larger than the upper hinge in the distribution

. = residuals between hinges

- = residuals smaller than the lower hinge in the distribution

What about the structure of the residuals for the average earnings data with which we have been dealing? In Figure 4.5 we use symbols -, ., and + to indicate, respectively, residuals smaller than the lower hinge, between the two hinges, and larger than the upper hinge. The figures show the structure of residuals for our additive polished fit. In fact, we see that the residuals are not random. The largest residuals are concentrated in the top row, for university education, indicating once again that the model is not

Figure 4.5:

The Structure of the Residuals from the Polished Additive Fit to the Earnings Data in Figure 4.2

Education	Age					
	16-24	25-34	35-44	45-54	55-64	65+
University	-	-	+	-	+	-
Secondary	.	.	.	.	.	.
Elementary	+	.	.	.	-	.

very adequate for predicting average salaries for people with that level of education. Even though the reduction in variation has been substantial, the model is less than completely satisfactory.

Exercises 4.2: Shown in Table 4.3 are reading achievement scores for 17-year-olds in the United States, for the years 1971 and 1974. Do a two-way analysis of this table, including one polishing iteration. Compute the goodness-of-fit ratio and examine the structure of the residuals. Interpret the effects of parental education and year.

Table 4.3:

Average Reading Achievement Scores for 17-Year-Olds in the United States: 1971 and 1974

Parental Education	Mean Percentages, Correct Responses	
	1971	1974
Post high school	93.2	94.0
High school graduate	89.3	90.7
Some high school	85.0	88.0
No high school	80.7	85.4

Source: U.S. Department of Commerce, Social Indicators 1976 Washington, D.C.: U.S. Government Printing Office, 1977, p. 306.

3.0 *The Additive Fit Extended*

As we have just seen with the example for average earnings, it may be that the simple additive fit does not do a very good job in predicting the actual data. When this occurs, there are some simple extensions of the additive model that may be taken. In the remainder of this chapter, we will discuss two such extensions: i) adding another term to the simple additive fit of equation (4.2); and ii) re-expressing the tabled data.

3.1 *Adding an Interaction Term*

Recall that with the simple additive model shown in equation (4.2) the assumption was made that the effects due to the row and column were purely additive. For example, with the earnings data we assumed that the effects of education could be added to the effects of age to give a good prediction of average earnings. But it may be that age and education *interact*. This would mean that in order to predict the tabled values we need to know something about both the education level and the age group for the cell being predicted. For example, we saw in Figure 4.2 that there is something special about a university education among those 16–24 years old which simply cannot be explained by the effect of a university education plus the effect of the 16–24 age group. Somehow the two features (university education and age) interact to produce average earnings which are much less than what we would predict from the simple additive model of age and education. When such interactions appear, the simple additive model can be extended by including an interaction or multiplicative term, as shown below in education (4.5):

$$\text{actual cell value} = \text{common} + \text{column effect} + \text{row effect} + \frac{k(\text{row})(\text{column})}{\text{common}} + \text{residual} \qquad (4.5)$$

In equation (4.5) the value k is a constant selected to achieve the best fit. Its manner of determination will be illustrated shortly. The (row)(column)/common terms, sometimes called *comparison values*, are similar in appearance to the expected cell value computations used to calculate Chi-square under the assumption of independence. For each cell a value will be computed which depends on both the row and column appropriate to that cell. The values used for row, column, and common effects in equation (4.5) are just the effect estimates from the polished fit.

Values of the constant k in equation (4.5) are chosen to achieve the best fit of the model to the data. To choose k, one first plots what is called a *diagnostic plot*. This is a graph of the residuals for each cell versus the computed comparison values. Shown in Figure 4.6 are the residuals from the polished fit of average earnings data in Figure 4.2 and their computed comparison values. As an example of the manner in which comparison values are computed, we find in Figure 4.2 that the common value for the polished fit is 9010, and that the row effect for elementary education is −2218 and the column effect for the ages 16–24 years is −4488. Thus the comparison value for the cell (elementary, 16–24) is given as

$$\text{Comparison value for (elementary, 16–24)} = \frac{(-2218)(-4488)}{9010} = 1105$$

The numbers in Figure 4.6 can be plotted with the residuals along the Y-axis of the graph and the corresponding comparison values along the X-axis. This is shown in Figure 4.7, along with the residuals and their

Figure 4.6:

Residuals from the Polished Additive Fit to the Earnings Data in Figure 4.2 and Corresponding Comparison Values to Be Used in a Multiplicative Fit

a) Residuals

Education	Age					
	16-24	25-34	35-44	45-54	55-64	65+
University	-4902	-2142	1601	2346	1795	-1528
Secondary	0	0	-33	0	0	0
Elementary	1423	266	0	-62	-583	73

b) Comparison values

Education	Age					
	16-24	25-34	35-44	45-54	55-64	65+
University	-2432	-394	354	652	427	-1953
Secondary	0	0	0	0	0	0
Elementary	1105	179	-161	-296	-194	887

corresponding comparison value. From this plot we can draw a straight line that will pass through the points as closely as possible. This is easily done with a clear ruler or with a piece of string. In Figure 4.7 one possible line has been fitted to the data. In the next chapter on regression analysis, we will discuss a systematic method of selecting a line to pass through the points, but for present purposes the line need not be refined. The reason we want to pass a line through the points is so that we can calculate the *slope* of the line. It is this number that is used as the value for k in equation (4.5).

The slope of a line refers to the steepness of the line, and to the direction it runs. As shown in Figure 4.7, it is easily computed as:

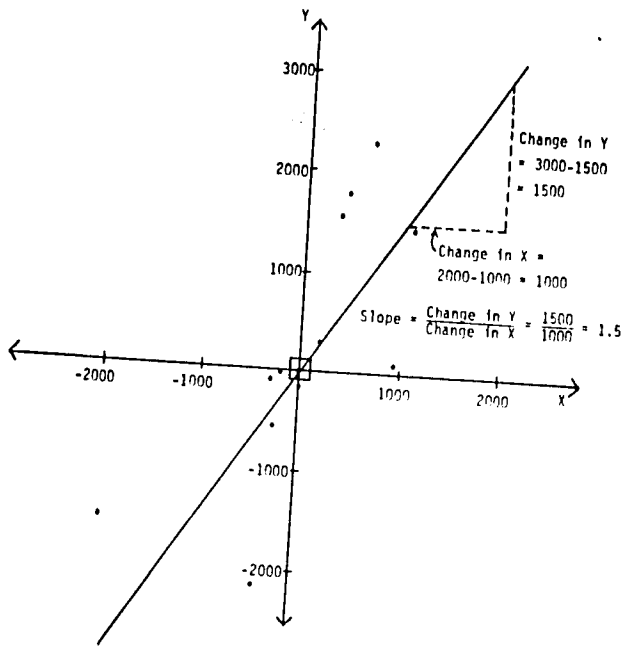
$$\text{Slope of line} \quad \frac{\text{Change in Y}}{\text{Unit change in X}}$$

Figure 4.7 shows that positively sloped lines are those which slant upward to the right, while negatively sloped lines slant down to the right (see Chapter 5 if this is not familiar). Slopes of zero are horizontal, while an infinite slope would be a vertical line. If we return to Figure 4.7, we see that the line fitted to the data has a slope of about 1.5.[2]

Figure 4.7:

A Diagnostic Plot of Residuals Versus Comparison Values, for a Multiplicative Fit to the Earnings Data in Table 4.1

a) Graph*



*5 points (0,0) indicated as ▼; 1 point (-2432, -4902) is off-scale

a) Data

X Comparison Values	Y Residuals
1105	1423
179	266
-161	0
-296	-62
-194	-503
882	73
0	0
0	0
0	-33
0	0
0	0
0	0
-2432	-4902
-394	-2142
354	1601
652	2346
427	1795
-1953	-1528

Figure 4.8:

A Multiplicative Fit to the Earnings Data in Figure 4.2

a) Values of 1.5 (row)(column)/common

Education	Age					
	16-24	25-34	35-44	45-54	55-64	65+
University	-3648	-591	531	978	641	-2930
Secondary	0	0	0	0	0	0
Elementary	1658	269	-242	-444	-291	1331

b) Residuals from the model in equation (4.5)

Education	Age					
	16-24	25-34	35-44	45-54	55-64	65+
University	-1254	-1551	1070	1368	1154	1402
Secondary	0	0	-33	0	0	0
Elementary	-235	-3	242	382	-292	-1258

c) Stem-and-leaf diagram of residuals; units in hundreds

```
-1* | 6
-1· | 33
-0*
-0· | 3200
+0· | 0000024
+0*
 1· | 1244
 1* |
```

N = 18

M	0		
H	-2	4	6
1	-16	14	30

Goodness-of-fit ratio (See Figure 4.3 also) = $\frac{48 - 6}{48} = 0.88$

d) Structure of residuals[a]

Education	Age					
	16-24	25-34	35-44	45-54	55-64	65+
University	-	-	+	+	+	+
Secondary	·	·	·	·	·	·
Elementary	·	·	·	·	-	-

[a]For meaning of symbols, see Figure 4.4.

We can use this value for k and proceed with the analysis of the additive model, which now includes the multiplicative term indicated in equation (4.5). Figure 4.8(a) shows the results of this analysis, where the new residuals are easily computed from the residuals for the simple additive fit:

residuals from equation (4.5) = actual values − common − row effect − column effect − k (row) (column)/common
= residuals from additive model − k (comparison values)

In other words, the residuals for the multiplicative model of equation (4.5) are just the residuals from the simple additive model minus k times the comparison values. The row, column, and common values, of course, are the same in the model of equation (4.5) as they are in the additive fit.

Has the new fit improved our analysis at all? From Figure 4.8(c) we see that the goodness-of-fit ratio is 0.88, a slight decrease compared to the additive fit. The structure of residuals is changed only slightly over the simple additive fit (compare Figures 4.8(d) and 4.5). The dominant feature is still the poor fit of the model to the university-educated group.

A comparison of the residuals in the simple additive model with those in the more complex model containing a multiplicative term are interesting to note. The significant changes occur in those residuals for the group of people older than 65 years. For example, the simple additive model seemed to predict the (elementary, 65+) cell fairly well, but the more complex additive model overestimates the average earnings of these people. It may be that some mechanism is operating which acts to depress the earnings levels of aged persons with little education more than we would expect with our model. Similarly, those with a university education who are over 65 have higher earnings than expected with the model, a complete reversal from the situation with the simple additive model. With the similarity in goodness of fit of the two models, and the definite structure in the residuals of each, it is difficult to decide between them, and hence to place much significance on those findings that differ between the models.

3.2 *Data Transformations*

In Chapter 2 it was shown that transformations of data can often be a useful tool for the analyst. For example, we saw that when symmetry in the distribution of a batch of numbers was important, transformations with logarithms, square roots, or other functions could often achieve symmetry. Similarly, when comparing measures of central tendency between groups, it was important to have roughly equal measures of spread. Once again, it was often possible to achieve this by transforming the data.

With tabled measures of central tendency, like those discussed in this chapter, it is also often possible to get a better fit to the data by first transforming the tabled data. Only one such transformation will be discussed here, that for logarithms, because the interpretation with this transformation is still easy to understand. Other transformations are used less often, and only after careful consideration of the benefit from a better fit as opposed to a decrease in the ease of interpretation of the results. For a good discussion of

when such transformations might be indicated, see the work by Tukey listed at the end of the chapter.

How do we know that data in a table would be best analyzed after transformation? Unfortunately, a table does not tell us what particular model will provide the best analysis. We must start with one analysis and then evaluate the results. One can proceed with further analysis if that seems necessary given the fit of the model to the data, the amount of refinement deemed necessary, the decisions to be made from the analysis, and a variety of other situational factors. In the example of average earnings that we have dealt with in this chapter, neither the simple additive model nor the additive model with one additional multiplicative term seemed to fit the data very well. In particular, there was still a good deal of structure left in the residuals, with the university education level the main problem.

If we take a simple additive fit up to the point of making a diagnostic plot of residuals versus comparison values, it is possible to determine if a better fit could be achieved with a data transformation. In particular, if the computed value of k is close to $+1.0$, we will often find that a logarithmic transformation will provide a better fit to the data than the additive fit. If the slope is as high as 2.0 or as low as 0.5, other transformations are indicated (see Tukey, 1977, for a discussion).

If the diagnostic plot indicates that a logarithmic transformation might provide a better fit to the data than an additive model, the first step is to return to the original tabled data and to convert all the numbers to logarithms. For the average earnings example, the diagnostic plot in Figure 4.7 showed a slope of 1.5, indicating that a logarithmic transformation is worth trying. The logarithmic transformations are shown in Table 4.4.

The analysis at this point proceeds just exactly as with the simple additive fit, and the various steps through the polished stage are shown in Figure 4.9. Figure 4.10 shows that the goodness-of-fit ratio is again quite high, this time 0.93. Better still, much of the structure has been removed from the residuals, meaning that the logarithmic model has done a better job of predicting the data than the additive model.

An additional benefit from this analysis is that we can now clearly see a result that was only indicated in the additive model with a multiplicative term, and was totally obscured with the simple additive model. In fact, it *does* appear that people who have an elementary-level education and are over 65 years of age tend to have average earnings that are depressed from the levels we would expect from their age and education level alone. The residual for this group, -0.09, is the second largest in the table. This is a result worth tracking down in further analysis or discussion.

We saw that the logarithmic transformation has resulted in a better analysis of the earnings data, but we have not yet talked about how to interpret things like the row effects or the various column effects. Shown in equation (4.6) is the analysis model we have been fitting to the transformed data:

$$\text{log of actual data} = \text{log of common value} + \text{log of row effect} + \text{log of column effect} + \text{log of residual} \qquad (4.6)$$

Table 4.4:

Average Earnings Data, as Shown in Table 4.1, and the Logarithmic Transformations of the Data

a) Original data

Education	Age					
	16-24	25-34	35-44	45-54	55-64	65+
University	4502	11,023	16,146	17,442	16,475	8758
Secondary	4522	8283	9630	10,214	9798	5405
Elementary	3727	6331	7445	7934	6997	3260

b) Logarithmic transformations

Education	Age					
	16-24	25-34	35-44	45-54	55-64	65+
University	3.65	4.04	4.21	4.24	4.22	3.94
Secondary	3.66	3.92	3.98	4.01	3.99	3.73
Elementary	3.57	3.80	3.87	3.90	3.84	3.51

This is just a simple additive fit using the logarithms of the tabled data. But recall that the logarithm of a multiplicative factor is the sum of the logarithms. That is

$$\log(ab) = \log a + \log b$$

Consequently, the additive model in equation (4.6) can also be thought of as a multiplicative model, since it can be rewritten as:

$$\text{log of actual data} = \text{log of (common value} + \text{row effect} + \text{column effect} + \text{residual)}$$

This fact is useful in interpretation.

Shown in Figure 4.11 is the last part of the polished logarithmic additive fit from Figure 4.9, and the corresponding multiplicative table in which each of the log terms in part (a) has been converted back into its corresponding number. For example, the row effect for elementary education is log (−0.13) = 0.74. Similarly, the residual for the cell (university, 16-24) is log (−0.22) = .60. In part (b), these converted terms are shown, along with an indication that the actual values are now not the *sum* of the common value, row effect, column effect, and residual, but rather the *product* of those terms.

Figure 4.9:

A Two-Way Analysis of the Average Earnings Data in Table 4.1, Using Logarithmic Transformations

Education	Age																							
	16-24	25-34	35-44	45-54	55-64	65+																		
University	3.65	4.04	4.21	4.24	4.22	3.94	4.13	-.48	-.09	.08	.11	.09	-.19											
Secondary	3.66	3.92	3.98	4.01	3.99	3.73	3.95	-.29	-.03	.03	.06	.04	-.22											
Elementary	3.57	3.80	3.87	3.90	3.84	3.51	3.82	-.25	-.02	.05	.08	.02	-.31											
							3.95	-.29	-.03	.05	.08	.04	-.22	0	-.29	-.03	.05	.08	.04	-.22	3.95			
							.18	-.19	-.06	.03	.03	.05	.03	.03	-.22	-.09	0	0	.02	0	.21			
							0	0	0	-.02	-.02	0	0	0	0	0	-.02	-.02	0	0	0			
							-.13	.04	.01	0	0	-.02	-.09	0	.04	.01	0	0	-.02	-.09	-.13			
															0	0	0	0	0	0	0			
															-.22	-.09	0	0	.02	0	.21			
															0	0	-.02	-.02	0	0	0			
															.04	.01	0	0	-.02	-.09	-.13			
															-.29	-.03	.05	.08	.04	-.22	3.95			

Figure 4.10:

Examining the Residuals from the Two-Way Analysis With Logarithmic Transformations, as Shown in Figure 4.9

a) Original Data (unit = 0.01)

35·	1
35*	7
36·	
36*	56
37·	3
37*	
38·	04
38*	7
39·	024
39*	89
40·	14
40*	
41·	
41*	
42·	142
42*	

N = 18

M	391		
H	373	401	28
1	351	422	71

b) Residuals (unit = 0.01)

-2*	
-2·	2
-1*	
-1·	
-0*	99
-0·	222
+0·	000000000124
+0*	
1·	

N = 18

M	0		
H	-2	0	2
1	-22	4	26

c) Goodness of fit ratio = $\frac{28 - 2}{26} = .93$

d) Structure of residuals

Education	Age					
	16-24	25-34	35-44	45-54	55-64	65+
University	·	-	·	·	+	·
Secondary	·	·	·	·	·	·
Elementary	+	+	·	·	·	-

For example, we see that for the cell (secondary, 35-44), the actual value is given as:

actual value = common value + row effect + column effect + residual
8913 (1.0) (1.12) (.95) 9483

Figure 4.11:

Comparability of an Additive Log and a Multiplicative Model, for the Two-Way Analysis in Figure 4.9

a) Additive log model

Education	Age						Row Effects
	16-24	25-34	35-44	45-54	55-64	65+	
University	-.22	-.09	0	0	.02	0	.21
Secondary	0	0	-.02	-.02	0	0	0
Elementary	.04	.01	0	0	-.02	-.09	-.13
Column Effects and Common	-.29	-.03	.05	.08	.04	-.22	3.05

b) Multiplicative model corresponding to additive log model in (a)

Education	Age							Row Effects
	16-24	25-34	35-44	45-54	55-64	65+		
University	.80	.91	1.0	1.0	1.05	1.0		1.62
Secondary	1.0	1.0	.95	.95	1.0	1.0	x	1.0
Elementary	1.16	1.02	1.0	1.0	.98	.91		.74
	x						x	
Column Effects and Common	.51	.93	1.12	1.23	1.14	.93	x	20.13

Note that the value of 9.13 differs from the original value of 9.30 because throughout the logarithmic analysis the log terms were rounded off to two decimal places, and this rounding creates the discrepancy. In any event, the multiplicative table (Figure 4.11b) is easy to interpret. The interpretation goes something like this:

Common value: the common value as found with the logarithmic model is 8.8913, and the interpretation is the same as with the

additive model. It is a term indicating the mid-range of average earnings for all age and education groups.

— *Row effects*: Again the interpretation of row effects is done through comparison of any two rows. For example, we see that the row effect for elementary education is 0.74, while that for secondary school is 1.0, indicating that people with elementary education earn on average about three-fourths what those with a high school education earn. Similarly, those with a university degree earn over 1.5 times what those with secondary degrees earn.

— *Column effects*: As with row effects, column effects are interpreted by comparing them with each other. For example, we see that those aged 65 and over earn about half of those aged 45-54, since the ratio of those column effects is (0.60)/1.20 = 0.5.

— *Residuals*: The residuals shown in Figure 4.11(b) indicate how far the actual terms are from the predicted terms. If the residual is 1.0, the actual and predicted terms are identical. If the residual is less than 1.0, the actual term is less than that predicted, with the ratio of actual to predicted being given by the residual. For example, the residual for the cell (university, 16-24) is 0.60, indicating that the actual value is only 60% of the predicted value. In fact, our model predicted too high a number for that cell. For the (university, 16-24) cell, the predicted value is:

$$\text{Predicted (university, 16-24)} = 8913 \times 1.62 \times .51 = 7364$$

On the other hand, the actual cell value was 4502. The ratio of actual to predicted is just:

$$\frac{\text{Actual}}{\text{Predicted}} = \frac{4502}{7364} = 0.61$$

Within the limit of the errors involved in rounding off, this is the same as the residual of 0.60 shown in Figure 4.11(b). If the residual is greater than 1.0, the actual value is greater than the predicted value, and the ratio of actual to predicted is indicated by the size of the residual.

In summary, then, it appears that the logarithmic transformation of the data has resulted in a better prediction of the data, and consequently in a model on which we could more reliably base decisions. From the analysis in Figure 4.11(b), we can make several useful statements, as already indicated, about the relative earnings of various age or education groups. We also know that for a few groups something else seems to be occurring. In particular the (elementary, 65 and over) and (university, 16-24) and (university, 24-34) groups were not predicted very well in the analysis. We may wish to find out why. The analysis has been useful in the sense that it has indicated particular groups where interesting mechanisms may be operating that result in earnings levels differing from what our model has predicted based only on age and education.

4.0 *Summary*

In this chapter we have looked at a series of techniques and procedures that are useful when the analyst wishes to dredge more information from a table of central tendency than is possible with just a visual analysis. Initially we looked at a simple additive fit to the data and emphasized the interpretation of row and column effects and residuals. It was demonstrated that a second iteration of the additive fit will provide a "polished" fit converging on stable effect estimates. It was also shown how the goodness-of-fit ratio and the structure of the residuals could be used to determine whether the additive model predicted the actual data very well.

If the prediction was not good, either because the residuals were large or because there was structure in the residuals, there are two additional procedures that can be taken. A multiplicative, or interaction, term can be included in the basic additive model. A diagnostic plot of residuals versus comparison values can be used to determine the constant to use in fitting the data. An additive model with a multiplicative term assumes that the effects due to row and column variables are not simply additive. This model hypothesizes that there are some cells in which interactions of the row and column variables are present.

Finally, it was shown that if the simple additive model does not provide a good fit to the data, the diagnostic plot can be used to decide if a data transformation will help. In particular we looked at the logarithmic transformation, again emphasizing the interpretation of the results.

5.0 *Terms for Review*

simple additive fit
common value
row effect
column effect
residual
polishing
goodness-of-fit ratio
structure in residuals
interaction
comparison values
diagnostic plots
slope of a line
additive logarithmic fit

6.0 *Practice Exercises*

1. Shown in Table 4.5 are average salaries by school-district size and enrollment trend for the North Central data set. What does an analysis of the data by inspection tell us? Perform a two-way analysis, fitting a simple additive model to the data. Carry the fit through one polishing iteration. Comment on the common value, the effects due to enrollment trend, and the effects due to district size.

Table 4.5:

North Central Average Salaries, by School-District Size and Enrollment Trend

Enrollment Trend	Size			
	0-299	300-799	800-1799	1800+
Fast Growth	257	269	284	298
Slow Growth	282	262	276	310
Stable	263	274	289	328
Slow Decline	261	276	294	334
Fast Decline	259	274	310	370

2. Compute a goodness-of-fit ratio for the polished fit in question 1. Also examine the structure of the residuals. Comment on how well the simple additive model has fit the data.

3. Compute comparison values and plot residuals versus these values. Estimate the slope of the line which best passes through the cluster of data points. Does it indicate that a logarithmic transformation of the data might provide a better fit to the data?

4. If you think a logarithmic transformation of the data would provide an improved fit to the data, do such an analysis. If a logarithmic transformation is not indicated, use the slope found in question 3 to help fit an additive model with an interaction term. Whichever model you decide to fit, carry the analysis through one polishing iteration, compute the goodness-of-fit ratio, examine the structure of the residuals, and interpret the results of the analysis.

Footnotes

[1]One advantage of using medians is that it is easily done where hand calculations must be performed. An additional advantage is that observations which are far from the summary will be untouched and hence show up clearly in the exposure part of the analysis when residuals are examined. Using means will tend to take any large deviations from the summary and spread this deviation around so that examination of residuals is not nearly so easy.

[2]The more refined fitting techniques in Chapter 5 would yield a slope somewhat higher than 1.5, but for illustrative purposes in this chapter the fit to the data in Figure 4.7 is sufficient.

Bibliography

Iverson, Gudmund R., and Helmut Norpoth. "Analysis of Variance." Beverly Hills: Sage, 1976. This monograph provides an easy introduction into analysis of variance, the counterpart technique to the two-way analysis techniques discussed in this chapter. Analysis of variance is generally used when inference from a sample to the population is of concern.

McNeil, Donald R. *Interactive Data Analysis*. New York: John Wiley, 1977. There is a useful discussion of two-way analysis in this book, with computer program listings in FORTRAN and APL for performing many of the procedures discussed in this chapter.

Mosteller, Frederick, and John W. Tukey. *Data Analysis and Regressio* Reading, Mass.: Addison-Wesley, 1977. This, and Tukey's book listed below, constitute the most thorough discussion available about two-way analysis and its extensions and complications.

Tukey, John W. *Exploratory Data Analysis*. Reading, Mass.: Addison-Wesley, 1977. There are several chapters in this book which describe in detail the range of procedures that are useful for two-way analysis. There is also a good discussion of the extension to three-way analysis, where three independent variables form the classifications in the table.

CHAPTER 5

Regression Analysis

Gary Lewis

1.0 *What's Ahead?*

Regression analysis has been called one of the most powerful procedures available to data analysts. Applied to interval level data, it is a technique that uses many of the same concepts developed in previous chapters. Data are first graphed, and then a model is used to try to fit a straight line or some other curve form to the set of data. (See Volume II, Chapter 7). The extent to which the model predicts the observed data is the basis for assessing the quality of the fit. If the fit is poor, the analyst can try data transformations to help correct some of the problems. All of these features — the plot of the actual data, the use of a model to predict the data, the use of residuals from the model to assess quality of fit, and steps to correct problems if the fit is poor, have been covered in past chapters.

In the present chapter, we discuss two variations of regression analysis, called simple and multiple, the latter being an extension of the former in that it includes more variables. Initially, the basic idea of regression analysis is discussed, with attention to the form of the model most often chosen, this being the linear model (a straight line). Secondly, various methods of fitting a straight line to a scatter of plotted points are described. These methods range from the simple to the more complex, and their choice for use depends upon the situation in which the analyst is operating. Thirdly, the quality of the fit of the model to the data is assessed by using residuals, where once again we are concerned about things like structure in the residuals. A fourth section describes how data transformations can sometimes alleviate the more common data problems. Simple guidelines for deciding upon an appropriate transformation are presented, and examples of such transformations are provided. The fifth and final section discusses one very in.portant extension of simple regression analysis to the case where we wish to relate one dependent variable to two or more independent variables. In this situation, multiple regression is indicated, and procedures for selecting a model, examining residuals, and transforming data are included. Since the purposes of simple and multiple regression vary, and can be abused by unwarranted interpretations of what the analysis seems to be saying, considerable effort is made when presenting examples to explore the limitations of the analysis as well as the benefits.

2.0 *The Basic Idea of Regression Analysis*

Table 5.1 offers a summary of admissions information for an institution of

higher education over a period of 7 years. It is a reasonable assumption that the number of new students who actually enrolled each year was dependent upon the number of students who were admitted for study. The data for those enrolled and admitted are graphed in Figure 5.1.

Table 5.1:

Institutional Admissions Summary, 1965-1971

Year	Applications	Admissions	Enrolled
1965	1178	604	416
1966	1097	648	504
1967	1006	558	426
1968	1078	630	447
1969	1166	595	445
1970	1095	581	411
1971	1059	694	511

It is clear from Figure 5.1 that there is a relationship between the number of those admitted and the number enrolled. We see from the scatter of points that, generally, the larger the number of admissions, the larger the number of students actually enrolling. In other words, the direction of the relationship is positive: low values of one variable are generally associated with low values of the second variable, and high values of one variable are usually associated with the larger values of the second variable.

Recall that in Chapter 3, when discussing tables of counts, we were interested in the directionality of a relationship, but were also interested in the strength and nature of that relationship. When we are dealing with interval level data, as in Figure 5.1, it is possible to fit to the scatter of points a straight line or some other functional form. The functional form tells us something about the nature of the relationship. For example, if a straight line seems to fit the data, we say that there is a linear relationship between the variables. If the data points do not lie more or less on a straight line, but give evidence of curves or bends, then some other functional form must be used to fit the data. For example, we might find that as X increases, Y also increases, first slowly and then more rapidly, as in Figure 5.2. In such a case an exponential form might provide the best fit to the data. An alternative to fitting such a functional form would be to transform the data for either or both of the variables, attempting to straighten out the data and to fit the resulting points with a line. Such data transformations are discussed later in the chapter.

Figure 5.1:
A Scatter Plot of the Number of Admissions and the Number of Students Enrolled, for Data in Table 5.1

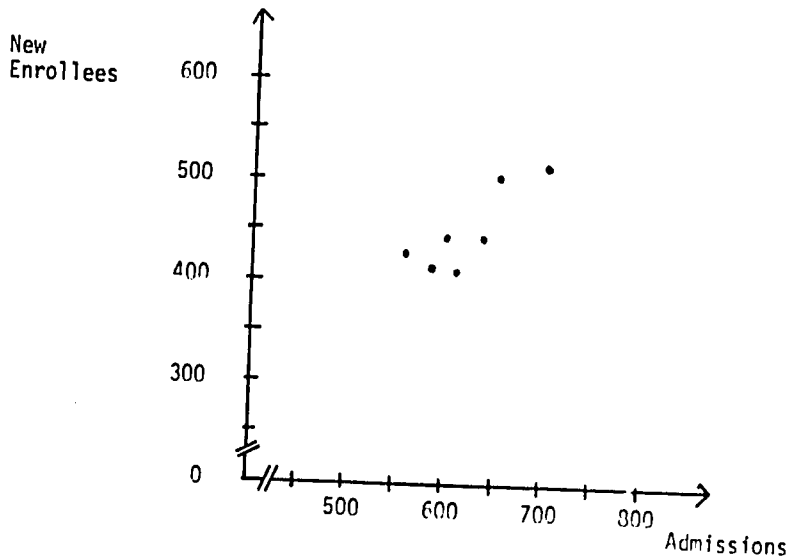
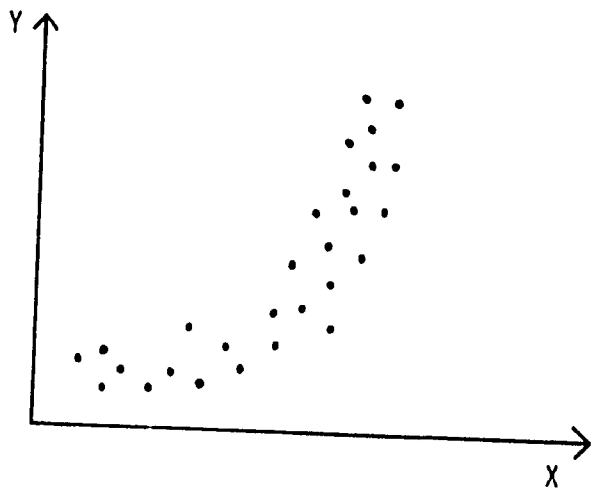


Figure 5.2:
An Example of a Non-Linear Scatter of Data Points

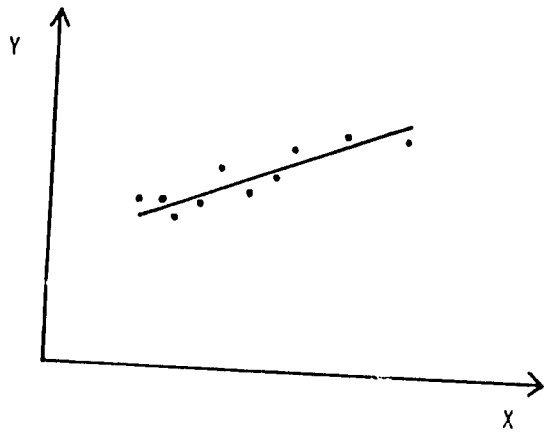


Besides telling us something about the nature of the relationship between the variables, the fit of a line to the data can also tell us something about the goodness of the fit of the line and the strength of the relationship. If, for example, the observed data points lie fairly close to the line fitted to them, the line accurately describes the nature of the relationship between the variables and indicates that the fit is a pretty good one. Such a situation is shown in Figure 5.3(a). In this case we say that the strength of the relationship is strong. When the data points do not lie close to the line, the strength of the relationship is weaker, as in Figure 5.3(b).

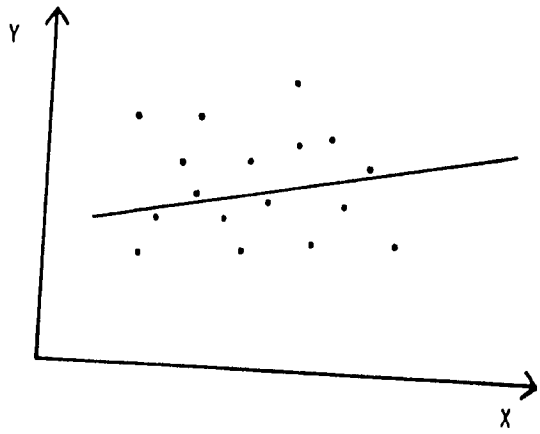
Figure 5.3:

Varying Degrees of Goodness of Fit and Strength of Relationships

a) Good fit of line to data; strong linear relationship



b) Poor fit; weak linear relationship



Fitting a line or some other functional form to graphed interval data is called regression analysis. From this introductory description it should come as no surprise that the basic approach to the analysis of such data is quite similar to approaches used in previous chapters. We will first propose a model, and then use the model to compute expected values. A comparison of the expected and the observed values is used to determine if the model did a good job at fitting the data. A close fit to the data is important when it comes time to interpret the results from the analysis.

This same "summarize and expose" technique was used with measures of association and prediction analysis for tables of counts in Chapter 3, and with the two-way analysis of tables of means in Chapter 4. The form of the model varies according to the type of data being analyzed, but the basic summarization and exposure approach is used throughout.

With regression analysis, the basic "model" applied to the data is the straight line. As indicated above, there are refinements to this basic model that become necessary when the data are not well fitted with a line, but as a first approximation the linear form is generally used. Before proceeding to the mechanics of fitting a line to a set of points, we first must learn something about lines. They are defined by two parameters, the *slope* and the *intercept*. The slope of a line indicates how flat or steep the line rises or falls. It is defined as:

$$\text{slope of a straight line} = \frac{\text{change in } Y}{\text{one unit change in } X} \qquad (5.1)$$

Shown in Figure 5.4 are two examples, one where the slope is quite steep and positive, and the other where the slope is fairly flat and in a negative direction. Note that the sign on the slope indicates directionality of the relationship. In Figure 5.4(b), for example, there is a tendency for the smaller X values to be associated with larger Y values, and for larger X values to be associated with smaller Y values.

The *intercept* is the second parameter needed to uniquely identify a single straight line. It is defined as that point where the line intercepts the Y-axis, or would intercept the Y-axis if it were extended to that axis. Figure 5.5 shows a variety of lines, all with the same slope but with varying intercepts.

If we are going to fit a line (our "model") to a set of data points, we must determine both the slope and the Y-intercept, since they uniquely define a given straight line. In general terms the equation of a straight line is given by:

$$Y = a + bX, \qquad (5.2)$$

where a is the Y-intercept and b is the slope. The dependent variable is Y, and the independent variable is X in this equation.

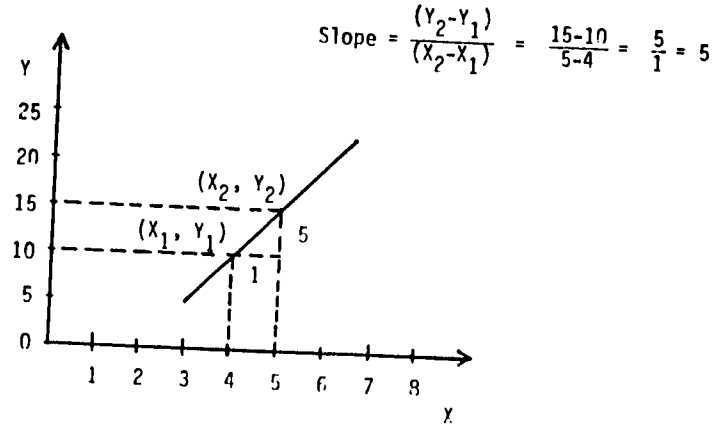
3.0 *Fitting Procedures*

How does one determine which line "best" fits the data? There are several methods for fitting a line to a set of points, from the fairly simple to the somewhat more complex. Before detailing these methods, however, we

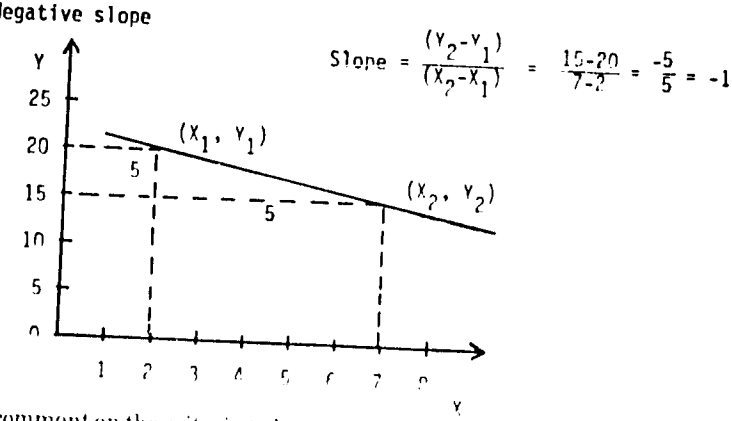
Figure 5.4:

Computation of Slopes

(a) Positive slope



(b) Negative slope



should comment on the criteria to be used when judging whether a fit to a set of data is good or poor. Intuitively, the closer we can pass the line to the points, the better the fit. Figure 5.6 shows two lines passing through the same set of points, and it is easy to see that line 1 seems a better fit than line 2, since the data points lie closer to line 1 than they do to line 2. In a few moments we will work out in a more rigorous fashion how this criterion of closeness of the data points to the line can be used to select a fitted line. For the present, however, it is sufficient to recognize that the closer a line passes to all the data points, the better the fit. By using this criterion, it is possible to select a line using the simplest of the approaches to be discussed here.

Figure 5.5:

Examples of Straight Lines with the Same Slope but Varying Y-Intercepts.

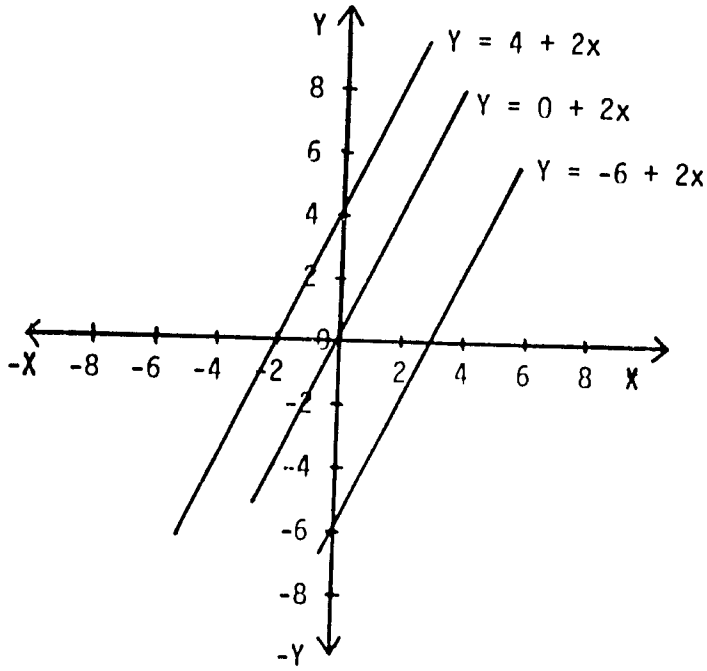
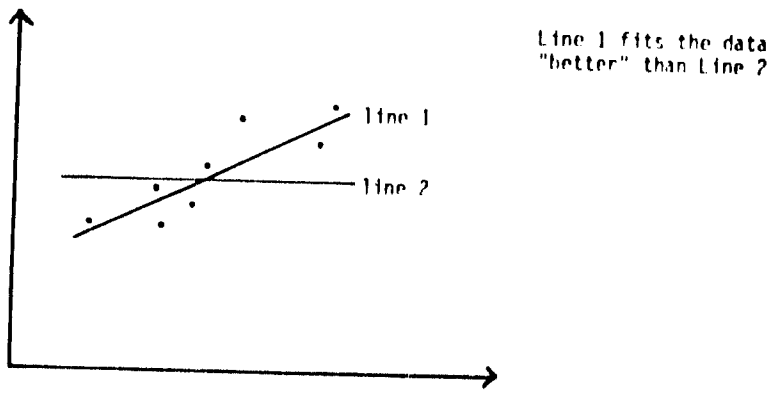


Figure 5.6:

An Illustration of Goodness-of-Fit, for a Linear Model



3.1 *The Inspection Fit*

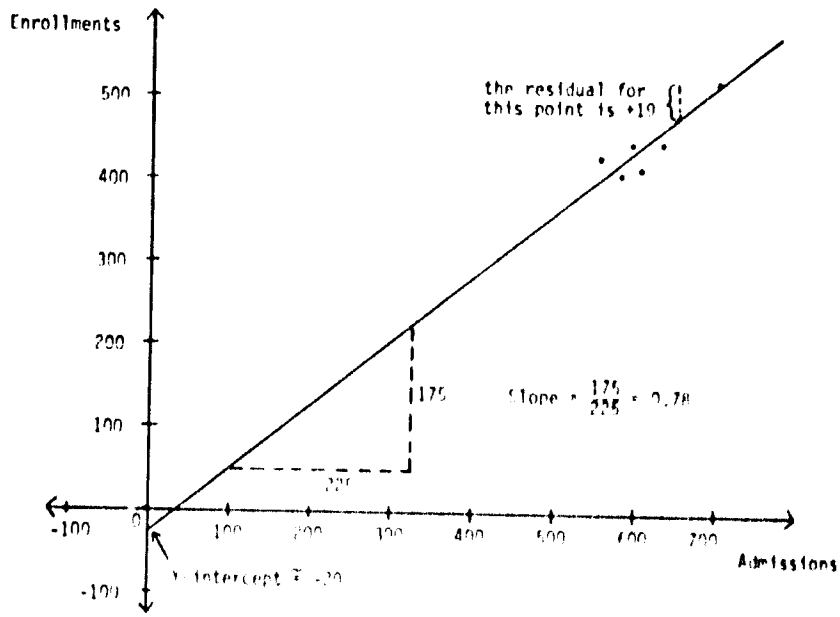
The quickest and simplest of fitting procedures uses the human eye and a clear ruler or straight edge. First, plot the data points on a piece of graph paper. Secondly, move the straight edge through the points until you reach an orientation where it looks as if the data points are closer to the straight edge than for any other orientation. Select this as your line and draw the line through the data points. Find the slope by measurement from the graph. Find the intercept as that point where the line passes through the Y-axis or would pass through the Y-axis if it were extended.

This procedure has been used to fit a line to the admissions and enrollment data of Figure 5.1. Shown in Figure 5.7 is the fit to the data, where we see that the equation of the line has been found to be.

$$Y = 20 + .78X$$

Figure 5.7:

An Inspection Fit of a Straight Line to a Plot of Enrollments and Admissions, for the Data in Table 5.1



Fitted Line is: Y = -20 + 0.78X,

where Y = Enrollments

and X = Admissions

Such an equation provides a *summary* or a *description* of the relationship between X and Y. We see, for example, that because the slope is positive, X and Y values generally increase together. Furthermore, the magnitude of the slope tells us that for an increase of 100 in the number admitted, there are generally 78 people enrolled. There are two qualifications to be made on this summary statement, however. First, it is only true over the data ranges involved. In the present case, we are dealing with admissions which range from the mid-500's to almost 700. Equation (5.7) summarizes the data we have in this range. It says nothing, however, about the (nonexistent) data for admissions larger than 700 or lower than the mid-500's. In other words the equation is a summary of the data points themselves, not of any extensions of the line to lower or higher values.

Secondly, the summary is an accurate statement of the relationship only to the extent that the line fits the data. If the data points are not close to the line, then we have less confidence in the summary than we would if the data were close to the fitted line. Methods of assessing this goodness-of-fit will be discussed in later sections of this chapter.

As we have just seen, one purpose of regression analysis is to obtain a summary of the relationship between two variables. *Prediction* is the second purpose for which regression analysis can be used. We could use the equation to predict the number of students actually enrolled, for a given number of students admitted. How well, for example, does the equation predict the actual number of enrollees in 1966? In that year there were 648 students admitted. We find that the equation for the fit by inspection predicts or estimates that 485 students would enroll:

$$Y = -20 + 0.78(648) = 485$$

In actual fact there were 504 students who enrolled in 1966, so the prediction from the equation was off by about 19 students (see Figure 5.7). This number, which is the deviation of the predicted Y value from the observed Y value, is called a *residual*. Residuals can be computed for each of the observed Y values. In a later section we will show how these residuals permit us to assess the goodness-of-fit of the data.

It is also possible to use the regression equation to predict Y values for a given X value that is not among the data points used to select the line. Suppose that the data we plotted was not for seven consecutive years. If we knew the number of students admitted for one of the missing years, we could use the regression equation to estimate the number of students who actually enrolled. This is a somewhat contrived use of regression analysis for prediction, because we are not likely to know the number admitted without knowing the number enrolled.

What about using the equation to predict the number of students we would expect to enroll if we admit a specified number of students? For educational planners this could be of much use, and was discussed under the rubric *projection* in Volume II, Chapter 7. Technically, we should not use the regression equation to predict to some future state, such as the number of enrollees to be expected if we admit 620 students next year. Regression is based on past

data and, strictly speaking, is valid for prediction of data within the range of values upon which the equation is based. But *if* we believe that circumstances have not changed recently to affect the dependent variable Y, then we can and do use regression as prediction, where prediction is to some future state. Similar projections are used constantly in planning.

3.2 *The Median Fit*

The median fit can be used to provide a somewhat more accurate fit to the data points, in the sense that the line determined with this method will generally be closer to the data points than one chosen by inspection. The series of steps needed to complete a median fit are listed below:

1. On the basis of X values, divide the data points into three mutually exclusive groups so that there are roughly the same number of points in each group.
2. Compute the medians for both the X and the Y values for the group with the smallest X values and for the group with the largest X values. Call these values (X_s, Y_s) and (X_l, Y_l), where the *s* refers to small and the *l* refers to large.
3. Use (X_s, Y_s) and (X_l, Y_l) to compute the slope of the line with the following equation.

$$\text{Slope} = \frac{(Y_l - Y_s)}{(X_l - X_s)} \tag{5.3}$$

4. For each X value, compute an expected Y value (called Y') using the following equation:

$$Y' = (\text{slope})(X) \tag{5.4}$$

5. Compute residuals as: residual $= Y - Y'$, and then find the median residual. This value is the Y intercept. Alternatively, one can draw a line on the plot of the original data points, with the line having the computed slope in step 3. Then move the straight edge parallel to this line until half of the data points are on one side of the straight edge and half are on the other. At this point draw the regression line through the data points. Extend the line to the Y axis to determine the Y intercept.

These steps have been worked out in Figure 5.8 for the admission and enrollment data that we have been using to this time. Note that this procedure produces the following equation for the regression line through the data:

$$Y = 100 + 0.88X$$

The interpretation of this equation, of course, proceeds just as for the fit by inspection. It may be used for both a summary description and for prediction.

3.3 *The Least Squares Fit*

Another commonly used procedure to fit a line to a set of data points is called the least-squares fit, which is introduced in Volume II, Chapter 7. The line determined by the least-squares fit minimizes the sum of the squares of

Figure 5.8:

A Median Fit to the Admissions and Enrollment Data of Figure 5.1

a) Rank order X values (admissions) and divide into three groups (remember to keep Y values, on enrollments, with appropriate X values)

Groups	X	Y
1	558	426
	581	411
2	595	445
	604	416
	630	447
3	648	504
	694	511

b) Compute median values for groups 1 and 3

Groups	X_m	Y_m
1	570	419
3	671	508

c) Compute slope

$$\text{Slope} = \frac{(Y_{m3} - Y_{m1})}{(X_{m3} - X_{m1})} = \frac{(508 - 419)}{(671 - 570)} = 0.88$$

d) Determine Y intercept, by computing residuals = Y - Y'

X	Y	Y' = 0.88 X	Y - Y'
558	426	491	-65
581	411	511	-100
595	445	524	-79
604	416	532	-116
630	447	554	-107
648	504	570	-66
694	511	611	-100

Median Residual = -100

the residuals. Recall that a residual is just the vertical distance, on the graph, separating the observed Y value from the expected Y value on the regression line (see Figure 5.7). Minimizing the sum of the squares of the residuals will produce a line that is as close as possible to all points. The least-squares fit also results in other attractive features that statisticians worry about, but will not concern us here.

The least-squares fit is somewhat more time-consuming to perform than either of the two previously discussed fitting procedures. The steps are outlined below, and, while simple, they can become time consuming when there are many data points. The computer can help in this regard if one is available, since there are many packaged programs which routinely compute the least-squares fit to a set of data points. Alternatively, many inexpensive hand calculators now have been programmed to provide a least-squares fit to a set of data. The data are simply entered into the calculator and a few buttons pressed and the slope and intercept are then computed and displayed.

For hand computation of a small number of data points, however, the following steps may be used to yield a least-squares fit:

1. Sum the X values. Call this ΣX.
2. Compute the mean of the X values. Call this $\bar{X}$.
3. Sum the Y values. Call this ΣY.
4. Compute the mean of the Y values. Call this $\bar{Y}$.
5. For each X value, compute X^2. Sum these X^2 values and call this ΣX^2.
6. For each (X,Y) pair, compute XY. Sum these XY values and call this ΣXY.

After these terms have all been obtained, one computes the slope and intercept of the least-squares fit as:

$$\text{slope} = \frac{N\Sigma XY - \Sigma X \Sigma Y}{N\Sigma X^2 - (\Sigma X)^2} \tag{5.5}$$

where N is the number of data points;

$$\text{intercept} = \bar{Y} - (\text{slope})(\bar{X}) \tag{5.6}$$

Figure 5.9 shows the computations necessary to produce the least-squares fit to the admissions and enrollment data discussed previously. The least-squares regression equation is:

$$Y = 29 + 0.78X$$

The simple inspection fit to the data was remarkably close to the least-squares fit. This is nice, but unfortunately it does not always occur. Of the three procedures, there is an increase in the amount of effort as we go from the fit by inspection to the median fit to the least-squares fit. Which of the three is to be used depends on the purposes, resources, and time of the analyst. What may be appropriate in one situation may be inappropriate in another situation.

4.0 *Examination of Residuals*

The regression equation can be thought of as a model which describes the

Figure 5.9:

Least-Squares Fit to the Admissions and Enrollment Data of Figure 5.1

a) Compute necessary terms

	X	Y	X^2	Y^2	XY
	604	416	364,816	173,056	251,264
	648	504	419,904	254,016	326,592
	558	426	311,364	181,476	237,709
	630	447	396,900	199,809	281,610
	595	445	354,025	198,025	264,775
	581	411	337,561	168,921	238,791
	694	511	481,636	261,121	354,634
Sum	4,310	3,160	2,666,206	1,436,424	1,955,374
Mean	616	451			

b) Compute slope

$$\text{slope} = \frac{N(\Sigma XY) - (\Sigma X)(\Sigma Y)}{N(\Sigma X^2) - (\Sigma X)^2} = \frac{7(1,955,374) - (4310)(3160)}{7(2,666,206) - (4310)^2} = \frac{68,018}{87,342} = 0.78$$

c) Compute Y-intercept

$$\text{Y-intercept} = \bar{Y} - (\text{slope})(\bar{X})$$

$$= 451 - (0.78)(616) = -29$$

d) Regression equation: $Y = -29 + 0.78X$

data. For any given X value, a Y can be computed and compared to the actual Y values. As stated in the previous section, the difference between an actual and predicted value is called a residual. It is the residuals which tell us whether or not the model we have chosen has done a good job of prediction. If the residuals are small and without noticeable structure, the model is said to fit the data well. If the residuals are large, or there is structure to the residuals, the model does not fit the data well.

There are several different things that can be done with regression residuals, and no regression analysis is really complete without doing at least the most basic of these, which is simply to compute the residuals and plot them against the X values.

4.1 *A Plot of Residuals*

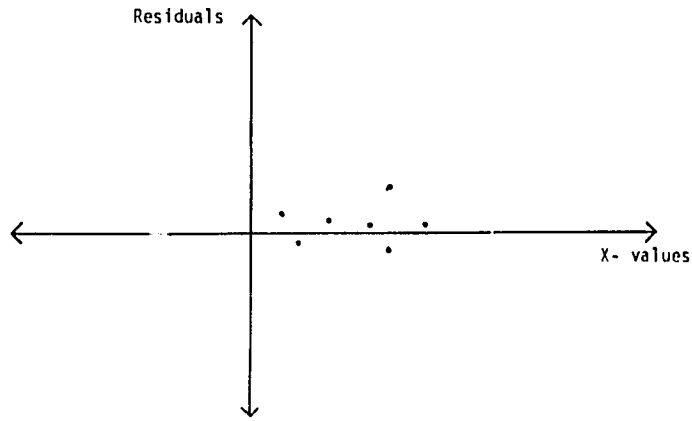
Computing residuals and plotting them against X values accomplishes several things. First, it allows us to assess whether the linear form appears to be a good fit to the data. If the linear form is a good fit, we should see a fairly random pattern to the residuals (no structure), as shown in Figure 5.10(a). If

there is structure in the residuals, as in Figure 5.10(b), the linear form does not provide a good fit to the data. In such circumstances we might want to think about transforming the data in such a way that a straight line would then fit the data.

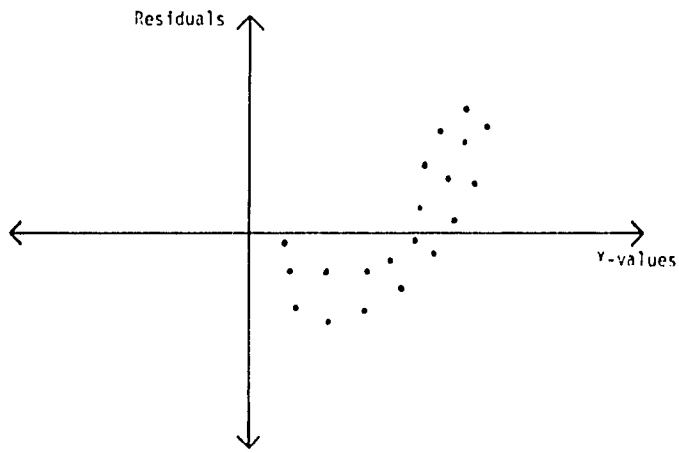
Figure 5.10:

Structure in Regression Residuals

a) Structure absent; linear form provides a good fit



b) Structure present; linear fit is poor



A plot of residuals can also be very useful for identifying *outlier* values. These are data points where the model has done a particularly poor job of predicting the observed values. Outlier values are interesting, particularly if the model seems to provide a good fit to most of the data points. In such circumstances outliers identify situations which may warrant further attention. For example, it may be that an outlier value indicates there has been severe measurement error, or that there was an analyst blunder caused by writing down the wrong number. However, an outlier might also indicate that this particular data point is uniquely different from the others, and that there may be different processes at work in this situation compared to the situations that gave rise to the other data points. This may or may not be of interest to the analyst, but it is worth the time to identify such outlier values and to reflect upon possible reasons for their appearance.

Figure 5.11 shows a plot of the residuals from the least-squares regression of enrollment on admissions. This plot indicates that the linear form provides a reasonable fit to the data. There does not appear to be any noticeable structure to the residuals. However, the two largest residuals happen to occur for the two earliest years in the sequence of years for which the data were collected. This might indicate that there was something unusual about those two years, and it might be worth checking to see if there was some change in admission procedures or some external factors influencing enrollments that might have led to these results. Since these two residuals do not appear to be widely out of line with the other residuals, (that is, they are not true outlier values), we would probably be content with the linear model if our investigation of those two years turned up nothing unusual.

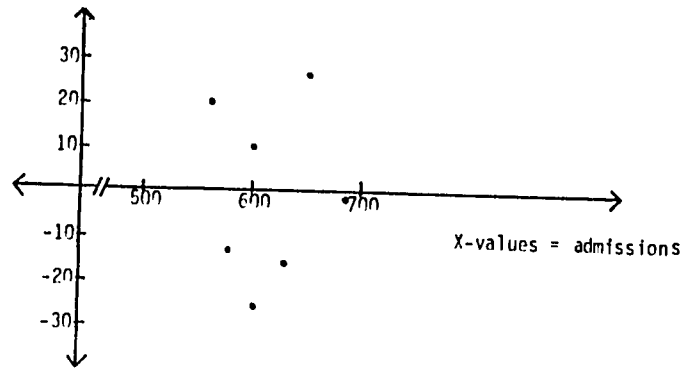
Figure 5.11:

Residuals from the Regression of Enrollments on Admissions, for the Data in Figure 5.1

a) Compute residuals; Y' = -29 + 0.78X

Year	Admissions	Enrollment	Predicted	Residuals
	X	Y	Y'	Y-Y'
1965	604	416	442	-26
1966	648	504	477	27
1967	559	426	406	20
1968	639	447	463	-16
1969	595	441	436	10
1970	581	411	424	-13
1971	694	511	512	-1

b) Plot of residuals



5.0 *Measure of Association*

Just as with nominal and ordinal data discussed in previous chapters, it is possible to develop measures of association between interval variables. One such measure, the Pearson product-moment correlation coefficient, is closely related in meaning to the regression lines just discussed. For interval level data we would interpret a strong association between two variables X and Y as occurring when the values of both variables tend to increase (or decrease) together. A weak association between the variables occurs when caange in one variable is not related to change in the second variable, when changing X does not tend either to increase or decrease Y

When the points cluster closely to the line, and there is some slope to the line, then there is a strong association between the X and Y variables. When, however, the points are scattered loosely about the regression line, or there is no slope to the line, then intuitively we feel that the association between X and Y is weak.

The Pearson product-moment correlation coefficient, or simply the correlation coefficient, is a measure of association which summarizes the strength of a *straight-line* relationship between two variables. Using the same terminology as for the least-squares fit of the previous section, the correlation coefficient r, is determined by the following equation:

$$r = \frac{N[\Sigma XY] - (\Sigma X)(\Sigma Y)}{[N(\Sigma X^2) - (\Sigma X)^2][N(\Sigma Y^2) - (\Sigma Y)^2]} \qquad (5.7)$$

Using the numbers from Figure 5.9, where a least-squares fit was made to the plot of enrollments versus admissions, we find that the correlation between enrollments and admissions is:

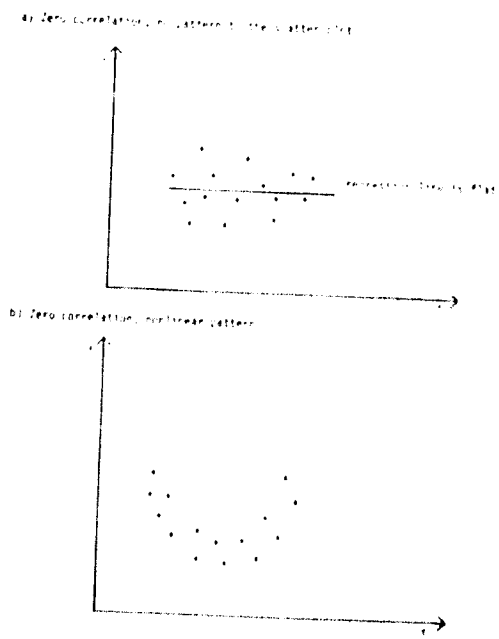
$$r = \frac{7(1{,}955{,}374) - (4310)(3160)}{[7(2{,}666{,}206) - (4310)^2][7(1{,}436{,}424) - (3160)^2]} = .87$$

How is this correlation coefficient of 0.87 to be interpreted? The correlation coefficient has a range of -1.0 to $+1.0$, with either extreme indicating a perfect association in the sense that a straight line fitted to the data would

pass through all the data points. The sign of the correlation coefficient is interpreted as are the signs on measures of association for ordinal data: a negative sign indicates that, as values of one of the variables tend to increase, the values on the other variable tend to decrease. A positive sign indicates that both variables tend to change in the same direction. A value of zero for a correlation indicates that there is no straight-line association between the variables. That does not necessarily mean that there is no regular pattern or relationship between the variables, only that there is not a straight-line relationship. Both scatter plots in Figure 5.12 would produce a correlation coefficient near zero, although there is clearly a very marked relationship between the variables in Figure 5.12(b); it just is not a straight line. A correlation of 0.87, as above, indicates that there is a fairly strong linear association between admissions and enrollments, with the positive sign indicating that as admissions tend to go up, enrollments also tend to increase.

Figure 5.12:

Patterns of Scatter Plots and the Lack of Correlation



One very nice feature of the correlation coefficient is that it can be shown, with a little algebra, that the square of the correlation coefficient gives an indication of the amount of variation in the Y values that is accounted for by the regression line. If we use the regression line to compute predicted Y values, it turns out that the square of the correlation coefficient is related to the variance of the predicted and observed Y values:

$$r^2 = \frac{\text{variance of predicted Y values}}{\text{variance of observed Y values}} \tag{5.8}$$

Because the predicted Y values must lie on the fitted line, their variation is always less than the observed Y values, unless the two sets of values coincide. Thus r^2 will vary between zero and one, and it indicates the amount of variation in the dependent variable that is "accounted for" by the regression line. With the enrollment and admissions example, $r^2 = (0.87)^2 = 0.75$. In other words, we have *explained* 75% of the variation in enrollments with the fitted regression equation.

Explaining or accounting for variance in the dependent variable does *not* mean that we have located the causal mechanism operating, only that there is an association between the two variables. A high r^2 value, as in our example, means that our model (i.e., regression line) has done a good job in fitting the data. This is just another way of stating that the observed Y values lie close to the fitted regression line.

Figure 5.13 shows that two lines with the same slopes and identical X values can still result in different correlation coefficients. In fact, it can be shown that the slope b, the correlation coefficient, and the standard deviations of the two variables in question (S_Y and S_X) can be related through the following equation:

$$b = r\frac{S_Y}{S_X} \tag{5.9}$$

The closer the fit of points to the line, the higher the correlation coefficient, and hence r^2 will be higher.

6.0 *Data Transformations*

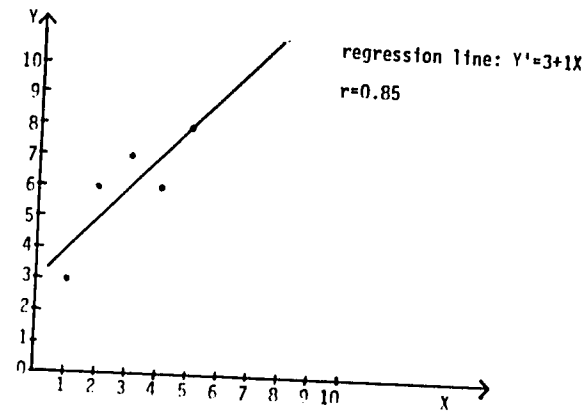
With the enrollment and admissions data that have been used thus far, a straight line seemed to fit the data reasonably well. But that is, of course, not the case for all sets of data. Shown in Figure 5.14 is a plot of an index called the Physical Quality of Life Index (PQLI) against the per capita Gross National Product for 45 countries in Africa. The PQLI, developed by the Overseas Development Council in Washington, DC, is designed to measure the effectiveness of social services. It includes information on life expectancy, literacy, and infant mortality. The minimum value possible is 0, and the maximum is 100. One would expect that as a nation's per capita GNP increased, more attention would be given to social services, and hence the PQLI would increase.

Figure 5.13:

An Illustration of the Change in the Correlation Coefficient With a Change in the Standard Deviation of the Dependent Variable

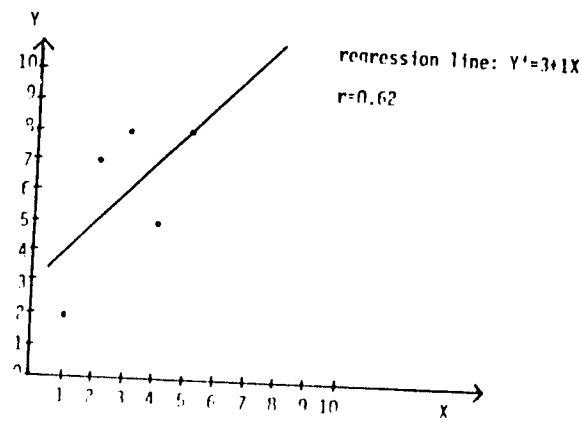
a) Data Set 1 (S_y = 1.7)

X	Y
1	3
2	6
3	7
4	6
5	8



b) Data Set 2 (S_y = 2.3)

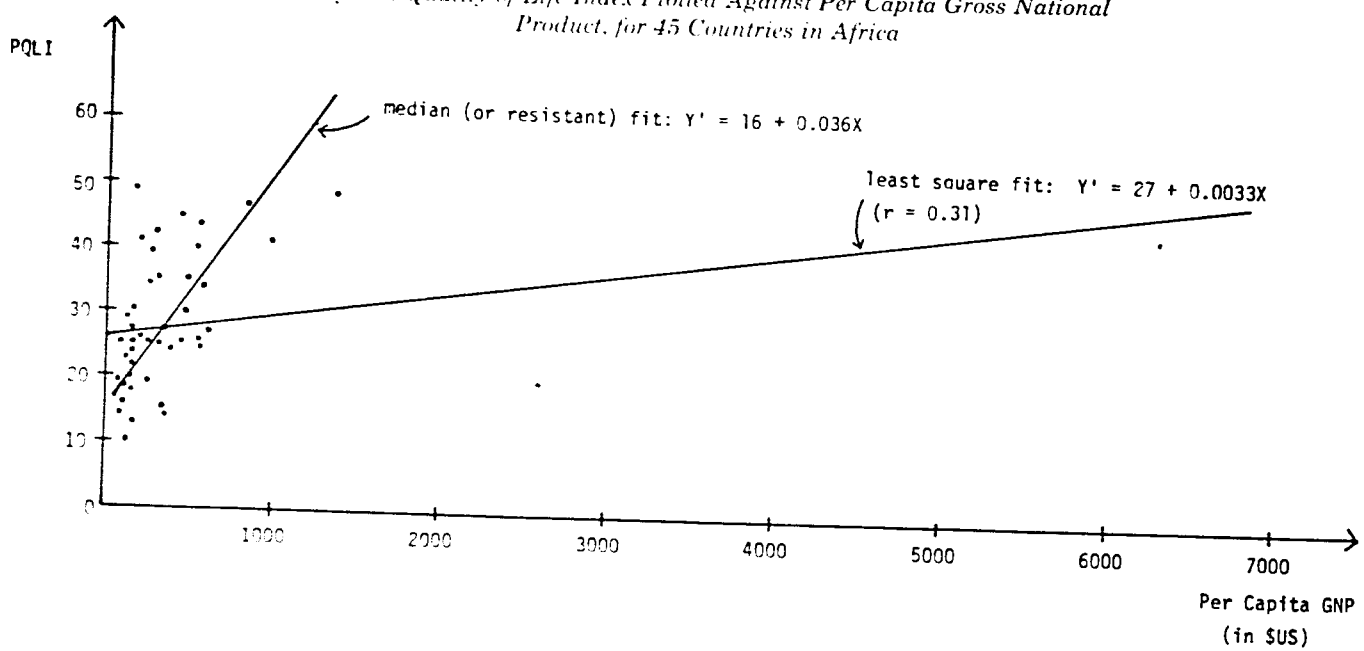
X	Y
1	2
2	7
3	8
4	5
5	8



The plot of the PQLI and GNP data is quite interesting, particularly when we fit a line to the data using different procedures. This plot is a good example of the effect of outlier values upon the least-squares fitting procedure. The data points for the two countries with large GNP's lie considerably removed from the rest of the scatter of points. The least-squares fitting procedure acts to minimize such large deviations and produces a least-squares fit of $Y = 27 + 0.0033X$. The median fitting procedure is more resistant to the effect of outliers, and the fitted line gives more importance to the preponderance of data points and partially ignores the effects of the outliers. The

Figure 5.14:

Least-Squares and Median Fittings to a Scatter of Points Showing the Physical Quality of Life Index Plotted Against Per Capita Gross National Product, for 45 Countries in Africa

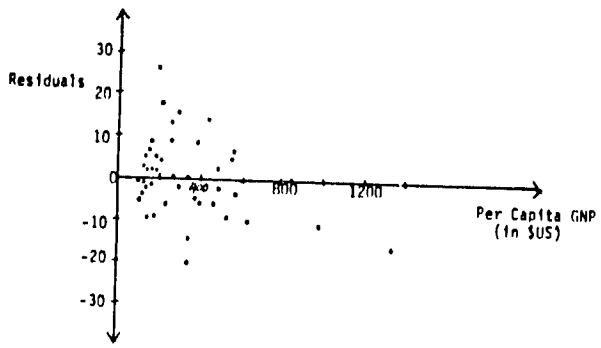


Source: Time, (March 13, 1978), pp. 26-27. PQLI is an index constructed by the Overseas Development Council and includes life expectancy, literacy and infant mortality. Per capita GNP figures are from 1976 World Bank data as shown in the Time article.

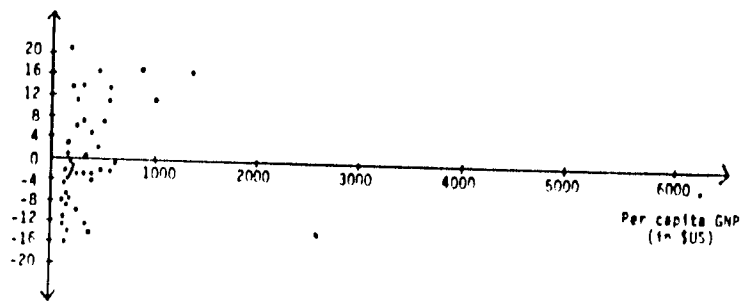
Figure 5.15:

A Plot of Residuals from the Resistant and Least-Squares Fits to the Data in Figure 5.14

a) Residuals from the resistant fit (excluding two outlier points)



b) Residuals from the least-squares fit



median, or resistant, line is $Y = 16 + 0.036X$. If one were to summarize these data using the two equations, there would be two considerably different descriptions, depending upon the fitting procedure used. The median fit results in a slope which was nearly ten times larger—steeper—than the slope of the least-squares line!

Because of the considerable influence of two outlier values, we would do well to consider each of these points separately. For example, we would want to recheck the numbers—to see that they had been copied correctly, or to go back to original sources and confirm their accuracy. We would also wish to consider whether there were factors operating in these countries which might make them very atypical of other African countries. If we found sufficient cause, it is possible that these two data points could be excluded from the remainder of the analysis. If we take this route, the resistant fit to the other data points appears pretty good. A plot of the residuals from this fit,

as shown in Figure 5.15(a), indicates little structure to the residuals, although at least one other possible outlier value is indicated (the country with GNP of 170 and a PQLI score of 48, indicating this country is doing much better than other countries in terms of the effectiveness of social services, given the level of GNP).

However, it may be that we would be unwilling to exclude the two countries with relatively high per capita GNP from our data set. It may also occur to us as we look at the plot of the data in Figure 5.14 that a tendency to plateau occurs. With an initial increase in per capita GNP, there is a fairly rapid rise in PQLI, but then a point is eventually reached where the PQLI index begins to flatten out. Credibility is given to this possibility in the plot of residuals. As seen in the residuals from the least-squares fit (Figure 5.15b), there definitely appears to be some curvature in the residuals around the level of $1000 per capita GNP.

What if we proceed with the assumption that there might be a plateau effect occurring, and that the two countries with large per capita GNP do not represent values to be excluded? In such a case, we have already seen that the linear fit is not adequate because there is structure remaining in the residuals. One can proceed with the analysis by transforming the data, either for the X or for the Y variables, or for both variables simultaneously. Data are transformed so that the linear form does provide a good fit to the data. By doing so, we increase our power of description, as will be shown with the examples to follow.

6.1 *Selection of the Appropriate Transformation*

In Chapter 2 we saw the effect that transforming data for a single variable could have on the shape of the distribution of values for that variable. We also developed a procedure for selecting an appropriate transformation from among all those available. With plots of data we have several options open for transformation. We could transform either or both of the variables. The selection of an appropriate transformation partially depends upon the approximate shape of the set of plotted data. Figure 5.16 shows a helpful schema developed by Tukey (1977) for selection of an appropriate transformation.

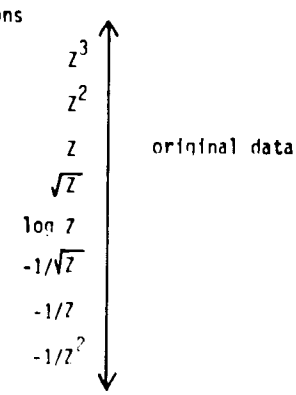
Also included in Figure 5.16 is the scale of transformations, as shown in Chapter 2. Recall that the original data in this scale are indicated by Z, and that transformations up the scale towards Z^2 emphasize differences among large numbers, while movement down the scale emphasizes differences among small numbers. Which direction to move along the scale depends on the general shape of the plot of the original data, and four shapes are shown in Figure 5.16.

The rule for movement on the scale of transformation is the following: move in the direction in which the curve bulges. Thus, for curve shape I in Figure 5.16, the bulge points toward higher Y values and higher X values. Thus we could attempt to straighten out the data by moving up the scale of transformations for either X or Y. The bulge for curve type IV, on the other hand, indicates that the correct direction to move would be up the scale for a Y

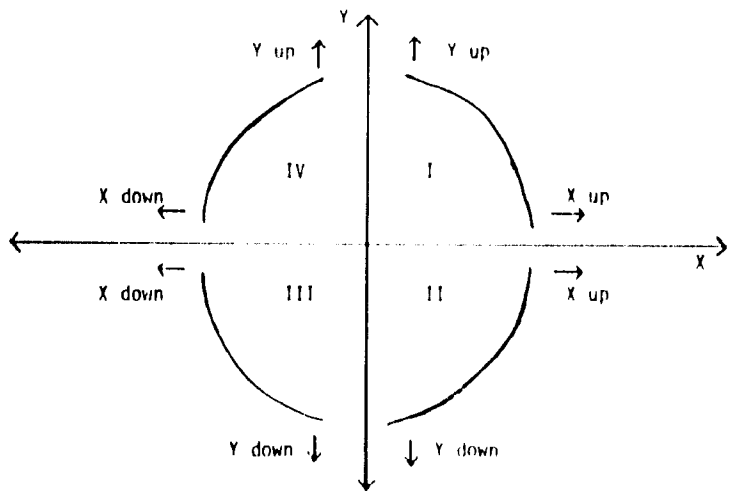
Figure 5.16:

A Schema for Selecting an Appropriate Data Transformation

a) Scale of transformations



b) Schema for selecting a transformation (arrows indicate directions to move on scale of transformations)



Source: Frederick Mosteller and John W. Tukey, Data Analysis and Regression Reading, Ma.: Addison-Wesley, 1977, p. 84.

transformation or down the scale for an X transformation (since the bulge of the curve points toward smaller X values).

Note that curve type IV is the general shape of the curve in the plot of PQLI versus GNP, meaning that if we are to straighten out the plot, we can do so by transforming the original Y values, perhaps as Y^2 or Y^3. We could also proceed by transforming the original X values, to $\sqrt{X}$, logX or some similar transformation down the scale of transformations.

Whether to transform the Y values or the X values is to some extent situationally dependent, but there are usually two factors that must be considered. First, one must consider the extent of the variation in the two variables. Variables with little variation in the values will result in less effective straightening under transformation than will variables where there is considerable variation in the values. Secondly, one must consider the meaning of the two variables, and which one is more meaningful if left in its original state. For example, in a plot of earnings versus time (in years), it probably makes more sense to transform earnings than it does time, because people would have less trouble thinking in terms of the log of earnings than they would of thinking in terms of the log of a year.

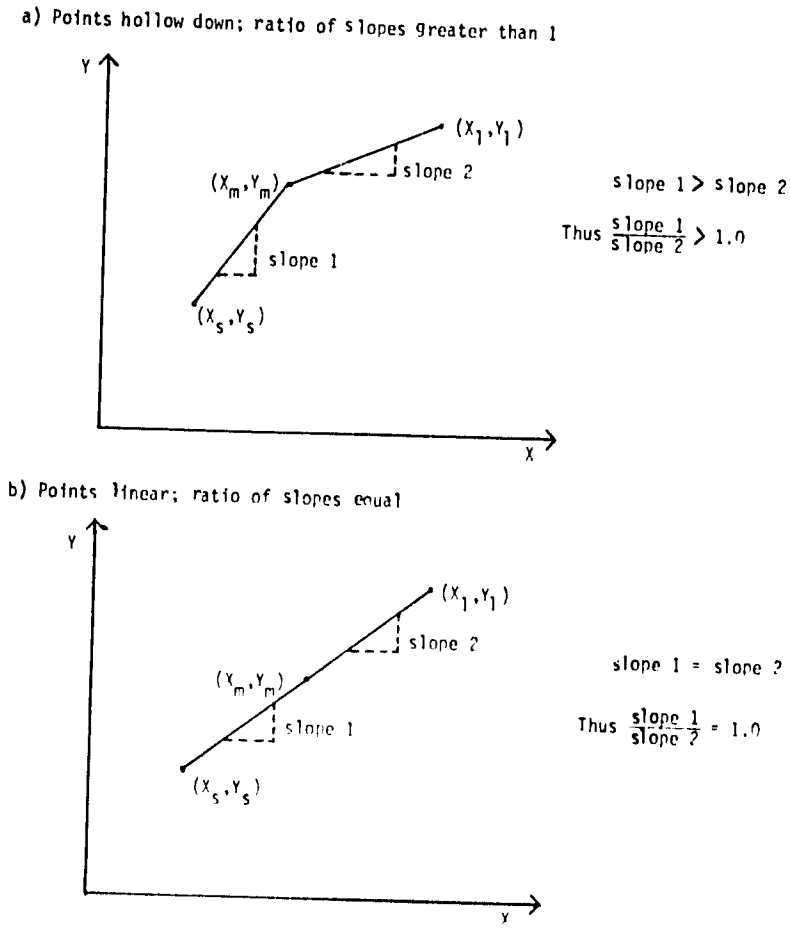
What direction should we move for transformation of the GNP and PQLI data? From the first criterion just discussed, it seems more appropriate to transform the X values (GNP).[2] The PQLI values vary only from a minimum of 11 to a maximum of 48, while the values of per capita GNP vary from a minimum of $100 to a maximum of $6310. In terms of ease of interpretation, it is also likely that one would prefer to leave the index scores for PQLI in their original form and to transform the GNP values, but this is open to individual preference. For the present illustration, assume that transforming GNP is the more useful route to take.

Having selected the variable to transform in our example, and the general direction to move on the scale of transformations, how do we select which of the possible transformations to try? One procedure would be to go ahead and try different transformations, selecting the one which best straightened out the plot. This, however, is time consuming for all but the smallest of problems. Even with the 45 data points as in the present example, this procedure would be a formidable undertaking.

Luckily there is a procedure that can do a pretty good job in deciding upon the appropriate transformation. To do this we make use of the three data points (X_1, Y_1), (X_2, Y_2), and (X_3, Y_3) that are used when making a resistant fit to the data. The idea is simple: using (X_1, Y_1) and (X_2, Y_2) one slope is computed. Using (X_2, Y_2) and (X_3, Y_3) a second slope is computed. If these two slopes are about equal in magnitude, the three points will lie on approximately a straight line (see Figure 5.17). If they are not equal, we transform the variable selected (here it is X), using a transformation from the scale of transformations, and again compute the slopes. This process requires only that three data points be transformed, rather than the complete set of data. When a transformation has been selected, then all the data points can be transformed. The end result is a considerable savings in time.

Figure 5.17:

An Illustration of the Slope Ratio Method for Selecting an Appropriate Transformation



This method of selecting the transformation is illustrated with the data for PQLI and GNP. Using the same procedure described earlier in the chapter, it was found that $(X_s, Y_s) = (130,20)$, $(X_m, Y_m) = (260,27)$ and $(X_l, Y_l) = (550,35)$. Figure 5.18 shows the transformed values for these three data points under a variety of different transformations. Also shown in this figure are the computations of the slopes, and the resulting ratio of the slopes.

Figure 5.18:

Selecting an Appropriate Transformation for the PQLI and GNP Data in Figure 5.14

a) Find (X_s, Y_s), (X_m, Y_m) and (X_l, Y_l), as in fitting a resistant line.

Using the original data set, it can be determined that:

$$(X_s, Y_s) = (130,20)$$
$$(X_m, Y_m) = (260,27)$$
$$(X_l, Y_l) = (550,35)$$

b) Select an appropriate X transformation (see text for discussion as to why X)

	X_s	X_m	X_l
X	130	260	550
$\sqrt{X}$	11.4	16.1	23.5
$\log X$	2.11	2.41	2.74
$-1/\sqrt{X}$	-0.088	-0.062	-0.043

c) Compute slopes

Transformation	Slope 1	Slope 2	Ratio of slopes
X	(27-20)/(260-130)=0.054	(35-27)/(550-260)=0.028	1.93
$\sqrt{X}$	7/4.7 = 1.49	8/7.4 = 1.08	1.38
$\log X$	7/0.3 = 23.3	8/0.33 = 24.2	0.96
$-1/\sqrt{X}$	7/0.026 = 269	8/0.019 = 421	0.64

d) Use ratio of slopes to select a transformation

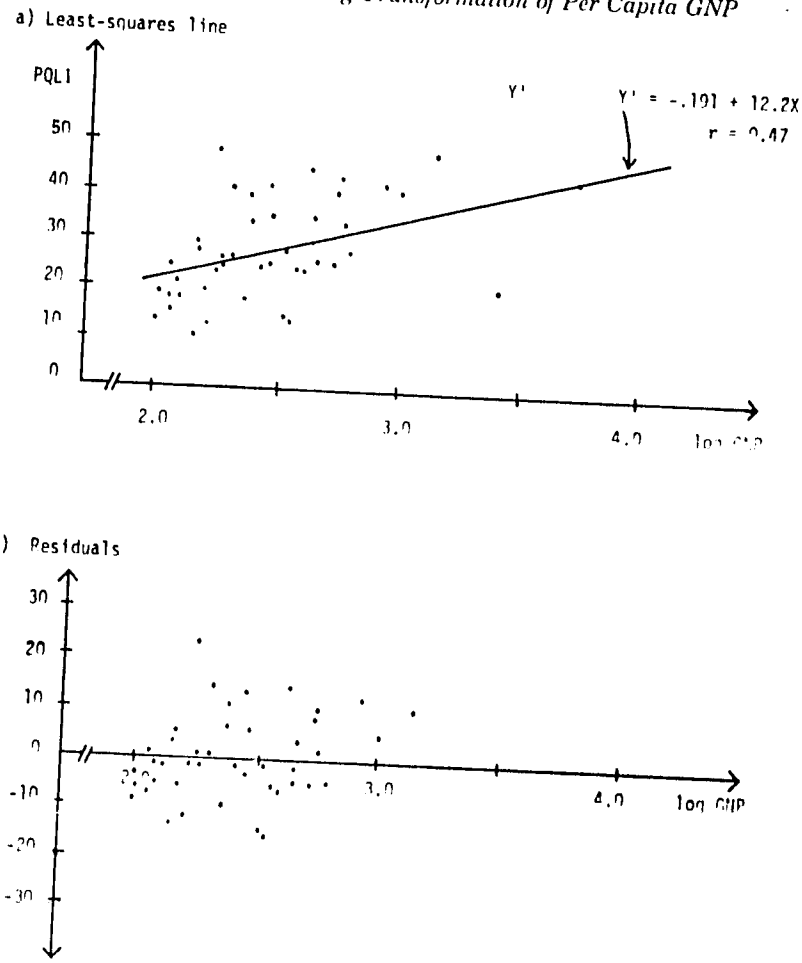
Since the logarithmic transformation produces a ratio of slopes closest to 1.0, this is the transformation to choose for the data above.

When the ratio of slopes is close to 1, it indicates that the transformation has succeeded in taking the curvature out of the three points used to compute the slopes. As shown in Figure 5.17, the log transformation appears to take out most of the curvature from the three data points used as representative points.

Using the log transformation on all the X data values produces the plot shown in Figure 5.19. When fitted with a least-squares line and a plot of residuals made (Figure 5.19b), we see that, even with a log transformation, there appears to be some structure left in the residuals. Low X values tend to have generally more negative residuals, while the higher X values tend to have positive residuals. The two extreme values in X (high per capita GNP) are still exerting considerable influence over the least-squares line. A resistant line fitted to the log transformation data would still have a larger slope than the least-squares fit.

Figure 5.19:

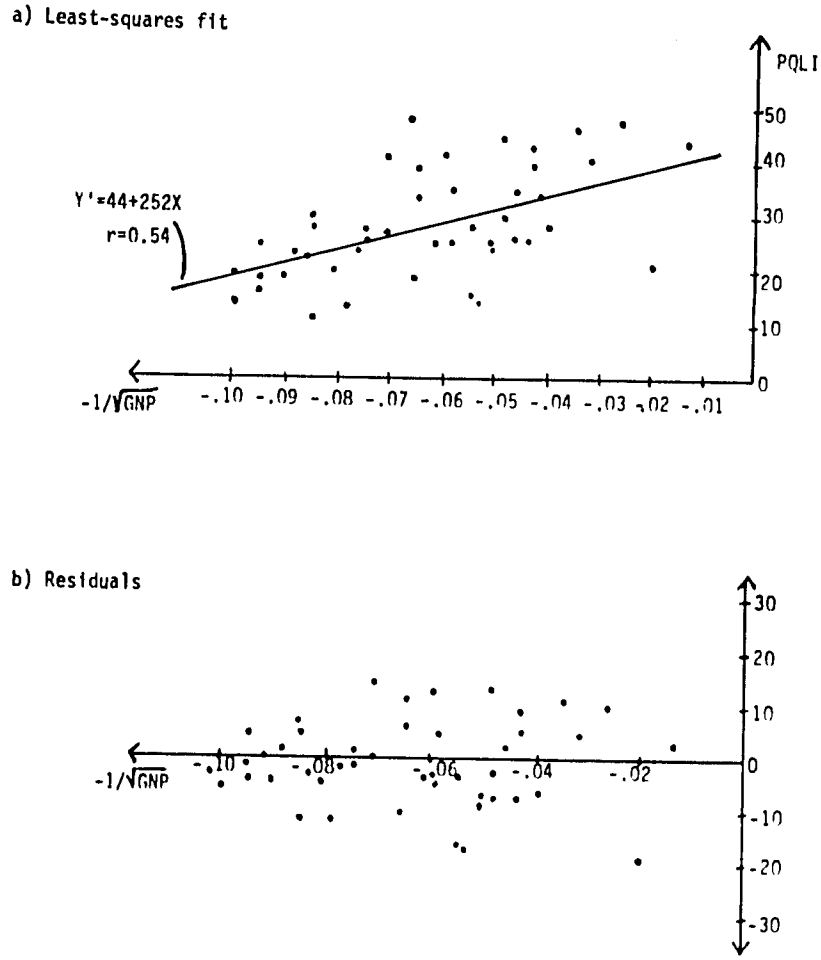
A Least-Squares Fit to the Data in Figure 5.14, Where the Independent Variable is now the Log Transformation of Per Capita GNP



If we had selected the next lower transformation on the scale of transformations, that of $1/\sqrt{X}$, the results are more encouraging (see Figure 5.20). The residuals show less pattern, indicating that this transformation is somewhat better than the log transformation. Our transformation selection procedure got us close to a good fit, but there can be no guarantee that it will result in the best transformation until all data points are plotted and residuals found.

Figure 5.20:

A Least-Squares Fit to the Data in Figure 5.14, Where the Independent Variable is now the Negative Reciprocal of Per Capita GNP



Transformations of a moderate sized data set, such as the example just discussed, can represent several hours' work with a hand calculator. Is it worth it? What in fact have we gotten out of such a transformation of the GNP and PQLI data? There are several things. First, we have a better description of the data than the original least-squares description in Figure 5.14. In that

fit, the amount of variance explained was $(0.31)^2 = 0.1$, or about 10%. Fitting the data after transforming X values by $-1/\sqrt{X}$ produced an r^2 value of 0.29, or nearly triple what the original linear fit provided. From a purely descriptive point of view, we have succeeded in summarizing the data better with the transformed data than with the original data.

Better description results in a second advantage. By describing the data better, we can also feel somewhat more secure in the identification of outliers. Such identification can be one of the most useful of results from such fittings. In Figure 5.20, it is quite clear that there are two outlier values, the first for the value (170,48) and the second for the value (2590,21). Assuming that neither value represents some error in measurement, computation, or blunder, then we have some useful information in these residuals. For example, it would be very interesting to do some additional research on the first country, to see how it is that they can achieve such high levels of efficiency in social services despite fairly low levels of per capita GNP. On the other hand, investigation of the second country might yield factors that inhibited the effectiveness of social services, despite a fairly high per capita GNP. Comparing and contrasting the factors in both countries might be even more productive.

A third benefit from such transformations concerns the form of the fit itself. Educational planners are never going to be concerned with trying to discover "natural laws," so the particular function that is used to fit the data may never be of concern or need interpretation. The general form may be of interest, however. In the present example, the best fit is provided by a transformation which, if re-transformed to the original data, would show a plateau in the fit. In Figure 5.21 PQLI is plotted against GNP again, with the best transformed fit from Figure 5.20 ($Y' = 44 + 252X$, where $X = -1/\sqrt{GNP}$) indicated on the plot. Is it in fact true that the effectiveness in social services plateaus after a certain level of GNP?

There is some evidence in the data set to support a plateau effect, as we have already discussed. But the present data are not sufficient to establish such an effect, and one would want to collect additional information. For example, the range beyond a per capita GNP of $1000 is pretty bare, with only three countries represented. It may be possible to use data from other countries to fill in some of the gaps in the data set used here.

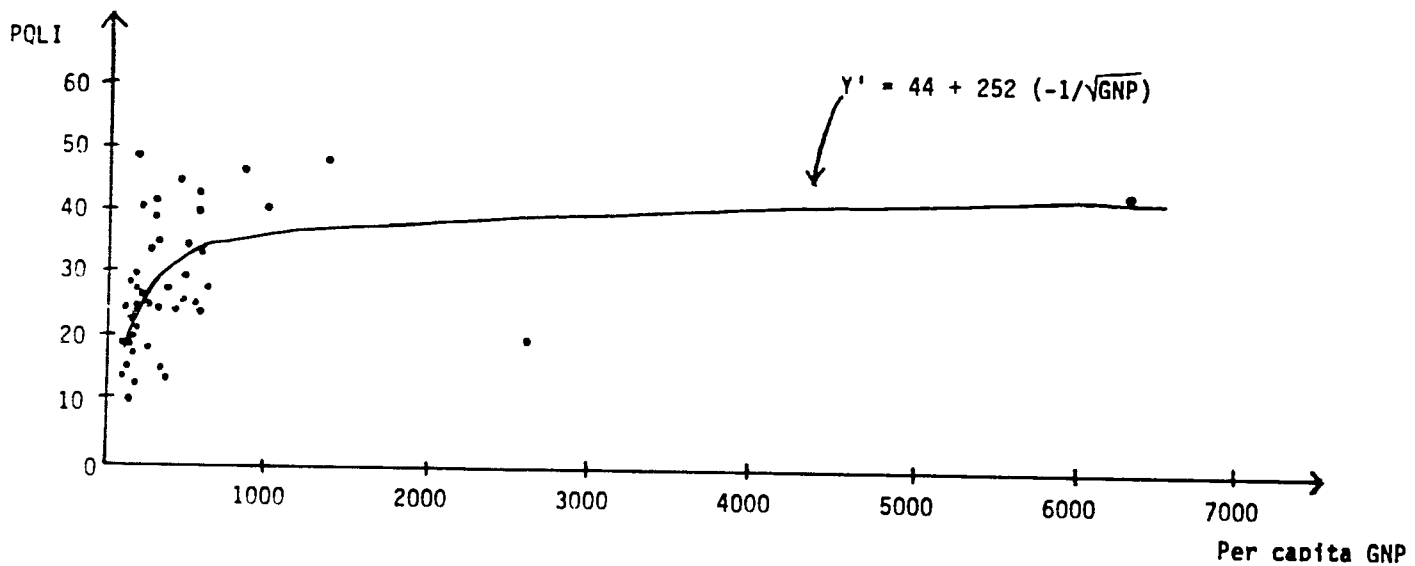
In sum, the effort required to make data transformations and then fit the transformed data is often repaid by the increased depth and richness provided by the analysis, by a better description of the data, identification of outlier values that may be of interest, and the meaning given to the general form of the function fitted to the data.

7.0 *Multiple Regression*

Up to this point we have related a single dependent variable to only one other variable, by fitting a line or some other functional form to the plot of the two variables. We saw that this could be done for both description and prediction purposes. However, in many cases it is likely that the dependent

Figure 5.21:

The Fit of Figure 5.20 Applied to a Plot of the Data in Figure 5.14



variable is related to more than one other variable, and that both variables are of interest to us. In such circumstances, where there are two or more variables that we believe to be related to the dependent variable, the basic regression procedures can be extended.

One extension of simple regression is called multiple regression, where the word *multiple* refers to two or more independent variables.[3] With multiple regression, the technique can be used for prediction and description, as with simple regression. It can also be used, however, to assess the relationship between one of the independent variables and the dependent variable, while excluding the effect of the other independent variable or variables. Regression as *exclusion* is a useful tool for an educational planner, although the interpretation of the results depends on factors that may be easily confused. We discuss these factors in a later section.

In this section we will again return to our simple example, first introduced in this chapter, dealing with admissions and enrollment data. We will also make use of the column of data in Table 5.1 called "applications," which is simply the number of applications received from prospective students in given years. Thus, for each year we have a data point that is a triplet (Y, X_1, X_2), where Y is enrollment, X_1 is admissions, and X_2 is applications. The data for the first year, for example, are just the triplet (416;604;1178).

These triplets can no longer be plotted on a two-dimensional piece of paper because there are three variables, but it is possible to imagine them plotted in three-dimensional space. Multiple regression with two independent variables amounts to fitting a plane to the scatter of points in three dimensions. This is entirely analogous to simple regression, where we tried to fit a line to a scatter of points in two dimensions. The basic form, with two independent variables, is:

$$Y' = a + b_1X_1 + b_2X_2, \tag{5.10}$$

where a is the Y-intercept and b_1 and b_2 are both slopes.

7.1 *The Fitting Procedure*

There are a variety of fitting procedures for performing multiple regressions. Which one is used depends fundamentally on how large the data set is, and on what kind of equipment is available for analysis. If the data set is large, a computer is required. What is large depends to some extent on the individual analyst, but anything beyond 50 units becomes quite burdensome for all but the hardiest. In such cases a computer will routinely calculate the coefficients that define the fit to the actual data. There are now available packaged computer routines which permit data transformations, plotting of residuals, and other techniques used by the careful analyst.

Assuming that the data set is small to moderate, however, it is possible and even advantageous in some respects to do multiple regressions by hand. This is aided enormously by the use of a hand calculator, and there are now calculators with built-in statistical programs which will compute correlations and the slope and intercept for a simple least-squares regression. In

what follows, various least-squares fitting procedures are indicated for a multiple regression with two independent variables.

7.1.1 A least-squares fit using a hand calculator having no statistical routines. Without a hand calculator that performs routine statistical computations, one needs to set up a table for the orderly calculations that must be done to determine the intercept and slopes that define a multiple regression equation. The basic table for these calculations is shown in Figure 5.22, for the admission, enrollment, and application data. The steps are the following:

Figure 5.22:

Calculations for Performing a Multiple Regression, for the Data on Enrollments, Admissions, and Applications as Shown in Table 5.1

a) Compute necessary sums of squares and cross-products

Enrollments Y	Admissions X_1	Applications X_2	Y^2	X_1^2	X_2^2	X_1Y	X_2Y	X_1X_2
416	604	1178	173,056	364,816	1,387,684	251,264	490,048	711,512
504	648	1097	254,016	419,904	1,203,409	326,592	552,888	710,856
426	558	1006	181,476	311,364	1,012,036	237,708	428,556	561,348
447	630	1078	199,809	396,900	1,162,084	281,610	481,866	679,140
445	595	1166	198,025	354,025	1,359,556	264,775	518,870	693,770
411	581	1095	168,921	337,561	1,199,025	238,791	450,045	636,195
511	694	1059	261,121	481,636	1,121,481	354,634	541,149	734,946
Means 451.4	615.7	1097						
Std. Dev. 40.6	45.6	59.7						
Sums 3160	4310	7679	1,436,424	2,666,206	8,445,275	1,955,374	3,463,422	4,727,767

b) Compute necessary terms for determining slopes

$$\Sigma x_1^2 = \Sigma X_1^2 - (\Sigma X_1)^2/N = 2{,}666{,}206 - \frac{(4310)^2}{7} = 12{,}477$$

$$\Sigma x_2^2 = \Sigma X_2^2 - (\Sigma X_2)^2/N = 8{,}445{,}275 - \frac{(7679)^2}{7} = 21{,}412$$

$$\Sigma x_1y = \Sigma X_1Y - (\Sigma X_1)(\Sigma Y)/N = 1{,}955{,}374 - \frac{(4310)(3160)}{7} = 9717$$

$$\Sigma x_2y = \Sigma X_2Y - (\Sigma X_2)(\Sigma Y)/N = 3{,}463{,}422 - \frac{(7679)(3160)}{7} = -3098$$

$$\Sigma x_1x_2 = \Sigma X_1X_2 - (\Sigma X_1)(\Sigma X_2)/N = 4{,}727{,}767 - \frac{(4310)(7679)}{7} = -303$$

c) Compute regression parameters

$$b_1 = \frac{(\Sigma x_2^2)(\Sigma x_1y) - (\Sigma x_1x_2)(\Sigma x_2y)}{(\Sigma x_1^2)(\Sigma x_2^2) - \Sigma x_1x_2} = \frac{(21{,}412)(9717) - (-303)(-3098)}{(12{,}477)(21{,}412) - (-303)^2} = 0.78$$

$$b_2 = \frac{(\Sigma x_1^2)(\Sigma x_2y) - (\Sigma x_1x_2)(\Sigma x_1y)}{(\Sigma x_1^2)(\Sigma x_2^2) - \Sigma x_1x_2} = \frac{(12{,}477)(-3098) - (-303)(9717)}{(12{,}477)(21{,}412) - (-303)^2} = -0.13$$

$$a = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2 = 451.4 - 0.78(615.7) - (-0.13)(1097) = 113.8$$

Thus $Y' = 113.8 + 0.78X_1 - 0.13X_2$

Step 1: Sum the values for each variable. Call these ΣY, ΣX_1, and ΣX_2.

Step 2: Find the means of each of the variables. Call these $\overline{Y}$, $\overline{X}_1$, and $\overline{X}_2$, for the dependent variable, and the first and second independent variables respectively.

Step 3: Compute Y^2, X_1^2, and X_2^2. For each variable, sum the squared values and call these ΣY^2, ΣX_1^2, and ΣX_2^2.

Step 4: Find the cross products YX_1, YX_2 and X_1X_2, for data in each triplet. Sum the cross-product terms and call them ΣYX_1, ΣYX_2, and ΣX_1X_2.

Step 5: Use the summation terms found in steps 2 through 4 above to compute the following:

$$\Sigma x_1^2 = \Sigma X_1^2 - (1/N)(\Sigma X_1)^2$$

$$\Sigma x_2^2 = \Sigma X_2^2 - (1/N)(\Sigma X_2)^2$$

$$\Sigma x_1 y = \Sigma X_1 Y - (1/N)(\Sigma X_1)(\Sigma Y)$$

$$\Sigma x_2 y = \Sigma X_2 Y - (1/N)(\Sigma X_2)(\Sigma Y)$$

$$\Sigma x_1 x_2 = \Sigma X_1 X_2 - (1/N)(\Sigma X_1)(\Sigma X_2)$$

Step 6: Use the values in step 5 to calculate the slope and intercept with the following equations:

$$\text{slope } b_1 = \frac{(\Sigma x_2^2)(\Sigma x_1 y) - (\Sigma x_1 x_2)(\Sigma x_2 y)}{(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2} \tag{5.11}$$

$$\text{slope } b_2 = \frac{(\Sigma x_1^2)(\Sigma x_2 y) - (\Sigma x_1 x_2)(\Sigma x_1 y)}{\Sigma(x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2} \tag{5.12}$$

$$\text{intercept } a = \overline{Y} - b_1\overline{X}_1 - b_2\overline{X}_2 \tag{5.13}$$

Shown in Figure 5.22 are the computations necessary for finding the slopes and intercepts for the enrollment data. As shown there, the fitted equation turns out to be:

$$Y' = 113.8 + 0.78X_1 - 0.13X_2$$

How to interpret these values will be saved for a later section.

7.1.2 A least-squares fit, using a hand calculator having statistical routines. It is far less tiresome to do a least-squares multiple regression if the analyst has at hand a calculator which will compute correlation coefficients. The data are simply entered into the machine in the manner appropriate for that machine, and correlation coefficients obtained. In our example we would

have the following three correlations: r_{XY_1}, r_{YX_2}, and $r_{X_1X_2}$. To compute the regression coefficients, you also need to have the standard deviations for each of the three variables. These are obtained from the pre-programmed routines in the calculator, or by use of the equations given in Chapter 2. With this information, the regressions coefficients can be computed as:

$$\text{slope } b_1 = \frac{r_{YX_1} - r_{YX_2}r_{X_1X_2}}{1 - r^2_{X_1X_2}} \frac{S_Y}{S_{X_1}} = b_1^* \frac{S_Y}{S_{X_1}}, \qquad (5.14)$$

$$\text{slope } b_2 = \frac{r_{YX_2} - r_{YX_1}r_{X_1X_2}}{1 - r^2_{X_1X_2}} \frac{S_Y}{S_{X_2}} = b_2^* \frac{S_Y}{S_{X_1}}, \qquad (5.15)$$

where r_{ij} is the correlation between variables i and j; S_i is the standard deviation of variable i; and b_i^* is known as a standardized regression coefficient.

$$\text{intercept} = a = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2$$

Figure 5.23 shows the computations of the regression coefficients using this approach. It is less tiresome than the previous procedure, but is still limited because of need to manually enter the data into the calculator several times. The values of b_1^* and b_2^*, the so-called standardized regression coefficients, can be used to compare the relative importance of the two variables in terms of affecting the dependent variable. As shown in Figure 5.23, b_1^* is .87, while b_2^* is −.20, indicating that X_1, admissions, is substantially more important in predicting enrollments, than is X_2, applications. Such regression coefficients run from −1 to +1, with values near zero indicating little predictive importance.

7.2 *Examining Residuals*

Just as with simple regressions, it is necessary to examine the residuals obtained from the least-squares multiple regression fit. The fact that there are more variables means that more plots must be examined. With two independent variables, we would minimally want to examine the following plots: (1) residuals versus X_1 and (2) residuals versus X_2. In each case we are looking for the same things as with the simple regression. Plots of residuals versus the independent variables should be examined for a pattern in the residual, particularly curvature that would indicate a data transformation is necessary.

Shown in Figure 5.24 are plots of the residuals from the least-squares fit in Figure 5.22. With only a few data points, it is hard to pick up trends, but none of the plots in Figure 5.24 indicates the need for transformations in the data. We would probably be content to summarize the data with the linear fit

Figure 5.23:

Computation of Multiple Regression Coefficients using Correlation Coefficients, for the Data in Table 5.1 and Figure 5.22

Example: $$r_{x_1y} = \frac{7(1{,}955{,}374) - (4310)(3160)}{\{[7(2{,}666{,}206)-(4310)^2][7(1{,}436{,}424)-(3160)^2]\}^{1/2}} = 0.874$$

In similar fashion, $r_{x_2y} = -0.213$

$r_{x_1x_2} = -0.0185$

b) Use correlation coefficients to compute regression coefficients

$$b_{ij\cdot k} = \left[\frac{r_{ik} - r_{ij}r_{jk}}{1-r_{jk}^2}\right]\left(\frac{s_i}{s_j}\right) ,$$

where $b_{ij\cdot k}$ = the regression coefficient for the regression of variable i on variable j, with both i and j linearly adjusted for variable k (see text)

r_{ij} = the correlation coefficient between variables i and j

s_i = the standard deviation of variable i

Thus $$b_{yx_2\cdot x_1} = \left[\frac{(-0.213)-(0.874)(-0.0185)}{1-(-0.0185)^2}\right]\left(\frac{40.6}{59.7}\right) = -0.13$$

Similarly $b_{yx_1\cdot x_2} = 0.77$

a) Compute correlations, using data in Figure 5.22 (Y = Enrollment, X_1 = Admissions, X_2 = Applications)

$$r = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]\}^{1/2}}$$

Figure 5.24:

Plots of Residuals from the Multiple Regression in Figure 5.22

a) Residuals

Y	X_1	X_2	Fit[a]	Residuals
416	604	1178	432	-16
504	648	1097	477	27
426	558	1006	418	8
447	630	1078	465	-18
445	595	1166	426	19
411	581	1095	425	-14
511	694	1059	517	- 6

[a] $Y' = 113.8 + 0.78X_1 - 0.13X_2$

b) Residuals plotted versus X_1

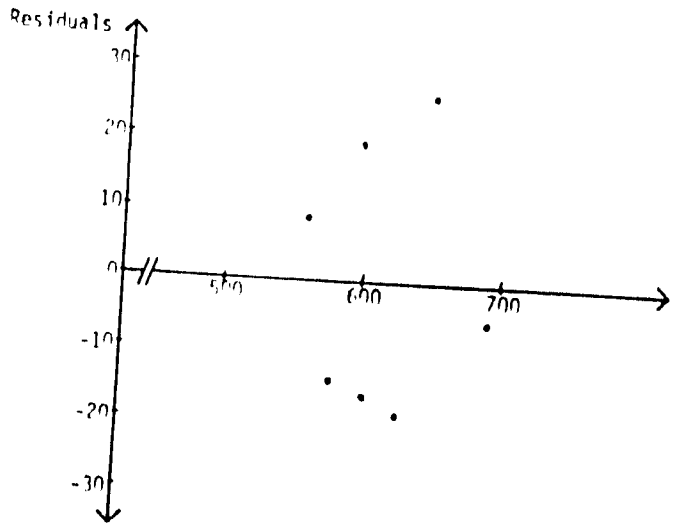
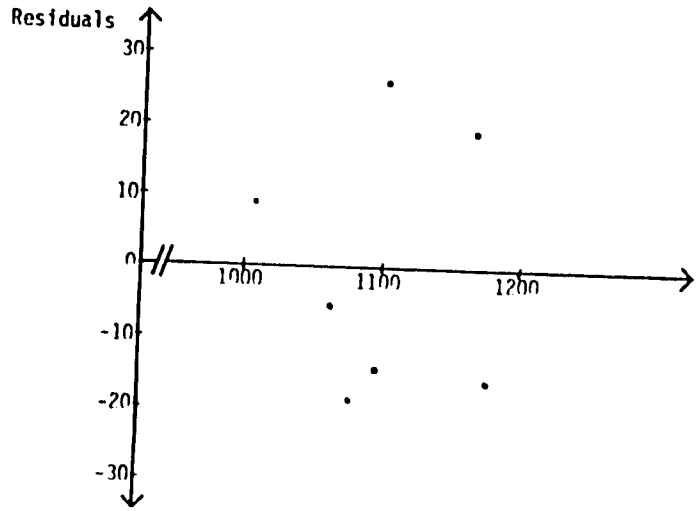


Figure 5.24:

continued

c) Residuals plotted versus X_2



Assuming that there is no structure found in the plots of residuals versus X_i, we feel reasonably content with the functional form used to fit the data, we can compute a statistic which summarizes the closeness of the fit. Recall that with simple regression analysis, the square of the correlation coefficient could be interpreted as the amount of variance in the dependent variable that was explained by the regression line. With multiple regression it is possible to compute a number called the multiple correlation, symbolized as R. This is the correlation between the dependent variable and the fitted values obtained by predicting Y with the regression equation. In Figure 5.24, for example, we could compute a correlation between the actual Y values and the fitted Y values. This correlation is found to be 0.90. Values for a multiple correlation coefficient run from a low of zero, indicating no correlation between observed and fitted values, to a maximum of 1.00, indicating the predicted and observed values are identical.

It can be shown, furthermore, that the square of the multiple correlation coefficient can be given the same interpretation as the square of the simple correlation coefficient. That is, R^2 is interpreted as the amount of variation in the dependent variable that is explained by the multiple regression equation. In the example with enrollments, $R^2 = (.90)^2 = 0.81$, indicating that about 81% of the variation in enrollments is explained by the two independent variables, admissions and applications.

It is also possible to compute R^2 directly from the simple correlation coefficients between variables. The formula is:

$$R^2 = \frac{r^2_{YX_1} + r^2_{YX_2} - 2r_{YX_1}r_{YX_2}r_{X_1X_2}}{1 - r^2_{X_1X_2}} \tag{5.16}$$

For the enrollment data, using correlations found in Figure 5.23, we can compute R^2 directly as:

$$R^2 = \frac{(.874)^2 + (-.213)^2 - 2(.874)(-.213)(-.019)}{1 - (-.019)^2}$$
$$= .80$$

7.3 *The Interpretation of Multiple Regression Coefficients*

So far we have discussed the easy part of multiple regression, the mechanics of fitting a particular scatter of points and then examining residuals, transforming the data if necessary, and computing summary statistics. Now comes the more difficult question of how to interpret the coefficients. To date we have called the b_p coefficients "slopes," an analogy to the case of the simple regression. The use of this term is not unwarranted. The coefficient for a given "independent" variable is a slope, but it cannot be interpreted independently of what other variables are in the regression equation. For example, we saw that the fit to the enrollment data was:

$$Y' = 113.8 + 0.78X_1 - 0.13X_2$$

We *cannot* interpret the value 0.78 as the general increase in the dependent variable with a unit change in X_1. The interpretation is more subtle than this and must include reference to the variable X_2. If there are more than two independent variables in a multiple regression equation, interpretation of the slope for any one of these variables must also be made with reference to the complete set of the other independent variables.

The more precise interpretation of slope coefficients like 0.78 goes something like this: it is the general change in the dependent variable for a one unit change in X_1, *after adjustment has been made for linear change in* X_2. This is an accurate interpretation, but unfortunately it does not mean anything at present. To put meaning into the statement we must first demonstrate another way to compute the slope coefficients. This turns out to be very useful for getting across the meaning of the coefficients, but it is not a very easy method for computing the coefficients, and hence was not included under fitting procedures in the previous section.

As we will illustrate with the enrollment data, if we take the following steps, we can compute the regression coefficient b_1 (similar steps could be done to compute b_2, but only one will be discussed here for illustration):

1. Perform a simple least-squares regression of Y on X_2. Use the regression line to compute residuals. These residuals can be thought of as "Y adjusted for the linear effects of X_2 on Y." In other words, these residuals are just the original Y values with

the linear effects of X_2 taken out, or excluded. Call this set of residuals Y_r.

2. Perform a simple least-squares regression of X_1 on X_2. Use the regression line to compute residuals. These residuals can be thought of as "X_1 adjusted for the linear effects of X_2 on X_1." In other words, these residuals are just the original X_1 values with the linear effects of X_2 taken out, or excluded. Call this set of residuals X_{1r}.
3. Regress the residuals Y_r on the residuals X_{1r}. The slope of this simple regression is just b_1, the value obtained from the multiple regression of Y on X_1 and X_2.

Let us interpret this set of steps as simply as possible. Essentially what it says is this: a slope coefficient in a multiple regression equation is just the slope obtained from a simple regression of Y on X_1, *when both* Y and X_1 have been adjusted for the effects of X_2. Figure 5.25 shows this more clearly, for the example with the enrollment data. The regression of the two sets of residuals (i.e., Y and X_1 adjusted for X_2) yields the following equation:

$$Y'_r = 0.013 + 0.78X_{1r}$$

The slope here is the same as the b_1 slope computed for the multiple regression of Y on both X_1 and X_2.

Computing the multiple regression coefficient in this circuitous manner aids in the interpretation of the equations. We can see now why it is that the slope coefficient b_1 is the change in the dependent variable for a unit change in X_1, *but* only after *both* Y and X_1 have been adjusted for the linear effects of the other independent variables in the regression. With this we can proceed to interpret the full multiple regression equation $Y' = 113.8 + 0.78X_1 - 0.13X_2$. We now see that once we adjust for simultaneous changes in applications, an increase in the number of admissions by 100 will generally result in an increase in the number of enrollments by 78.

One other note of caution needs to be sounded regarding the interpretation of regression coefficients. In particular, what happens when the independent variables are correlated? This condition, called *collinearity*, can cause the size of the regression coefficients to become uncertain indicators of the importance of that variable (if the correlation is strong enough). In our example we were fortunate that the correlation between the two independent variables was small (0.02). But what if the correlation had been much stronger? In this case, we could not substantially change the values of one of the independent variables without affecting the values of the other independent variable. Consequently, the size of the slope coefficients would not tell us very much about the likely impact caused by some intervention to change the value of one of the independent variables. Recall that a slope can only be interpreted after adjustment for simultaneous change occurring as a result of the other independent variables. With a moderate or high correlation between the independent variables, our intervention to change one of them will affect the other. The outcome, in terms of the impact on the dependent variable, is not likely to be indicated by the slopes of the regression coefficient.

Figure 5.25:

An Illustration of the Interpretation of Multiple Regression Coefficients, Using Data from Figure 5.22

a) Regression of Y on X_2 is: $Y' = 610 - 0.145X_2$

b) Regression of X_1 on X_2 is: $X_1' = 631 - 0.014X_2$

c) Residuals

Y	X_1	X_2	$Y_r = Y - Y'$	$X_{1r} = X_1 - X_1'$
416	604	1178	-23.7	-10.6
504	648	1097	52.6	32.3
426	558	1006	-38.6	-59.0
447	630	1078	-7.2	14.0
445	595	1166	3.6	-19.7
411	581	1095	-40.7	-34.7
511	694	1059	54.1	77.8

d) Regression of Y_r on X_{1r}: $Y_r' = 0.013 + 0.78X_{1r}$

If, however, the correlation between the independent variables is weak, we might be able to get away with thinking about the size of the slope coefficients as the effect upon the dependent variable of a unit change in one independent variable, while the other independent variables are held *constant*. In our example, we have a small correlation between admissions and the number of applications. Within the ranges of the values in the data set then, we might consider the impact of a decision to increase admissions by 100. We could interpret the regression coefficient on admissions (0.78) as indicating that, if

applications could be held constant, we would expect about 78 new enrollees. Since the number of applications is partially beyond the control of an educational institution, the fact that we can interpret b_1 in the above manner may not make any sense in terms of what is possible from a policy standpoint. And in most cases we are not lucky enough to have such low correlations between the independent variables. In these cases we should not endeavor to interpret the coefficients as the likely effect if we could hold the other independent variables constant.

8.0 *Summary*

This chapter has covered a variety of techniques and procedures appropriate for the analysis of interval level data. When we wish to relate one dependent variable to one independent variable, simple regression analysis may be indicated. This involves, as a first approximation, fitting a straight line to the scatter of points. Several different fitting procedures were described, including a resistant fit using medians and a least-squares fit, which is very commonly used. Examination of residuals can tell us if there is structure to the residuals, and hence if the fit is a good one or a poor one. If there is a pattern such as curvature in the residuals, data transformations may be indicated. Guidelines were presented to help determine what transformation might be used. It was also shown that summarization and prediction are the two purposes of simple regression analysis that will be most used by educational planners. The presence of outlier values, if the regression fit is good, can also be of considerable interest.

Multiple regression is an extension of simple regression to include two or more independent variables. Several least-squares fitting procedures were described for the case with two independent variables. Residuals from a multiple regression continue to play an important role in determining the quality of the fit to the data, as do data transformations for correcting some of the more obvious problems in the data. Interpretation of the regression coefficients in multiple regression is more complicated than simple regression, because the interpretation depends upon what other variables are present in the equation. The limitations of this for policy and program intervention were discussed.

9.0 *Terms for Review*

summarize and expose
linear functional form
slope
intercept
residual from regression line
resistant fit
least-squares fit
outlier values
correlation coefficient

square of correlation coefficient
scale of data transformations
multiple regression
multiple correlation, R
R squared
collinearity
standardized regression coefficient
regression as description
regression as prediction
regression as exclusion

10.0 *Practice Exercises*

Table 5.2 shows data from the North Central data set on enrollment fluctuations. The data points were randomly selected from the 436 school districts represented in the data set. Shown in the table is information on both the average salary of the professional staff in the district and the size of the district (in numbers of students in schools).

When investigating the impact of fluctuating enrollments on the cost of education, it was discovered that the size of the school district was also closely related to cost. Suppose you as an analyst decide to regress average salary on district size, perhaps with the idea that this will be used in a later multiple regression that would include information on both size and enrollment trends. Perform the following steps and comment on the results of your observations at each of the stages:

Step 1: Plot average salary and school district size.

Step 2: Fit a resistant line to the scatter of data points. What are the slope and intercept values? Interpret the slope.

Step 3: Fit a least-squares regression line to the scatter of points. What are the slope and intercept values? Interpret the slope. Does this least-squares line differ much from the resistant fit? If so, why might this be the case?

Step 4: Plot residuals from the least squares fit against size. Is there structure to the residuals? If so, decide on an appropriate data transformation that might eliminate the structure present in the residuals.

Step 5: Compute the correlation coefficient between average salary and school district size. Interpret the square of the correlation coefficient.

Table 5.2:

Average Teacher Salaries and School District Enrollments, for a Randomly Drawn Sample of 25 Districts in the North Central Data Set

Average Salary	District Enrollment
284	276
268	204
299	192
263	491
291	639
285	752
278	379
304	614
293	591
270	538
290	610
251	478
296	448
266	322
248	352
304	644
298	342
277	674
344	1667
302	1019
283	1383
280	4374
300	3943
301	6073
424	8457

Footnotes

[1]Another example using information on PQLI and per capita GNP may be found in Morris D. Morris, "The Physical Quality of Life Index (PQLI)," *Development Digest*, vol. 18, no. 1 (Jan., 1980), pp. 95-109. This work came to the attention of the present author too late in the publication process to permit more extended discussion and comparison with the analysis provided in this chapter.

[2]With this example, there is a third criterion applicable to selecting which variable to transform. The scale of transformations is appropriate for use with non-bounded positive numbers, such as counts or amounts of things. With bounded numbers, such as percentages, or in this case with the PQLI index scores that are constrained between a minimum of zero and a maximum of 100, transformations other than those dealt with above are sometimes useful. These transformations are somewhat more complicated and not dealt with here.

[3]As Mosteller and Tukey are fond of saying, these variables are often *not* independent in the sense of being individually manipulable by someone. It is usually the case that an intervention will disturb both variables, unless in a strict experimental situation. Hence some analysts prefer to call these variables "carriers." "Independent" has been used here because it is the traditional term and is found in other sources the analyst may consult.

Bibliography

While incorporating some of the more standard material on least-squares regression, this chapter is heavily indebted to the exploratory techniques and procedures found in the following books:

Hoaglin, David C. "A First Course in Data Analysis," mimeo, 1976.
McNeil, Donald R. *Interactive Data Analysis*. New York: Wiley, 1977.
Tukey, John W. *Exploratory Data Analysis*. Reading, Mass.: Addison-Wesley, 1977.

For an excellent discussion concerning the various uses and abuses of regression analysis, see:

Mosteller, Frederick and John W. Tukey. *Data Analysis and Regression*. Reading, Mass.: Addison-Wesley, 1977.

CHAPTER 6

How Educational Statistics Can Be Utilized for Policy Analysis: A Case Study

Noel McGinn and Ernesto Schiefelbein

1.0 *Objectives of this Unit*

The purpose of this unit is to suggest ways in which planners can use *existing data bases* to provide information of value to decision makers and managers. The plan is to discuss how to:

— *Identify*, using the present statistical system, those *variables* of interest to decision makers and managers;

— Design and carry out a procedure for *sampling* existing data files to produce a usable data set;

— Design and carry out a procedure for rapid hand *tabulation* of data from the sample; and

— Conduct *analyses* of these data and present simple reports of findings.

The unit is designed, then, to suggest *ways in which existing data can be sampled and analyzed to generate useful information*. It does not supplant units on educational research procedures, nor does it pretend to cover general questions of information needs. The unit is addressed to planners with these kinds of questions:

— How can I organize my statistical office to respond to the information demands made by other parts of the Ministry?

— With no increment in budget, how can I proceed to carry out some basic analyses?

— How can I determine whether existing data are useful in assessing aspects of the functioning of the system?

— Can ordinary educational statistics be used to assess system policies? How?

The unit assumes that the Office of Planning knows what it would like to do, and what it needs to do, but that it has been unable to do so because of constraints of staff (both number and level of training) and time.

The procedures described also require the following *critical assumptions*:

— Most policy decisions allow a wide margin of error, and therefore few decisions require highly precise information.

— Decision makers and managers need information *now*, and prefer

quickly supplied analyses, even if they are more limited in scope than richer information that is produced more slowly.

— Most of the information needs of managers focuses on performance within a cycle, rather than across two or more cycles.

The *cyclical* data collected by managers serve several purposes. Various client audiences need to be informed about the performance of the system. The administrator monitors the progress of projects, assessing the need for changes in direction or inputs. Or the data may be used to detect problems that require the design of new projects.

In each case, the manager and decision maker seek information that is *specific rather than general*. It is a specific project that is evaluated and corrected, a specific audience that is informed, a specific set of inputs that are adjusted. The manager is not concerned about generalizing to all possible situations he may meet; instead he tries to solve specific problems that occur.

This concern for specificity is underlined to call attention to the distinction between the kind of information analysis described in this paper and the kind dealt with in units on social science research methods. Those latter methods are most helpful where the planner wants to know how to handle a *class* of problems. The methods treated in this paper will be most helpful when the planner has his problem well-defined and wants to diagnose some aspect of system operation.

2.0 *Can a Planning Office Be Useful to a Ministry?*

The question will not always receive an affirmative answer. Many planning offices do not fit into the organizational structure of the ministry, and planning is treated as an adjunct service to the operating divisions of the ministry rather than as a central part of the organizational process. Typically, this results in one of two outcomes. In some cases planners are totally isolated from the operation of the ministry and are limited to carrying out research that may eventually have some impact on policies of the ministry (but which more often is ignored). In other situations the *planners* are so overloaded with service responsibilities that they cannot do *planning*, that is, they cannot carry out the kinds of evaluations and assessments that, combined with goal-setting, allow for specification of future actions of the ministry.

One possible solution to this problem is to *redesign the ministry*. The office of planning could be located above all the operating divisions of the ministry, as part of the minister's st[illegible] with power to collect data and to impose its decisions. Late arriving [illegible] planning have gained this position of privilege in only a few co[illegible]

Another alternative is [illegible] *planning into* each of the *operating divisions*. This can be made to happen even where the formal office of planning remains as a service unit adjunct to the operating divisions. Each division head develops his own planning staff that not only collects but also analyzes data and produces plans based on those analyses. This kind of solution is, however, suboptimal. It wastes the resources invested in the office of plan-

ning and impedes a system-wide coordination of educational effort.

A third alternative is for existing offices of planning to *break out of the trap* in which they find themselves, the trap in which either they are isolated or overloaded with a series of tasks that limit the extent of their contribution to the system. This breakout requires first an analysis of the information requirements of the various decision makers and managers in the system and an assessment of unmet needs. That analysis requires the planner to understand the perspective of the manager.

To summarize:

— Planning often is difficult because of the position of the office in the ministry.

— Planners may be able to define a more effective role, by an analysis of information needs of managers and decision makers.

2.1 *The Educational Planning Office in Paraguay*

The Department of Educational Planning (DPE) had for some time been impatient with its inability to influence the operation and direction of education in Paraguay. The country is undergoing profound change, and education must be organized to meet the challenges of the future. But the DPE found itself overwhelmed with the annual routines of data tabulation, often unable to respond to urgent outside requests for information, let alone carry out studies that would form the basis for planning.

To understand the situation we must first look at how education is organized now, and what is happening to the country.

2.1.1 Problems Facing Education in Paraguay. The structure of the educational system of Paraguay is conventional in design. A six-year primary school system feeds into a three-year "basic" secondary cycle. In principle, students can opt for a three-year terminal vocational or agricultural program, but few students do so, and the official objective is to phase out these programs at basic level.

Senior secondary education is given in four kinds of three-year institutions, all of which allow for continuation after finishing secondary school. There are two autonomous universities in Paraguay, with programs ranging from four to seven years. Teacher training is done in postsecondary institutions run by the Ministry of Education and Worship (MEC).

Table 6.1 presents enrollment figures for the years 1971 through 1975, for the three levels. Primary education has grown slowly, and for three years did not increase at all. The MEC estimates that about 85% of the school-age population attend school. This estimate is based on the 1972 Census; there is no regular school age census. Secondary and higher education have grown rapidly in comparison with primary school, but these levels still enroll a small proportion of the population attending school, as Table 6.1 indicates.

Despite the apparent lack of dynamism in primary school enrollments, there have been significant changes in the distribution of students between urban and rural areas. As Table 6.2 shows, the trend in Paraguay is counter

Table 6.1:

Enrollment by Level, for Paraguay: 1971–1975

Year	Primary	%	Secondary	%	University	%
1971	436,987	86.5	58,130	11.5	10,182	2.0
1972	443,680	85.7	65,552	12.1	11,642	2.2
1973	451,580	85.0	66,746	12.6	12,682	2.4
1974	454,852	84.1	71,619	13.3	14,317	2.6
1975	452,673	83.0	75.424	13.8	17,199	3.2

Source: Ministerio de Educación y Culto y Anuario Estadístico 1971/75

to that experienced in most countries. In 1975 there were proportionately more students in school in *rural* areas than in urban areas, while i: 1965 the opposite was the case. The table also indicates that enrollments in private primary schools are increasing, also contrary to the experience in many countries.

Table 6.2:

Percentage Distribution of Enrollment in Primary Schools, by Zone and Sector: Paraguay, 1965–1976

Year	Zone		Sector		Total	
	Urban	Rural	Public	Private		
1965	52.3	47.7	89.6	10.4	100.0	(356,998)
1970	47.4	52.6	87.1	12.9	100.0	(424,179)
1971	46.4	53.6	86.5	13.5	100.0	(436,857)
1972	45.4	54.6	86.4	13.6	100.0	(443,680)
1973	43.9	56.1	86.6	13.4	100.0	(451,530)
1974	43.6	57.6	86.0	14.0	100.0	(454,853)
1975	43.3	57.7	85.3	14.7	100.0	(452,249)
1976	42.2	57.8	84.4	15.6	100.0	(465,540)

Source: Ministerio de Educación y Culto

The explanation of this phenomenon appears to lie in the unique geopolitical situation of Paraguay. Paraguay shares with its neighbors, Brazil and Argentina, a river boundary, and its vast potential of hydroelectric power compensates for the lack of any natural resources other than good land. Relatively underpopulated, Paraguay's positive balance of trade now comes exclusively from forestry and agriculture, but soon could include industry.

In recent years Brazilian farmers have been attracted to Paraguay in increasing numbers, and have purchased large extensions of land in Paraguay along the border with Brazil. This situation threatens the political

stability of Paraguay. The government's response has been a massive program of land reform, including settlement of large numbers of colonists (some Paraguayan, others European and Oriental) in areas along the borders with Brazil and Argentina. These areas of colonization have officially been designated priority zones for the nation. All-weather roads have been built, political and physical services provided, and other efforts made to increase the presence of the government in these areas. The government's official policy is to favor education in these areas over all other regions of the country.

The MEC is formally responsible for all primary and secondary education in the country. There are 15 agencies or departments that report to the Director General of Education, who reports to the Minister. In addition, the directors of each of the departments report directly to the Minister. The Department of Planning is one of these 15 agencies, on the same organization level as the operational units (Primary, Secondary, etc.) as well as the administrative agencies (Personnel, Construction, etc.).

2.1.2 Planning Office Response. Concerned about its ability to respond to information needs of the various departments in the MEC, the Department of Educational Planning (DPE) hired consultants to confer with the directors of the departments and their staff, "to describe and analyze the flow of information and the process of formulation and or taking of decisions" within the MEC. The consultants interviewed each of the directors of the major departments, in each case asking about relationships with the DPE, problems experienced in making and carrying out decisions, needs for information, and their own efforts to gather and analyze data for planning.

In their report the consultants concluded:

> There are various areas in which the work of the DPE could be improved. Some of these areas were pointed out by the Director or by members of the Department; some were indicated by directors of other departments; and some were mentioned by all. The areas mentioned can be put into two general categories: (1) planning and control; and (2) statistical information.

The report continued to suggest that a number of directors had observed that coordination among the various departments of the Ministry was a problem, that although there were a lot of data, there was not much information, and that someone needed to assume the role of coordinator if the Ministry was to move ahead. One of the directors cited in the report said:

> We should have a mechanism to develop concrete plans of operation . . . also the plan should indicate (1) the needs of the country, and (2) a policy to follow with respect to these needs, as well as different alternatives. The Office of Planning should tell us. This is possible and this is not.

Another was quoted as follows:

> We need more help from them (DPE). They are the brain of the entire system. In its role as rudder, Planning should be doing the planning that will help us carry out programs. In order to do our planning for next year we need some guidance from the Office of Planning as to what we should do.

Noting that although many called for coordination by the DPE, it was likely that any immediate steps to exert control would meet sharp resistance, the report recommended beginning a long-term process of development of communication and control links within the Ministry. The first step in the process would be the improvement of the statistical analysis services of the DPE.

> The steps recommended for immediate action are concentrated in improvements of the information system of the Ministry. The main reasons for beginning at this point are: (1) This area represents a strong-felt need for various departments, especially Primary and Secondary, and the DPE itself. (2) This is a sphere of activities generally recognized as an intrinsic responsibility of the DPE. (3) The topic in itself is not controversial. (4) There is a strong chance of success. The director of the DPE agreed with this analysis and the process began.

In summary:

- Political and economic factors beyond the control of the government are forcing changes in Paraguay, and in turn changes in education are required.
- To assess its contribution to the system, the Planning Office studied the process of decision making in the Ministry.
- This study showed that although there were data available, lack of analysis meant no coordinated effort.
- The Planning Office was seen as a candidate for the role of coordinator.

3.0 *Why Ministry Statistical Systems Don't Meet the Planner's Needs*

Most offices of educational planning are mislocated because they developed after ministries of education had already set up their basic bureaucratic control mechanisms, and after ministries had begun to collect statistics about their operation. Although many countries did not have national systems of public education until the twentieth century, the rudiments of public administration were laid down (in most cases) during the colonial periods. Few countries had educational planning offices prior to 1960. Planners and researchers trained in "modern" techniques of collection and analysis (in many cases in countries with bureaucratic traditions quite different from their own) find themselves up against a statistical system the rationality of which is often difficult to detect.

But even if there were not historical differences in the development of educational bureaucracies and offices of planning, there would be differences in the approach to statistics and data-gathering. *Most planners* face situations in which their *task is one of change*, plotting a new course for the system, either because previous targets have not been reached, or because new targets have been set. *Administrators* in the operational divisions, on the other hand, most often define their *task as one of maintenance*, and some-

times as efficiency. They are concerned with executing orders that have been given to them and carrying out the predetermined tasks of the educational system. Sometimes they are concerned with doing this at least possible cost, or with minimal waste of resources.

Administrators are concerned about *control* of the operation of an ongoing system. Planners are concerned with the *allocation of resources* to produce a new variant of the system. Planners are concerned with history in terms of its implications for the future while administrators are concerned with its implications for the present. If the planning task is defined in terms of long-term forecasting and policy planning, differences in information requirements become obvious. Policy analysts and planners are more concerned about *total system* (or program) performance than about operating characteristics of elements within a system or *program*. The need for specifying the distribution of resources, and the long cycle in the feedback loop that makes correction difficult or impossible, pushes planners to opt for highly reliable data. The administrator, closer to the scene and responsible for the small details of operation, wants to know *not* how the total system is moving, or even how a given program is working. His immediate concern is whether a given element in a program is contributing what it should.

Administrators typically cannot be concerned about outcomes or outputs of programs. They are into the next cycle of operation before there is time to look at the effects of the previous one. Their solution to this problem is to follow highly specified programs of operation and to worry most about inputs.

Given that statistical systems in most countries were developed by operational division administrators, it is not surprising that the kind of data they collect appears of little utility to the planner. Most statistical systems are concerned with keeping tabs on the basic inputs to the educational process: numbers of students, teachers, classrooms, books (if there are any). There is little concern for assessing the *quality* of inputs; it is their quantity that is important. The director wants to know the number of teachers and whether they met their classes, because that information is linked to the payroll. Or he wants to know the number of students per section, not as a mere total, but to determine whether the teacher assigned to that section should be kept there, or the section eliminated because it is too small. There is often a close tally of teachers' qualifications, not to determine whether better teachers are assigned to urban or upper-class areas, but rather to determine if the director is paying the appropriate salary for that level of training.

In summary, then:

- Planning offices were created long after educational bureaucracies were formed; hence planning often is ignored in administration
- Furthermore, most administrators are concerned about maintenance problems, while planners deal with change. Hence, they have different information needs.
- The statistics collected by ministries often do not, therefore, seem useful to planners.

3.1 *Data Collection in Paraguay*

The educational statistics compiled by the DPE are collected through two forms, one for primary education, and one for secondary education. In addition, DPE also processes some data provided by the Department of Regional Centers (12-year institutions covering both primary and secondary education) and by Technical-Vocational education.

These are the only cyclical data collected by DPE, but the MEC collects more. For example, the Personnel Department maintains a personal file on every teacher in the system. The Construction Department has information on schools in terms of available classrooms and physical condition. The Guidance Department makes a continuous assessment of student achievement in pilot programs. The Materials Department has records on the distribution of textbooks in school districts. The Curriculum Department has data on the experiences of new curriculum projects. Both Primary and Secondary Education collect data other than that included on the forms used by the DPE.

Although in principle these other data sources are available to the DPE, *in fact* only the data collected directly by them are readily accessible. *No systematic effort* has been made by DPE *to gain access to data from other departments*. DPE staff have little or no way of finding out about the existence of these data, because their own normal workload makes it impossible to visit with other departments, and because until very recently (March 1978) there had been no regular form of communication across departments within the MEC.

As it is, the DPE is in fact only a *processor* of data about primary and secondary education. The data actually are collected by supervisors and directors attached to each of these departments. A standard "planilla" (data form) was used to collect data through March 1978. The "planilla" collected information in March (beginning of the school year) on each school site identified by number, name, town, district, and province, and whether or not the institution is public or private, rural or urban. It indicated whether the school functioned in morning, afternoon, or evening session, or a combination of these, and asked for the number of sections which functioned for the morning, afternoon, or evening session.

The basic data were collected in five major blocks. The first block provided information on all teaching (regular and special) and directorial personnel. This included age, sex, civil state, category (training level), grade taught by morning, afternoon, or evening session, number of students, number of assignments, payment, years of service, and years in the present school. The block had 23 rows and 12 columns, and at maximum it would yield 276 cells of information.

The second block showed personnel movement in the month of March. It indicated the names of those who leave or join the school by date, and noted the reasons for leaving (transfer, retirement, removal).

The third block recorded the number of students by grade who appeared for examinations, and whether they passed or failed. Results were classified by the sex of the student.

The fourth block was the standard age-grade enrollment matrix, classified by sex. The fifth block was in the same form, but showed the number of students by age and grade who repeated the grade. These yield perhaps the most basic and essential information for planners who must assess the flow through the system and the performance of the system. The age-grade matrices of the fourth and fifth blocks are shown in Figure 6.1, because these are the fundamental educational data for planners.

Figure 6.1:

Data Collection Forms for Age-Grade Matrices of Enrolled and Repeaters, in Paraguay

(a) Students enrolled by grade, sex and age

Block IV

Age	Grades							
	Pre-School	1	2	3	4	5	6	Total
	M F	M F	M F	M F	M F	M F	M F	M F T
Less than 7 Years								
7								
8								
9								
10								
11								
12								
13								
14								
15+								
Total								

(b) Students enrolled who repeat[a]

Block V

Age	Grades							
	Pre-School	1	2	3	4	5	6	Total
	M F	M F	M F	M F	M F	M F	M F	M F T
Less than 7 Years								
7								
8								
9								
10								
11								
12								
13								
14								
15+								
Total								

[a]A repeater is a student enrolled two or more times in the same grade in a different school year.

The form was first printed before the creation of the DPE in 1968. Statistical data handled by the DPE are collected at two points in the year, corresponding to the opening of the academic year in March, and to the close of the year in November. Separate forms are used. The March form is printed in January of the year in question. The decision on the line item for the printing of forms must be made in June of the previous year and transmitted to the Ministry of Finance prior to the finalization of the national budget in September.

About the 1st of February the Departments (Primary and Secondary) distribute the forms to their supervisors. Each supervisor is given three forms for each of the "schools" under his control. For purposes of the data collection procedure, a "school" is defined as a morning or an afternoon session. About half the schools in the country are on double shift, and therefore fill out two or more sets of the forms.

About the 14th of February supervisors begin to distribute the forms to the directors of the nuclear ("cabecera") schools, who then distribute them to the satellite schools. According to the rules, the forms are to be returned to the nuclear school by the 15th of March, two weeks after the opening of classes in the 1st week of March. The forms then are sent up to the supervisor, who returns them to the capital, in principle by the 5th of May. One copy of each set of forms remains with the school, one with the supervisor, and the final copy is sent to Primary (or Secondary). This process is described in Figure 6.2.

The forms pass first through the Planning Unit in the Department of Primary (or Secondary) Education. The unit takes from the forms information concerning class size, which is used to determine if additional teachers must be appointed in some schools, or if some classes are so small that they should be collapsed and the teacher reassigned. Although it is possible to create new positions for teachers during the year, each position so created must be justified with a petition to the Minister of Education and the Ministry of Finance. Hence, relatively few positions are created during the year. Instead, Primary makes every effort to anticipate in the annual budgeting process how many new teacher positions it will need for the coming year.

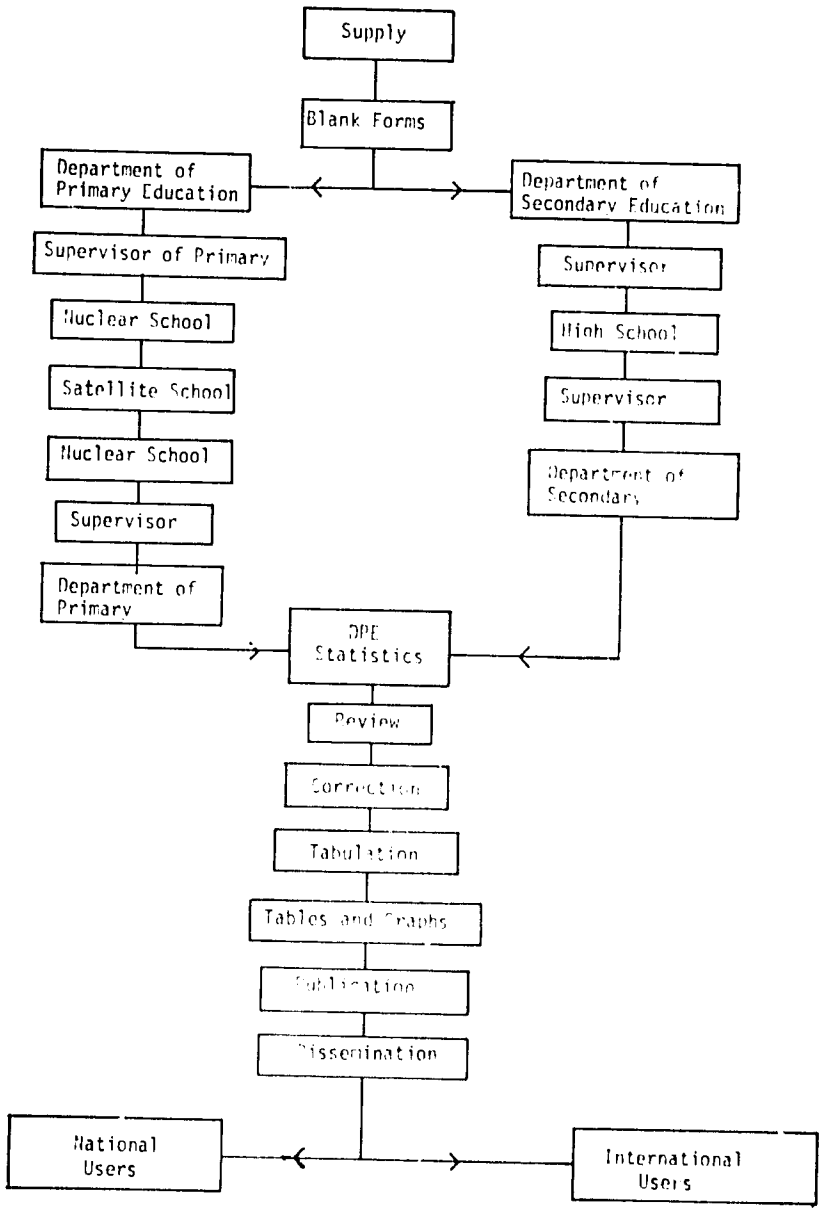
After information is taken from the forms, they are passed on to the DPE, which then begins its own process of tabulation. About 85% of the forms usually arrive at the DPE by August. The DPE normally begins tabulation of the data forms at that time, following its involvement in the annual budgeting process in June and July.

Some of the procedure of tabulation is described in Figure 6.2. Because the volume of work is so much greater for data for Primary than Secondary, it is this process which is described below. The forms are first completed by Province (or State), and by District within Province. Each school has a number. Historically, this number distinguished between complete (all six grades) schools and those that are incomplete (without all six grades). Because schools tend to become complete over time this number is no longer useful for that distinction. There are now more than 1000 complete schools, whose numbers begin with 1001, but incomplete schools were numbered beginning with 1001.

Paraguay has 19 provinces and 45 supervisors. The provinces average 15

Figure 6.2:

Flow of Statistical Data From Source, for Paraguay



districts, but the number changes from time to time according to movement of enrollments. For example, provinces that have been losing enrollments now have fewer districts than they did several years previously.

With the forms organized by province and district, a tally is taken and compared with the list of schools from the previous year. If a form is missing, a query is sent to Primary. If the DPE has received information about the creation of a new school, that name is added to the list. This process of organizing the forms, and working toward a more complete set of data forms, takes place during a period of approximately 35 days.

The next 30 days are spent correcting the numbers on the data forms. The DPE staff review all the totals, and "square" (make row totals equal column totals) the tables. The staff have desk calculators. Some of the calculators are electronic and have a memory, but this function is not used in the tabulation process.

The next period, about 30 days in length, involves the first round of tabulation of the data. Each person in the Statistics Unit of the DPE is a specialist in one or more provinces, or in one or more of the tables to be produced from the forms. The tabulation process involves transferring data, one school at a time, from the basic data form to a work sheet. These work sheets are organized by districts, within provinces. Each work sheet is summed at the bottom and the sums collected to produce totals. Districts are classified as urban or rural (corresponding with civil definitions) and schools as public or private. There can therefore be four work sheets for a given district. The tables generated from the work sheets present sums for each province, broken down by urban vs. rural and by public vs. private. The data sheets are then compiled again and summed to produce national totals. Figure 6.3 is an example of one of the 12 basic work sheets that are used.

This process generally is finished in rough form by the end of the calendar year. The month of January is used to tidy up the results, prior to taking summer vacations in February.

A different form is used to collect data at the end of the academic year in November. The form repeats all the questions on directorial and teaching personnel. It asks for the names of personnel who left or entered the school after March, and the reasons for those moves. It asks for average attendance by grade in the month of July. It records the number enrolled in March and new students enrolled each month after that by sex and grade. Finally, it records the number of students passing and failing final examinations by grade and by sex.

The processing and tabulation of these forms often begins in late December and continues into the following year. After the required tables are put together, the Statistics Unit personnel begin to construct the annual statistical report. A draft version of this is in circulation by August (of the year following the year in which the data were collected). After the draft version has been reviewed several times, it is sent to the reproduction unit, where it is produced on mimeograph stencils. The final version is distributed in early March of the third year.

The head of the Statistics Unit described some of her problems in the

Figure 6.3:

Basic Work Sheets for the Compilation of Educational Data in Paraguay

Total of Teachers' Assignments by Category and Special Teachers

Prov.

COD. No.	Teachers								Directors								Special Teachers					Totals			
	1	2	3	4	5	6	7	Total	1	2	3	4	5	6	7	Total	Art	P.E.	M.	O.	Total	T Tea.	T Dir.	T Spec.	Total Gen.

following way:

> The major problems we have are these. First, Primary holds the data forms when they are returned in order to calculate the number of over-age students by zone. Second, the computation of desertion rates is questionable. At times we have negative rates, that is, the final enrollment is larger than the original enrollment. Third, the data are not reliable, especially in Asunción (the capital). The teachers add on students in order to hold their jobs, that is, they overestimate the enrollment.

In addition to producing the annual statistical report, the DPE also responds to requests for information. Most of these come from within MEC, with others from Defense and other ministries, or from private institutions and individuals. All requests are honored, but some take much longer than others. The Director of the DPE has for some time wished to improve the DPE's ability to respond to these requests, especially those coming from the Minister or from other departments within the MEC. Work on the annual statistical report often is interrupted, but, even so, the time to meet requests for information is considered too long.

The *DPE had, until this project, never done any research using the statistical data base*, although information in the annual statistical reports often was used for diagnostic studies required by the international agencies.

In summary:

- The DPE processes basic statistical data for Primary and Secondary. This is the major activity of the office.
- Processing is delayed because Primary and Secondary use the data themselves before handing it over, and because the DPE has to interrupt tabulations to meet requests for information.
- Tabulation is done by hand, on all statistical forms received from Primary or Secondary. There are some doubts about the quality of the data.
- DPE makes no use of the data base for planning, although it does use tabulations printed in the statistical annual

4.0 *How to Choose Variables for Analysis in Statistics*

Faced with the difficulties of managing and analyzing a large and awkwardly designed data system, designed to meet needs now perhaps forgotten, the planner has at least *three possible responses* available. The first is common but not useful. The planner can *despair*, and forego ambitions to contribute to the direction of the Ministry through systematic analysis of data. The Office of Planning becomes essentially an Office of Statistics for the Ministry, a general service organization that contributes little to detection of problems and possibilities, or to the generation of a more effective organization. This response is not uncommon. Davis (USAID Harvard Project Paper 74) and McGinn and Warwick (USAID Harvard Project Paper 75) describe how the earlier versions of the educational planning office in El Salvador were limited to data-gathering.

The second option is to convert the Office of Planning into an Office of *Research*. This may follow from despair (i.e., "We can't do anything anyway, so at least let's do some interesting research"), or it may be seen as a means of producing long-term improvements in the education system. Planners may feel that their contribution to the system is best made over the long term, through research findings that demonstrate the costs and benefits of programs; or forecast demands for education and training of a different type or of greater or lesser quantity; or demonstrate more effective ways for implementing programs. Snyder (USAID/Harvard Project Paper 83) describes how this came about during one stage in the development of the educational planning office in Peru.

As noted in the beginning of this chapter, social science research can serve an important need and contribute to educational development, as it becomes more closely linked with policy and decision making in the Ministry. This is more likely, it would seem, if the Office of Planning plays an active role in research design and analysis, rather than giving over the task to an external agency more distant from the operation of the Ministry.

But the time required to do effective long-term research tends to impede the involvement of the Office of Planning in other kinds of policy and decision making in the Ministry. By its nature, research tends to last longer than one cycle and to impose requirements that isolate researchers from the action of developing a system. These tensions are discussed by McGinn and Warwick (USAID Harvard Project Paper 19) in their paper on Sector Analysis in El Salvador.

The third option is to *develop analytical procedures* (not called research here in order to make the distinction from research, which tends to be theory-based) that can be carried out in synchronization with the decision-making cycle, and that can take advantage of the existing data bank of the Ministry. Rather than collect more data, it is possible for planners to analyze more thoroughly what they have. Although antiquated and inefficient data systems, as those described, cannot answer all questions, they can be made to provide information of more immediate use in policy making and to suggest better research designs for longer-cycle studies.

What are the *major problems in using existing data sources?* There are at least three. First, the *sheer volume* of statistics collected by ministries makes analysis difficult. Any program to use these sources to analyze system operation would require sampling procedures. A second difficulty is the apparent *low level of reliability* of information; there have been few studies to assess the adequacy of official statistics. One of the first steps would have to be an assessment of sources and magnitude of errors.

A third problem is that many of the variables included in most statistical systems *seem inappropriate* for the types of questions that planners might want to ask. Certainly the *variables* on statistical forms are different from those that researchers would design: they have no theoretical origins and appear to be conceptually and operationally complex. The inappropriateness of variables is probably the most difficult problem facing the researcher who would use these data to inform decisions.

The variables of major interest will be those related to concerns of decision

makers and managers. Chapter 5 in Volume I and USAID/Harvard Project Paper 25 discuss the questions likely to be most pertinent, as, for example: Should a given program continue, or should it be changed? If it is to be changed, what are the critical elements needing attention? What quantity of resources should be allocated to this program rather than to another? Is the failure of a given program serious enough to warrant action?

The procedure described here differs from what might be called a conventional research approach. Conventional research properly begins by collecting a set of questions of the kind listed above. It then proceeds to identify the kind of information that would be useful to the decision maker to answer those questions. The analysis proposed here begins on the other end. Because the data have already been gathered, the planner-researcher first has to ask, "What do I know?" and then look for problems to which that information can be addressed.

That statement is an exaggeration, of course. The first pass through the data will deal with only a few variables, and the selection of which variables to examine will be determined by some notion of what most needs attention. But it is important to recognize that, because in this mode of research the planner does not have control over the data collection process, the procedures for analysis are different from conventional research. The researcher is not in this case analyzing data to see how they apply to a conceptual model already developed (and for which research instruments were specifically designed). Rather, he analyzes data to see what models might be built from them, what problems they might help to resolve. Or, in other words, emphasis is placed on seeing the fullest possible range of implications of the data — inventing hypotheses that could be proved or disproved — rather than a straightforward application to hypotheses already stated.

This emphasis requires that the planner-researcher think constantly in terms of both the *operating system* to which the data refer and the *policy instruments available* to the decision maker or manager. These are reflected in the data. Even though they were not designed for that purpose, the researcher has to see the linkage between the statistical data collected through the Ministry's forms and elements of the educational system under study. The process requires more imagination than is involved in conventional research.

In summary:

- Planners need not *withdraw from the field*; they can contribute directly to the improvement of education, through analysis of existing data.
- To be useful to managers, data analysis must include variables that managers manipulate through policy and decision making.
- To identify these variables, one must look at the functioning of the system. In contrast to conventional research, one *develops* models from the data rather than using the data to test predetermined models. (This is shown in earlier chapters.)

4.1 *Variables Included in the Analysis in Paraguay*

If anything, the data forms used in Paraguay provide too many variables. Figure 6.4 presents a partial list of the variables that could be taken from the March form alone. What appears as a simple device is capable of generating more information than the DPE can handle.

Figure 6.4:

Variables that Could be Taken from the March Data Form

Province
Public-private
Urban-rural
Shift (and shifts)
Number of classes
Number of teachers
Proportion of men/women teachers
Age of Teachers
Civil state
Level of training of teachers
Class size
Total enrollment
Stability of Teachers
Presence of specialized professors
Characteristics of the principal
Students taking make-up exams
Students who pass/fail make-up exams
Enrollment by grade by sex
Average age by grade
Students over or under-age for grade
Number of repeaters by grade
Ratio repeater/total enrollment by grade

The DPE had, however, at least two means for winnowing out the number of variables to analyze. First, the interviews with the directors of the departments in the Ministry, and with the Vice-Minister, described the information they use in making decisions, and the kind of information they would use were it available. The directors of Primary and Secondary education made specific requests for analyses that would help them in assigning teachers.

A second source of insights into the variables to analyze came from the head of the Statistics Unit in the DPE. This person's years of experience with the system, and with the data, gave her insights into what variables were likely to be most important. She stated:

> Region or province is most important, because of the differences in growth. For example, Alto Parana is growing very rapidly . . . On the other hand, Cordillera, Caazapa and Paraguari will probably lose enrollments . . . A second factor is zone that is, urban or rural. All new schools are being built in rural areas . . . Sex doesn't vary much across the country. There doesn't seem to be any systematic variation. But people who ask for data always want breakdowns by sex. About 85% of the schools are public. Only Caaguazu shows some variation from the national average. In every new colonization the schools begin as private. Caaguazu has a lot of new colonizations. After a year, the colony tries to get the Ministry to make the school public and take over its support. Often these are very small communities where the Rural Development Agency won't build a school. As a result, there are a lot of incomplete schools and the number continues to increase.

The DPE staff used these two sources to develop a statement of what would be important to include in their analysis of the statistical data. The following is an excerpt from a report published by the DPE.

Selection of Variables

> At this time we want to examine some of the factors that could play an important role in determining student achievement. Also included are some factors that traditionally have been considered important, but which many studies done in other countries have suggested have only a minor importance.
>
> *Turnover of teachers*, for example, could be disconcerting for students who see contradictory elements in what one teacher demanded and the requirements of his replacement.
>
> The *availability of specialized teachers* in some areas could make an important contribution to the formation of the students. Not all teachers are capable of teaching Art, Music, or Physical Education.
>
> The use of *multiple-grade classes* (two or more grades taught simultaneously in the same room) is related to the Ministry's desire to provide complete (all six grades) schools in all areas. The growing interest of the Ministry in the use of multiple-grade classes justifies its inclusion as a variable.
>
> *Class size* in Paraguay is considered to be important, with 30 as the desirable upper limit. On the other hand, many studies (in other countries) have failed to find any significant relationships between class size and student achievement.
>
> *School size* often is a factor in the level of efficiency of use of resources — large schools tend to be more efficient. In this case, we also are interested in how school size is related to the four variables named above.

Hypotheses

To the extent that this is an exploratory study it does not require a detailed theoretical framework. On the other hand, previous studies in Paraguay and other countries provide at least some guidelines as to how to carry out the analysis of the data to be collected. *The results will be examined as well in terms of the tendencies or relationships they suggest that could guide future studies.* [Emphasis added] The results will, of course, be compared with the data from the 1974 and 1975 samples yet to be drawn.

1. School size should have an influence on the presence of specialized teachers. The larger the school, the more likely it is that there will be specialized teachers, because a minimum number of students are required to justify hiring these kinds of teachers.

2. School size should influence class size. The smaller a school is, the more difficult it is to have average size classes, and the larger a school is, the more likely that classes will be large. We can also expect that as school size is smaller, there are more multiple-grade classes. Class size also should be related to the presence of multiple-grade classrooms, as this may be used as a means to get classes of a minimal size (i.e., large enough to justify a teacher).

3. The zone will influence the presence of multiple-grade classes. The dispersion of population in rural areas will make schools smaller in total enrollment, and increase the number of multiple-grade classes and incomplete schools (which motivate the Ministry to create multiple-grade classes). Another way of looking at this will be to see if urban areas have more schools that use two or more shifts, supposing that a school goes from one to two shifts when there is pressure to enroll more students.

Besides the "independent" variables named above, we selected four classificatory variables. These can be used to describe some basic characteristics of the schools in the sample.

Province: using data grouped by province allows an examination of the representativity of the sample with respect to variables usually tabulated by the Statistics Unit.

Sector: the categories "public" and "private" can be used to control the influence of different sponsorship on the relationships studied.

Zone: the categories "urban" and "rural" are indispensable to examine the problems of geographic dispersion that preoccupy planners and administrators in the Ministry. These categories are especially important when looking at determinants of incomplete schools, school and class size, and multiple-grade classes.

Number of shifts: this can be treated either as an independent variable or as a classificatory variable. We are especially interested in seeing whether double shifts work better than multiple-grade classes in terms of providing complete schools.

In summary, the salient points are:

— The Paraguayan data forms provide a rich source of variables for analysis.

— The DPE used two means for choosing variables.
  a) Interviews with decision makers as to what information they use and would like to have;
  b) Knowledge of trends in system performance.

— Variables were chosen because they might be related to student performance, an important objective for the system.

— All the independent variables could be directly affected by policy; in addition, classificatory variables were included.

— The variables were used to describe how the system operates, i.e., to produce a "model" of the system, stated in the form of hypotheses.

5.0 *Sampling the Data Source*

In the previous section we described a means for reducing the complexity of overly rich data sources, by means of focusing on certain problems of greatest concern to decision makers. These problems are identified not by examination of theory or models but by talking with those concerned with management of the system. Models are then built using policy variables that managers can manipulate in an attempt to find solutions to those problems. As these models use only a limited set of variables, the planner is freed from the burden of having to analyze a very long list of possible measures.

In this section we describe the second step taken to reduce complexity. In this step the planner can *sample the sources* of data so that the analysis that will be done will contend with a much smaller mass of data than contained in the original data base.

This sampling is done principally to speed the analyses that the planner will carry out. It takes considerably less time to tabulate 300 data forms than it does to tabulate 3000. The facility of tabulation has two other advantages. First, it is possible that the reliability of tabulations improve because more time and care can be spent on recording results. Second, with released time the planner can perform more analyses than would otherwise have been possible. He can tabulate his variables in more ways, and he can return to the data base and draw out other samples and make more significant comparisons.

For this to be done effectively, the planner must anticipate the kinds of analyses that will be of greatest interest and use. The *sampling framework* should itself include one or more of the classificatory variables of interest. The size of the sample must be sufficient to provide data at an acceptable level of reliability. (In general, these levels are considerably more relaxed than those imposed by social science researchers.)

Concern for *reliability* is one of the reasons most frequently cited by persons who continue to analyze data drawn from the *universe* or all the data sources. A *census*, in which an attempt is made to count every unit in the universe, appears intrinsically to be more trustworthy than a sample, especially one based on a small fraction of the total.

In some instances, that intrinsic appeal is justified, as, for example, when it is important to know the *exact* number of cases in a total. But that situation almost never holds for the busy administrator, who needs to know only whether he should do more or less of something, who has neither the time nor the control to measure out in exact drops the dosage that will just remedy the system's problem.

For almost all decisions that are made in a Ministry of Education, the information required need not be exact or precise, and a well-drawn sample is sufficient. (In fact, in the case of estimates of enrollment, it is possible that in some systems a sample of schools would do a better job of estimating the real total than is done by requiring — but not getting — enrollment data from every school.)

What are the attributes of a well-drawn sample? The major characteristic is that the good sample is *representative* of the universe from which it was drawn. Under most circumstances, this representativeness can be achieved by drawing a *simple random* sample from the universe. By *random* is meant that cases are drawn in such a way that every case has an equal chance of being drawn into the sample. *Simple* means that cases are drawn directly from the entire universe.

It is possible, of course, to deliberately draw an unrepresentative sample, in which certain kinds of cases or units have a greater likelihood of appearing. This would be a good thing to do if you were trying to study some phenomenon that had a low rate of occurrence: by weighting the sample to include 2, 3, or more times as many cases of that type than others, you would increase the possibility of ending up with enough cases to study.

Sometimes it is helpful to draw a *stratified* sample. If you know that variation is distributed among the cases in relation to some classificatory variable, then sampling randomly within the categories of that classificatory variable would allow you to draw a smaller sample than you would have to draw using a simple sampling technique, in order to have the same reliability. For example, if the inputs that schools receive vary considerably by political district (e.g., state, province, department), and you want to look at inputs, then you could sample within these districts. More information on sampling is contained in Volume III and in the Appendix at the end of this volume.

If simplicity is the order of the day in the drawing of a sample, it also makes sense when proceeding to tabulation of the data. Many of the analytic techniques used in conventional social science research projects are so complicated that they cannot easily be carried out by hand. And not only do they require expensive data-processing equipment, they also require the insertion, between the planner and the data, of an *expert* who can handle the equipment. For many, and perhaps most, of the planners who would talk to managers, this degree of complexity is unwarranted.

Most of the analyses that will be of use to managers and decision makers require no more than simple cross-tabulation of variables. In most cases, tests of statistical significance are of little use — the problem is either so clear that one needs not measure "tendencies" or it is not there. Small samples, and use of discrete (e.g., public vs. private) instead of continuous (e.g., number enrolled) variables, facilitates simple tabulation and comparison. Doing analysis by hand, or with simple data-processing equipment (e.g., card sorters) allows the analyst to "play" with the data, and encourages thinking instead of reliance on a computer to somehow sort out the important facts. With practice, the analyst can begin to "interact" with the data as if, in fact, he had built an operating "model" of the system.

In summary:

— The amount of data to be handled can be reduced by sampling the data base.

— For most purposes, the simple random sample is best.

— For most purposes, the best way to tabulate data is the simplest way, by hand.

5.1 *Sampling the Statistical Data Base in Paraguay*

5.1.1 Design of the Sample. A random sample was drawn of 10% of the almost 3000 forms sent by primary schools in March 1976. First, the forms were separated by province. For each province, the 3rd form was drawn, as well as every 10th form following (e.g., 13th, 23rd, etc.). This produced a total sample of 305 schools (or forms), slightly more than 10% because of rounding. One province (Amambay) was excluded from the analysis because the data forms were not found. The total time spent on the drawing of the sample, using seven persons, was two hours.

5.1.2 Characteristics of the Obtained Sample. The sampling procedure appears to have generated data that are faithful to those obtained from the census of all schools in the country. This can be appreciated by looking at Tables 6.3 through 6.5, where results obtained from the sample are compared with previously tabulated results from *all* the data forms. Data are presented in percentages to facilitate comparisons.

According to the census, 83.1% of the primary schools in Paraguay are public; the sample reported 83.6%. The census reported that in 1976, 57% of public schools were complete (i.e., had all 6 grades). Using data from the sample only, one would have concluded that 59% of the public primary schools are complete.

Table 6.3:

Distribution of Schools by Number of Grades and Sector, for Universe and Sample

Highest Grade Offered in the School	Sector			
	Public		Private	
	Universe	Sample	Universe	Sample
1st and 2nd	14%	14%	26%	16%
3rd and 4th	29	27	28	34
5th and 6th	57	59	46	50
Total	100%	100%	100%	100%
(Cases)	(2459)	(255)	(501)	(50)
Distribution	.831	.836	.169	.164

Table 6.4:

Distribution of Schools by Number of Grades and Zone, for Universe and Sample

Highest Grade Offered in the School	Zone			
	Urban		Rural	
	Universe	Sample	Universe	Sample
1st and 2nd	3%	1%	19%	18%
3rd and 4th	3	4	35	36
5th and 6th	94	95	46	46
Total (Cases)	100% (545)	100% (73)	100% (2415)	100% (232)

Table 6.5:

Distribution of Schools by Sector and Zone, for Universe and Sample

Sector	Zone			
	Urban		Rural	
	Universe	Sample	Universe	Sample
Oficial	66%	71%	87%	88%
Private	34	29	13	12
Total (Cases)	100% (545)	100% (73)	100% (2415)	100% (232)

Differences between the universe (census) and sample estimations are greater in the comparison of private schools. The census reported 26% of the schools with only first or second grade, while the sample suggested that only 16% of the schools were in this category. "Errors" of this size are to be expected given the number of cases in the sub-group of private schools in the sample. Using the formula $SE = \sqrt{(p \times q)/n}$, in which SE is the standard error of the estimate proportion, p is the proportion of cases of one type (and q the proportion for the other) and n is the number of cases, one can calculate the *confidence intervals* of proportions of dichotomous variables. In the case above, the standard error of the proportion of cases of private schools with only grades 1 or 2 is: $(.16 \times .84)/50 = .052$. A reasonable assumption is that about 95% of all the different samples of size 50 that could be drawn would have p values that are two SE greater or lesser than the p value obtained.

Therefore, one can be 95% certain that a given proportion of 0.16, for a sample of 50 cases, will vary between .056 and .264 (.16 ± 2 X .052). Thus, the "error" in the sample estimation of this proportion was within the confidence interval that one would have specified. Two points are to be drawn from this. First, as sample sizes go down, the size of errors goes up. Second, it is possible to estimate the largest possible error that would be made.

As it turns out, the "error" described above is the largest discrepancy between the estimations obtained from the sample and the values of the census as reported in Tables 6.3, 6.4 and 6.5. In other words, the sample appears to do a pretty good job of estimating the results that were obtained by carrying out the laborious procedures required to tabulate all 3000 data forms. For those cases in which the sample provides results *not* tabulated for the universe, the best estimator of the universe value is that obtained for the sample, and the formula for sampling error can be used to estimate confidence intervals.

5.1.3 Tabulation Procedure. Given the small number of cases in the sample and the fact that all of the variables were categorized, that is, either dichotomized or with three levels, it was possible to carry out all the analyses by hand.

The separation of the sample forms by provinces was maintained throughout the tabulation process. For each province, the forms were divided into two groups according to whether the school was public or private. Each of these groups was then divided into two, according to whether the school was located in an urban or a rural area. Seven persons, working two hours, were able to do this for all 305 forms.

In the next step, each of the forms was identified by writing in large letters on its face, indicating the number of shifts in the school (one, two, or three) and the highest grade available in the school (grouped also in three levels). With the completion of this step, it was then possible to divide all of the data forms by province, public/private, rural/urban, shifts, and highest grade. Figure 6.5 presents a "tree" showing the process that separated the forms into the 24 groups *possible* for each province.

Some of the possible groups did not in fact exist, and others had very few cases. The three levels of shifts were collapsed into two (one shift, two or more) because there were so few schools with three shifts. Only one of the 24 groups had more than 10 schools, and most of them had four or less. The small number of forms in each group made it possible to classify other variables within that group mentally, that is, without having to tally them or add them on a calculator. The persons working with the data found it easy to add cases "in their heads", only writing down the sums for each variable on work sheets (see Figure 6.6 for one example) that were prepared in advance for each province. As can be seen, the work sheet allowed, in addition to the five variables used to make the groupings, tabulation of two more variables. (And another work sheet was prepared that added another variable in the left-hand column.) In other words, by filling out this work sheet, the staff was tabulating *six* variables at once for each province.

Figure 6.5:

Steps in the Selection and Tabulation of the Sample

Figure 6.5: Steps in the Selection and Tabulation of the Sample

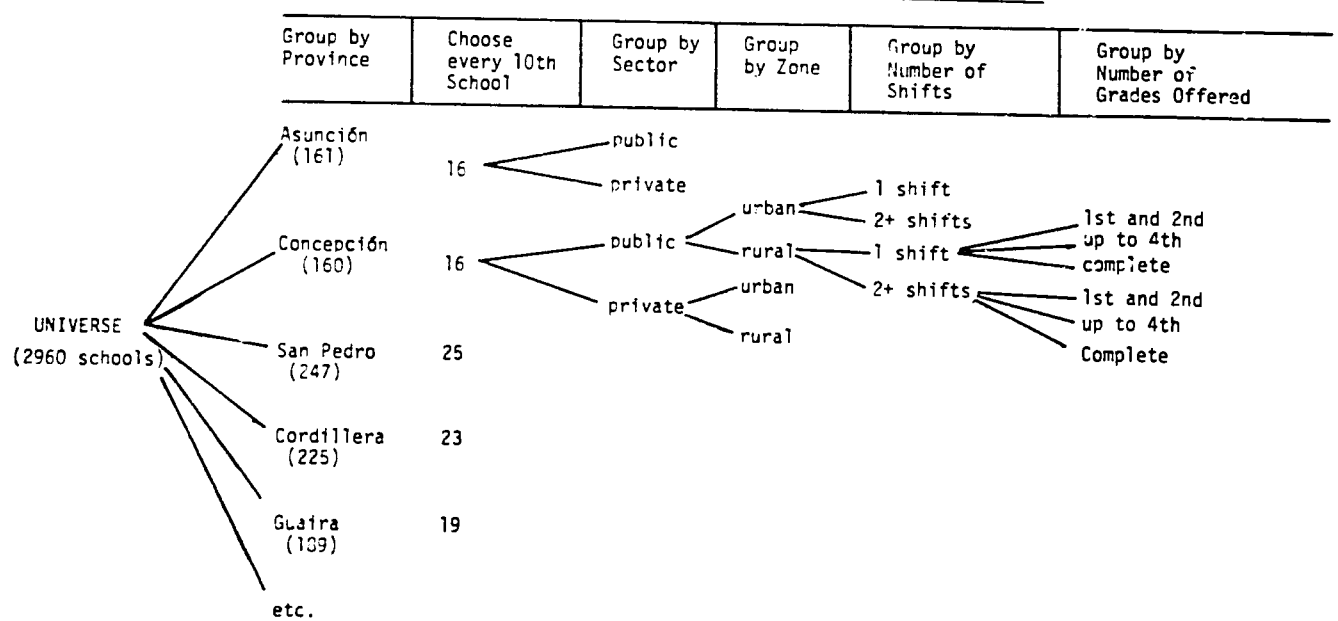


Figure 6.6:

Worksheets for Tabulation of Sampled Data

PROVINCE: _____

SHIFT	HIGHEST GRADE	PUBLIC								PRIVATE								TOTAL
		Urban				Rural				Urban				Rural				
1 SHIFT	1st & 2nd																	
	3rd & 4th																	
	5th & 6th																	
2 SHIFT	1st & 2nd																	
	3rd & 4th																	
	5th & 6th																	
3 OR MORE SHIFT	1st & 2nd																	
	3rd & 4th																	
	5th & 6th																	
TOTAL																		

The major source of difficulty in the use of this kind of work sheet is getting each sheet to "square," that is, to obtain equal totals summing the rows or the columns. Because the main classificatory variables were the same from analysis to analysis, once the correct column and row totals were obtained for the first work sheet, it was easy on successive sheets to spot the location of an error, to identify the cell in which it appeared, and therefore to return to the original data form for the correct value.

Once work sheets for all the provinces had been filled out, the staff tabulated the totals for each cell and produced a summary sheet for the nation. This process of aggregating data across provinces was facilitated by using heavier lines every two or three cells.

Once the national totals were obtained, the staff proceeded to put together tables like those that follow. The addition of more variables from the data forms was a relatively simple matter because once the row and column marginal totals for the provinces and the nation were obtained for the classificatory variables, it was easy to make rapid and accurate tabulations. The entire process of sampling, sorting, tabulation, and preparation of approximately 30 tables, beginning with an untrained staff, took less than one week. At the end of the week, the staff was able to continue the process on its own without supervision.

In summary:

- A random sample was drawn from the data forms collected in March 1976, by selecting every 10th school within each of the departments.
- This sample produced results that compared closely with those obtained from the universe.
- Tabulation was done by first sorting the sampled forms into groups by province; then each group was sorted again by sector, then each of those groups again by zone; and again by number of shifts; and finally by number of grades.
- Tabulation on the variables selected was done by hand by province, cross-tabulating up to 6 variables at a time, by using the groups formed in step 3.
- The entire process, sampling, sorting and tabulation, and the training of staff to do it, took less than one week of work. An equivalent analysis using all the data forms probably would have taken 6 months.

6.0 *Analysis Procedures*

Unlike conventional research designs, which generally are intended to "close" alternatives by providing definitive answers, the analytical process recommended for the planner is one in which problems are discovered. The effort is to generate more hypotheses rather than to test those invented before the research began. The planner-analyst seeks to learn new ways to model the system rather than to test whether his existing model is adequate. At the end point, the two approaches do not differ — they both seek to improve

models and specify how to improve a system. But the operations research model used here reaches that end by induction, while the conventional social science approach is basically deductive. Previous chapters used statistics to "expose" the problem. In operational terms, this means that the planner analyzing basic statistical data tries to follow variables "through" the system's cycle.

Strictly speaking, the inductive process could begin anywhere. But some time is gained by beginning with policy variables as suggested earlier. The planner looks to see how variations in the level of a policy variable are associated with other factors in which he is interested. Ideally, he has some measure of desirable system outcomes (e.g., student achievement). In many cases he will have some proxy for that, or merely assumptions about what will lead to favorable outcomes. He goes through his set of variables looking for those that seem to be most highly associated with his outcome measure, discarding those that seem to make no difference. Then he treats those policy variables themselves as outcomes and looks for variables that are associated with them, either as inputs, or as conditioning factors. He exposes, selects, compares, rejects, or keeps variables, attempting, with each selection, to increase the ability of the "model" he is constructing to "explain" what is going on. Recall the additive model and the steps of "polishing" the fit, or the use of data transformations to improve the descriptive quality of a model.

The sampling and tabulation techniques described earlier facilitate this process because they allow the planner-analyst to run through a number of variables in a relatively short period of time. He is encouraged to try out a variety of hypotheses, to dig deeper and deeper into the operation of the system as seen through the basic statistical data.

This approach differs from the "fishing expedition" or exploratory research of the social sciences, in which the investigator tabulates and examines at one time a number of variables, chosen without any serious attempt to model or conceptualize the system. In the approach recommended here, the planner-analyst begins with a small set of variables, chosen because they do make some operational or conceptual sense. If these variables do not "work," he abandons them and looks for others. Eventually, he may have gone through a long list, but in the process he develops a coherent model of the system.

This may be a clear case where one can only "learn by doing." But some consideration of what was done in Paraguay may be useful to further illustrate the perspective. We will conclude by discussing a few of the many analyses done in Paraguay.

6.1 *Are Multiple-Grade Classes and Double Shifts Useful in Paraguay?*

Using the sampled data described previously, planners in Paraguay decided to evaluate policies concerning the use of double shifts and multiple grade classes. The following is an excerpt from the report they prepared, based on the analysis of these two policies.

One of the problems that most concerns the Director of Primary Education is how to provide access to school to those children in rural areas who currently do not attend school. The problem has several facets. First, the dispersion of the population: in rural areas it is difficult to put together enough students to justify having all six grades. On the other hand, it is known that the incomplete school encourages repeating and dropouts, and that many children finish schooling without having completed the primary cycle. Sometimes, to justify the assignment of teachers to an upper grade (where there are fewer children), the Ministry will put together children from several grades — for example, a 5th and 6th grade are taught by the same person.

In other cases the problem of access is one of space. The Ministry cannot build schools fast enough to handle the demand (and for that reason private education is growing more rapidly). There are cases where there is enough space in the school, but not enough teachers because there are not enough funds to hire them. In both cases the double shift would appear to be a reasonable solution, as it permits a better utilization of the space that is available and of the teacher who is already in the rural area.

The DPE decided to study how these two different policies have been applied. Have they served to reduce the number of incomplete schools? Double shifts have been used for a number of years, and multiple-grade classes became more common several years ago. In the last three years of their use, what have been the results? Could the Ministry of Education use them more widely to increase access to school?

In order to answer these questions, the DPE used a type of 'panel' design in which data forms for the same schools were analyzed for 1974, 1975, and 1976. The expectation was that the comparison over time would show an increasing use of multiple-grade classes (Ministry officials presumed they were being used with greater frequency), with a consequent impact on enrollments. A quick glance at the data for the three years showed that, in fact, no significant changes in use had occurred. As a result, the more detailed analysis was limited to data only from 1976.

First, the DPE looked for the relationships between multiple-grade classes and the number of complete or incomplete schools, by zone. Table 6.6 shows that almost all of the urban schools are complete, and that of these 93% do not have multiple grade classes. In other words, complete schools in urban areas do not result from the use of multiple grade classes.

In the rural area, most of the schools are incomplete and do not have all 6 grades. Of the complete schools, 45% have multiple-grade classes, while 54% of the incomplete schools have such classes. One could ask, if multiple-grade classes are intended to make it possible to offer all six grades, why are they more common in incomplete schools?

One possible answer is that use of multiple-grade classes has something to do with the total number of students enrolled or with school size. One might expect that in very small schools, even the use of multiple-grade classes would not make it possible to offer all six grades (i.e., they would probably be used for the lower grades). Table 6.7 shows the analysis done to look at that possibility. First look at the small schools, those with 70 or fewer students. Those with multiple-grade classes are more likely to be complete than are those that do not have multiple-grade classes: 74% of those without

Table 6.6:

Distribution of Schools by Zone, Number of Grades, and Use of Multiple-Grade Classes

School Situation	Urban Zone		Rural Zone		Total Number of Cases Estimated To be in Universe
	Incomplete	Complete	Incomplete	Complete	
With multiple Grade Class	(1 case) 7%		54%	45%	1210
Without	(3 cases) 93		46	55	1840
Total	N.S.	100%	100%	100%	3050
Distribution	1%	23%	41%	35%	100%

N.S. Insignificant number of cases

Table 6.7:

Distribution of Schools by Size, Multiple-Grade Classes, and Number of Grades

Grades Offered by the School	School Size (Enrollment)						Total Number of Cases Estimated To be in Universe
	Up to 70		71 to 150		151 and more		
	Multiple Grade		Multiple Grade		Multiple Grade		
	Yes	No	Yes	No	Yes	No	
1st or 2nd	23%	74%	0	4%	0	0	420
Up to 4th	54	21	48	57	0	1	880
Complete	23	5	52	39	100	99	
Total	100%	100%	100%	100%	100%	100%	3050
Distribution	21%	11%	13%	15%	6%	33%	100%

multiple-grade classes have only first and second grade, in comparison with 23% of those with multiple-grade classes. Much the same appears to be the case with medium-size schools – from 71 to 150 students – although the relationship is not so marked because of the distribution of cases. Once a school has 150 or more students, it appears to become complete with or without multiple-grade classes. One could conclude from this table that, independently of multiple-grade classes, small schools are more likely to be incomplete, but that multiple-grade classes do help to make it possible to offer all six grades, even in very small schools.

If this is the case, one might then ask why it is that not all schools offer multiple-grade classes. Table 6.7 can be used to calculate that there are about 248 schools in Paraguay with 70 or fewer students, with only first or second grade, without multiple-grade classes (.74 × .11 × 3050). Why does this situation exist? Is it a result of lack of space or lack of teachers?

There are approximately 144 small schools with multiple-grade classes and only first and second grade (.23 × .21 × 3050). Why does this occur? By consulting another table [not shown here], it is possi-

ble to see that all of these schools operate at only one shift a day, which suggests that in that case, space is not the main problem.

The DPE then put together another table to look more closely at the result of double shifting. The results are presented in Table 6.8. The assumption is that double shifts are used when there are more students than can be handled in the existing space in one shift. Only rural schools are included in Table 6.8, since we have seen that the problem of incomplete schools is principally a problem for rural people. As Table 6.8 shows, *all* single-shift schools are incomplete, whether they have multiple-grade classes or not. One can conclude that these incomplete schools have space, but either fail to attract enough students to operate all six grades, or do not receive enough teachers.

Table 6.8:

Single or Double Shifts in Rural Schools, by School Situation

School Situation	One Shift		Two or More Shifts		Total Number of Cases Estimated To Be in Universe
	Incomplete	Complete	Incomplete	Complete	
With multiple Grade Class	72%	--	48%	45%	1150
Without	28	--	52	55	1170
Total	N.S.	--	100%	100%	2320
Distribution	13%	--	41%	46%	100%

N.S. Insignificant number of cases

About half the schools that operate two or more shifts are incomplete, and there is little difference between complete and incomplete schools in terms of whether they have multiple-grade classes or not. One might conclude that adding another shift permits schools without much space (e.g., only 2 or 3 classrooms) to offer all six grades, even when each grade is taught separately. About half the complete schools do not have multiple-grade classes; one would presume that two shifts are necessary in these schools because there are not enough classrooms to offer all six grades in one shift. If that is true, it is also likely that more students could be served in those schools by building additional classrooms. There are approximately 586 rural schools in Paraguay in this situation (.55 × .46 × 2320).

On the other hand, there are approximately 457 schools that are incomplete, even though operating two or more shifts and with multiple-grade classes. In these schools, one would presume, the problem is lack of students because of the isolated location of the school. Why are there so many schools (.52 × .41 × 2320 = 495) that do not have multiple-grade classes, even though they obviously have enough students to operate two shifts?

Finally, Table 6.8 calls attention to the fact that there are 84 schools operating only one shift, and without multiple-grade classes, that are incomplete. These schools have enough physical capacity to

offer more grades (by running two shifts) — they either need to receive more teachers or to ask those currently in place to teach in multiple-grade classes.

These and other tables provide a rich source of suggestions as to how the Ministry of Education, and specifically the Department of Primary Education could begin to identify schools — and the zones and provinces in which they tend to be located — that need special attention. The tables provide an easy means of evaluating the consequences of policies such as double shifts and multiple-grade classes, as well as possibilities for extending the application of those policies.

In addition to the report from which the above was extracted, the DPE wrote four other reports based on the analysis done during one week, using the sampling and tabulating techniques described above. One of the reports constructed age-grade matrices for several consecutive years, and used that information to recalculate repeater rates. The rates obtained by using this technique appeared to be more reliable than those usually reported; they were about twice as large as the rates estimated using the census data, which underestimated repeaters and enrollment load.

Another report looked at the distribution of class sizes in schools, and showed that many schools have small class sizes, even though there are many students. By increasing these classes in size to the present average, the Ministry could receive about 12,000 more students (2.5% of the present enrollment) without hiring another teacher or building another classroom. In a third report, it was concluded that specialized teachers are not distributed according to where they are most needed but according to other criteria.

On the basis of the experience gained in this exercise, the DPE has redesigned, together with Primary and Secondary, the basic statistical forms used in the Ministry. Primary has decided to change some of its policies with respect to assignment of teachers. Consideration is being given to the use of such data in the annual budgeting process.

7.0 *Conclusions*

This case study has reviewed the experience of one Office of Planning in the analysis of existing Ministry statistics for the purpose of informing decision makers and managers. The objective of this review has been to show that it is possible to use statistics, even when of apparent limited reliability, to shed light on aspects of operation of the system. The recommendation is not that the planner consider these statistics as the only possible source of data for planning; there are many other kinds of statistics available in the Ministry of Education, and for some problems conventional research is the best and necessary source of data. But Ministry statistics can be used, and that use can be fairly inexpensive and quick.

The case study has shown how a relatively untrained staff, with no expensive calculating or computing equipment, was able to organize its data files to conduct useful analyses in a brief period of time. The analyses carried out are within the reach of any ministry of education, regardless of the size of the educational system or the state in which the data are collected and stored.

Furthermore, the case study has shown how this approach to data analysis can serve as a training experience for staff and lead to progressively more sophisticated analyses. Two other comments on the Paraguayan experience should be noted. First, as a result of a close look at their data, DPE staff members began to notice systematic irregularities in the data reported. They formulated a series of hypotheses as to the causes of these irregularities, in general with considerable skepticism about some, because the speculations were based on assumptions that the teachers deviated significantly from Ministry rules on registration of students. A visit to various schools to observe the process and test the hypotheses taught the DPE staff a good deal about what really happens in the field. As a result, they are changing their procedures.

Second, one of the most dramatic changes has been in relationships with other departments of the Ministry. As a result of these analyses, and the circulation of information produced, other departments are now beginning to see the DPE as contributing to Ministry objectives. The DPE has found as a result that it is now possible to request and expect collaboration from departments that previously ignored communications.

Finally, the case study demonstrates once again a cardinal rule of planning. Serendipity is a major source of advance in planning, despite its rationalist basis. The good planner does everything possible to allow serendipity to appear, and to be in a position to take advantage of it when it does. The kinds of analyses described in this chapter, and in the previous chapters, uncover serendipities, recognize them, and permit action to be taken more rapidly.

APPENDIX

Selecting a Simple Random Sample

1.0 *Why Sample and What to Look for in a Sample*

Situations may arise where the analyst is unable to deal with the entire population of interest. It may be, for example, that the universe is very large and collecting and analyzing information would be excessively time consuming and expensive. Or it may be that the analyst does not have access to all the units in the population. In such circumstances, a sample or subset of the entire population is used, and information is collected and analyzed for that subset.

In selecting a sample, one is concerned that the sample is *representative* of the larger population. For example, one would not want to exclude girls from a sample of secondary students if statements about both boys and girls are to be made based on the sample data. One is also concerned with the *precision* of the sample selected. Precision refers to how closely a statistic computed from the sample (for example the mean of one of the variables) would approach the same term if the data for the entire population were used. Obviously we strive for increased precision when selecting a sample. While it is not possible to know exactly how precise a given sample is, there are methods to estimate the precision.

There are a variety of ways of selecting a sample so that it is likely to be both representative and fairly precise. In this appendix, only one such method will be discussed. Called *simple random sampling*, it is quite commonly used when selecting samples. It is called simple because there are more elaborate methods of random sampling. It is called random because each unit in the sample is selected in a random fashion from the population

In fact it is the random component which normally insures that the sample is representative. The assumption is made that a truly random selection process would eliminate biases in selection that may cause the sample to be unrepresentative of the entire population. For example, suppose the population of interest is a single class of students. We might non-randomly select a sample of these students by choosing those in the first row of seats, or those who volunteered to be in the sample. Under such circumstances we stand a larger risk of introducing biases into the sample; the sample may not reflect the actual composition of the class. It may also yield statistics that are not very precise. If, on the other hand, we had some random process for selecting a sample of students, the mere randomness of the selection process may eliminate reinforcing factors that cause lack of precision or unrepresentative samples. It does not entirely eliminate problems, however, because it is still possible to select a sample which might yield a statistic that would not be close to the population value.

2.0 *Tables of Random Numbers*

One technique that is commonly used to select a random sample requires the use of tables of random digits. Table A.1 shows several hundred random digits taken from a book of such numbers. Their use in drawing or selecting a sample is really quite simple:

a. The population to be sampled must be clearly identified. Is it all 4th grade students in the country; all women who teach in institutions of higher education; all schools in a given city; or all academic departments in a university?

b. A listing must be made of each unit within the population. For example, in the North Central data set, there were 436 school districts. If we wished to draw a sample from these districts, each district would first be listed, by name.

c. The sample is selected from the population using a random sampling procedure. This can be done in a variety of ways. For example, if the population units are written down on separate pieces of paper, placed in a container, and well mixed up, then selection can be made by drawing out the appropriate number of units from the container.

A convenient procedure for selecting a random sample relies on tables of random digits such as those shown in Table A.1. First, each of the units in the population list is assigned a sequential number, say from 1 to the number of units in the population. For the North Central data, each school district would be given a number from 1 to 436, one number per district and no number occurring more than once. Secondly, we use the table of random numbers to select which of these districts will be represented in the sample.

To do this we must first agree how to read the random digits. This could be done in a variety of ways, but only one need be selected. It, however, must be used consistently in drawing any one sample. Note in Table A.1 that the random digits are arranged in columns of five digits. To draw a sample for the North Central data, we will need 3 of these digits in any 5-digit sequence, because our numbers only range to a maximum of 436. If we were to draw a sample with 1100 people in the population, then we would need to use 4 digits from each 5-digit sequence. Note that the population unit assigned the number 7, for example, would be represented as the 3-digit sequence 007, or 78 as the 3-digit sequence 078.

Suppose we agree to read the last 3 digits in each 5-digit sequence, and agree to read down a column until the bottom of the page is reached, then moving to the column to the right of the starting column and to the top row in that column. But first we must make a random start in the table. This is done by closing the eyes and putting the finger down on the table of random digits. Wherever the finger lands will be the first 5-digit sequence to be used. We then read down the column, retaining all 3-digit numbers less than or equal to the maximum number in the population (e.g., 436 or less in the North Central data). One number may only appear once. For example, if we draw the number 142 twice, the school district assigned that number will be

Table A.1:

A Table of Random Digits

97536	10086	85020	01457	36123	63298	94553	34885	96718	02866
61364	50594	03771	54929	10697	78180	41958	77570	48798	50099
33815	44213	79352	90824	00982	84185	07982	60782	46932	41577
12380	71341	07728	89117	17776	68254	97580	91291	98582	05278
18773	80244	08238	28556	33787	32665	58681	93348	95619	63540
20689	34533	36739	93357	93311	92935	50547	11202	67566	28765
37935	51951	62884	61210	99044	08740	63113	69419	85137	13273
52185	26595	47285	68194	03212	81678	35589	23620	77114	82437
56358	99115	55855	65467	75535	69053	86450	67627	31643	42905
21269	69886	61463	95656	45151	07085	71831	57161	35969	90243
05973	63390	60980	15719	69493	61678	68406	42297	83767	35940
01142	06250	83688	61275	43629	96727	92783	45253	74389	40718
65827	55202	46406	94370	63183	21405	57107	12552	71800	41209
78076	97548	60869	93203	86007	13341	61564	13987	36663	89124
51130	17431	72885	87836	74686	87221	98267	70301	77489	59410
55696	70981	25750	92783	06806	31277	00231	78365	31082	64261
06877	41782	98528	75421	40631	75505	65036	37900	50105	54363
52413	39539	[illegible]	95054	62670	02883	29955	29859	06486	15642
82069	45994	[illegible]	47062	36298	43167	58227	88397	59718	92825
69123	03166	[illegible]	52408	11175	37165	96239	41992	24483	80629
04657	18785	[illegible]	12267	75890	84585	80666	58079	36696	96150
33729	52930	46087	81507	34444	06738	10502	26495	37555	58693
27375	99350	96162	96477	40068	99641	24504	11771	48149	44411
05799	48043	73863	27601	89225	02633	58188	76026	06776	03168
57629	07568	20655	49808	59111	95332	67985	10376	[illegible]	18976
02469	52207	[illegible]	06942	97839	60910	96017	41711	93883	20372
22322	87127	67410	79471	52193	24147	12964	27836	65238	31631
41764	78277	12900	29663	70248	29336	86928	21351	02894	16396
91266	51881	40595	22374	61612	62649	42017	61396	09897	[illegible]
17089	29965	38282	67098	62633	92070	94014	88900	24099	36332
13979	29626	44887	14088	89315	32728	57835	49346	55093	69179
02814	39801	99695	61001	52515	60220	72569	57583	26157	39463
42758	84138	81386	37046	31102	61625	41398	04785	53781	40565
26137	38253	21094	86058	76160	03696	33680	86648	63277	31125
65831	16379	25421	18829	70302	21343	75097	55501	53686	89903
05304	83419	57630	41873	80428	58395	83511	82283	19572	06136
85147	06545	15737	85315	77842	94927	72439	24390	36376	34539
77075	47053	90309	35516	50862	88326	09885	67118	12600	96005
95042	02173	88237	04547	06492	98156	18563	03693	53335	67559
85900	10914	56422	30785	63352	28668	94760	33898	19134	53955
28598	45814	72041	83478	06101	52584	57170	71881	37797	62962
32598	22489	97691	53206	77568	15070	65174	47873	81838	81434
12754	33283	50766	95608	26106	81417	55207	18496	58643	22786
26224	60672	37069	89927	24409	07257	29808	37029	21608	11433
03084	47595	12282	64559	50487	37878	58708	32092	24322	30188
71590	71151	55141	28586	77919	57023	50355	02603	05849	21896
21725	19925	16634	48190	87721	73833	37374	22351	81961	34338
66492	85232	52169	93745	02169	20468	17935	11916	78073	26975
98979	77646	08709	47698	65642	58880	60521	13038	73005	53568
71951	59733	86619	00099	62052	49022	49047	11196	36895	25605

Source: Rand Corporation. A Million Random Digits with 100,000 Normal Deviates. Glencoe, Ill.: The Free Press, 1955. Reprinted by permission.

represented in the sample, but only once. We agree to discard the second or subsequent duplicate numbers, continuing to draw numbers until we reach the size of the sample desired.

Suppose we work through an example showing how to select a 10% sample of the North Central school districts, using a table of random digits. Further suppose that the random start procedure resulted in the 5-digit start of 46406 from Table A.1, which is in the 3rd column and the 13th row. The last three digits are 406. This is less than 436, so we agree to keep this as part of the sample, meaning the school district assigned the number 406 in the population list will be a member of the sample. Reading down the third column as we agreed to do, the next number is 869. This exceeds 436, so is not used. The next number less than 436 is 032, indicating that the 32nd unit in the population will be a member of the population. Continuing in this way we get the following list of random digits, as shown in Figure A.1(b). Those exceeding the number 436 are crossed out, while duplicates have been circled and all but one of each duplicate crossed out. In this manner we select 10% of 436 or 44 districts to be in the sample. The first 44 numbers drawn within the range of 001 to 436, and which are not duplicates, are used.

Figure A.1:

Using a Table of Random Digits to select a Random Sample

a Steps in the Process

Step 1: Close eyes and place finger on Table A.1 to get random start. Suppose it is 46406 in the third column and thirteenth row.

Step 2: Write down the last three digits in each five digit number, beginning with the random start, and proceeding down a column and across the page column by column.

Step 3: Reject those numbers greater than the highest number assigned to units in the population. In the North Central example there are 436 units in the population, so we will reject random numbers greater than 436. In the table in part b, these have been indicated by crossing them out.

Step 4: Use a stem and leaf diagram of the remaining numbers to check for duplications. A single number can only be used once. In the following table, duplications have been indicated by circling and crossing out all but one of each duplicate. Numbers not crossed out indicate the units to be selected from the population, for inclusion in the sample. A 10% sample of the North Central data means we will select the first 44 useable digits in the table.

b. Table of Useable Random Digits

406	162	386	~~766~~	~~824~~	370	~~477~~	046	~~608~~	~~982~~	183
~~869~~	~~863~~	094	069	117	203	~~601~~	058	~~927~~	~~776~~	007
~~885~~	~~655~~	421	282	~~553~~	~~836~~	~~808~~	~~829~~	~~659~~	~~787~~	~~686~~
~~750~~	~~920~~	~~630~~	141	357	~~783~~	~~942~~	~~873~~	~~586~~	311	~~806~~
~~628~~	410	~~737~~	~~634~~	210	421	~~471~~	315	190	044	~~631~~
032	~~900~~	309	169	194	054	~~663~~	~~516~~	~~745~~	212	~~670~~
~~908~~	~~595~~	237	~~709~~	~~467~~	062	374	~~547~~	~~698~~	~~655~~	298
~~535~~	282	422	~~619~~	~~656~~	408	098	~~785~~	099	151	175
~~333~~	~~887~~	041	~~457~~	~~719~~	267	088	~~478~~	123	~~493~~	~~890~~
087	~~695~~	~~691~~	~~929~~	275	~~507~~	001	206	~~697~~	~~629~~	~~444~~

3.0 *Sample Size.*

The previous steps will permit the selection of a simple random sample from a reasonably sized population. However, how does one know what size of sample to select? How many units from the population should be selected? This is a simple question to ask, and a very practical question too. However, it is not a simple question to answer. In fact the size of the sample to select depends upon the following things:

a. What we wish to estimate with the sample: a mean, proportion or some similar statistic.

b. How precisely we wish to estimate the statistic, that is, how close we wish the sample calculation to approach the population figure.

c. What sampling procedure was used to select the sample. A simple random sample will require a different number of units for a given degree of precision than will other sampling methods.

d. Finally, and most disconcertingly, the sample size depends upon the variation in the population of concern. Populations with wide variation in the variable of interest will require a larger sample for a given degree of precision than if the variation is smaller. Shown in Table A.2 are some sample sizes needed to achieve a precision of plus or minus 2%, in 95 out of every 100 samples drawn, for two different assumptions concerning the variation in a population variable X. Table A.2 shows how large a sample must be so that if 100 samples were drawn (each one using the full set of population units to select from), 95 of them would estimate the mean of the variable X within plus or minus 2%. There are two points of interest in Table A.2.

— First, to achieve a specified degree of precision, the absolute level of the sample size is more important than is the relative size of the sample compared to the population size. For example, to achieve the desired precision with a population of 100 and $S_X^2 = 0.01\bar{X}^2$, we must select 48% of the population to be in the sample. But to achieve the same precision with a population of 10,000, we must only select 93 units or 0.9% of the population. So, for a given level of precision, it is much more important *how many* units are in the sample than what percentage of the total population the sample represents. This is why it is possible to do a national survey with only a small number of people in the sample. It also indicates that one way to increase the precision of estimate with a sample is to increase the sample size.

— Second, note how the size of the required sample changes with assumptions about the variation in the variable of interest. To achieve a precision of plus or minus 2%, in 95 samples out of 100, requires 79 units in a sample from a population of 500 if the variance is $0.01X^2$, but 180 units if the variance is triple that.

Table A.2:

Sample Size Needed to Achieve a Given Precision, for Simple Random Sampling

Total Number in Universe	98% Precision, 95 Samples Out of 100			
	Number in Sample, Variance of $X=0.03\bar{X}^2$	Sample Size as % of the Population	Number in Sample, Variance of $X=0.01\bar{X}^2$	Sample Size as % of the Population
100	74	74.0	48	48.0
200	117	58.5	64	32.0
500	180	36.0	79	15.8
1,000	220	22.0	86	8.6
5,000	267	5.3	92	1.8
10,000	274	2.7	93	0.9
50,000	280	0.6	94	0.2

It is disconcerting that before one can select the proper sample size something must be known about the population that is to be studied. Usually this is done by running a small pilot study to estimate the variance in the variable of interest, or by using other studies to estimate the variance. Barring recourse to these types of studies, one could also simply use a rule of thumb which states that at least 100 units are necessary in a simple random sample to obtain an "acceptable" level of precision.

4.0 *Other Sampling Procedures*

The simple random sampling procedures outlined in the present Appendix will suffice for many situations. There are other, slightly more complex sampling schemes which are also often used. These include stratified, cluster, or multistage random sampling techniques. Descriptions of such techniques may be found in the texts listed in the Bibliography below.

In Chapter 6, the planners in the case study on the use of educational statistics in Paraguay used another very simple and inexpensive form of sampling called systematic sampling. When a list of units in the population is available, systematic sampling can be very convenient. The essential feature is that the list is arranged in some order and the sample is selected by agreeing to take a unit at periodic intervals. In the Paraguay case study

there were 3000 units (i.e., forms sent by primary schools) which were separated by province. The analysts wanted a 10% sample or about 300 forms selected for the sample. One in every ten forms was selected for inclusion in the sample, with the starting point randomly chosen. They could, for example, have written the numbers from 1 to 10 on pieces of paper, put them in a hat, stirred them up, and selected a number as the starting point. As indicated in Chapter 6, the starting point was 3. This the 3rd, 13th, 23rd, etc. form in each province was chosen to be in the sample. This results in a sample of the appropriate number, with a random element incorporated. If there is no order to the list of population elements, the sample drawn will essentially be a random sample, although any periodicities in the listing could make it a non-random sample and might bias estimates of parameters of interest.

Bibliography

In increasing order of difficulty, the following books all discuss aspects of sampling theory and practice.

Slonim, Morris James. *Sampling*. New York: Simon and Schuster, 1960.

Stuart, Alan. *Basic Ideas of Scientific Sampling*. 2nd ed. New York: Hafner Press, 1976.

Warwick, Donald and Charles Lininger. *The Sample Survey: Theory and Practice*. New York: McGraw-Hill, 1975.

Kish, Leslie. *Survey Sampling*. New York: Wiley, 1965.

Coda on the Passing of An Institution

These four volumes bear the institutional signet of the Harvard Center for Studies in Education and Development (CSED) which now, after almost two decades of activity in education and development, ceases to exist with the close of the USAID/Harvard Project and the publication of these books. Founded in 1961 as a research center in the Harvard Graduate School of Education, CSED later developed an extensive instructional program, at both doctoral and master level, in the specialities of education and development and educational planning. CSED trained over a hundred professionals and scholars. The majority of graduates were educators from developing countries. CSED also ran four major overseas projects and a number of smaller activities. The one major Cambridge based project was the USAID/Harvard project in educational planning methodologies, now ending after five years. From the overseas development and research projects have come a score of books and hundreds of research studies and journal articles.

During the nineteen years CSED functioned, Adam Curle, Russell Davis and Noel McGinn served as directors and William R. Charleson and Charles N. Myers Jr. were executive directors; and more than a hundred others have served as research staff, instructional staff in the planning program, and administrative and support staff.

With these volumes CSED ceases to exist. There is no irony in the fact that the Center, established to study the dynamics of development and the process of change, should itself undergo change and pass from the scene. What could be more natural? No creative artist has ever found an ending that improves on a good death. In its recorded work, and in the memories of those who were associated with CSED, nothing has ended; everything goes on.

USAID/HARVARD PLANNING PROJECT PAPERS

This is a list of titles and authors of 82 project papers (numbers 15 and 34 were never completed) written in the USAID/Harvard Planning Project. Slightly more than 50 of these papers appear in some form in the four volumes published here.

I. General Issues and Review Papers

Serial Number		
1	Russell Davis Barclay Hudson	Issues in Human Resource Development Planning: Research and Development Possibilities
2	Suzanne Prysor-Jones	Education and Equality in Developing Countries
3	Donald Snodgrass	The Distribution of Schooling and the Distribution of Income
4	Russell Davis	Planning Education for Employment
5	Russell Davis	Usefulness of Macro Level Manpower Requirements
6	Andre Daniere	Identification of Country Objectives, and Reconciliation with Objectives and Strategies of U.S. Assistance
7	Russell Davis	With a View to the Future: Tracing Broad Trends and Planning
8	Russell Davis	A Delphi Perspective on Future Assistance to Education in Developing Countries
9	John Beardsley	Beneficiary Participation in Rural Education Programs
10	Noel McGinn Donald Snodgrass	A Typology of Implications of Planning Education for Economic Development
11	Douglas Windham	Planning: An Alternative Conception
12	Samuel Bowles	A Class Analysis of Resource Allocation in Education
13	Donald Warwick	Integrating Planning and Implementation: A Transactional Approach
14	Thomas Welsh	Goals and National Purpose in Planning
15		(never completed)

II. Systematic Treatment of Issues and Methodological Problems

16	David Kline	Introduction to Nonformal Education and Its Planning
17	Donald Snodgrass Debabrata Sen	Manpower Planning Analysis in Developing Countries: A Critique

18	Andre Daniere	A Survey of Education Sector Assessment in Six Latin American Countries
19	Noel McGinn Donald Warwick	Sector Analysis in Education: An Analytic Case Study with Recommendations
20	Hans Picker	The Relevance of Education: A Programmatic Definition
21	Donald Warwick	Analyzing the Transactional Context for Planning
22	Jean DeHasse	Morphological Mapping: A Methodology for Increasing the Fit Between Planning and Implementation
23	Russell Davis	School Mapping: Planning School Services for Areas Within a Country
24	Noel McGinn	Types of Research Useful for Educational Planning
25	Noel McGinn	Information Requirements for Educational Planning: A Review of Ten Conceptions
26	Russell Davis	Information for Planning, and Analysis Capacity in Afghanistan
27	Noel McGinn	Data Sources and Information Available in Paraguay
28	Elizabeth Truesdell Mary Hyde	Information Systems
29	Russell Davis	Development and Use of Systems Models for Planning
30	Russell Davis	Policy and Program Issues Raised by the Application of Compound Models
31	Lascelles Anderson	Allocation Models in Educational Planning
32	Barclay Hudson	Compact Policy Assessment and the Compass Method: Practical Application of Dialectical Theory to Educational Planning and Forecasting
33	Ernesto Schiefelbein	School Mapping: Applied in the Chilean Educational Reform
34		(never completed)

III. Instructional Materials on Research and Planning

35	Russell Davis	Population Projection and Demographic Analysis: A Manual for Instruction and Reference
36	Elizabeth Truesdell Mary Hyde	A Computer Coded Model for Population Projection (An Interactive System)

37-39 Enrollment Projections in Educational Planning: Foreword

37 Russell Davis	Enrollment Projections in Educational Planning
38 Ernesto Schiefelbein	Estimation of Enrollment Flow Rates: Five Methods and Their Applications
39 Elizabeth Truesdell Mary Hyde	Computer Coded Models for Enrollment Projection, Targeting, and Efficiency Assessment (An Interactive System)
40 Russell Davis	Manpower Planning Methods, Unit I
41 Russell Davis	Manpower Planning Methods, Unit II
42 Russell Davis	Manpower Planning Methods, Unit III With an additional Unit IV Supplement on Curve Fitting

43-52 Basic Empirical Research and Evaluation Methods for Educational Planning

43 David Kline	The Relation of Educational Research and Educational Planning
44 David Kline	Problem Identification and Formulation
45 David Kline	Guidelines for Selecting A Research Method
46 David Kline	Planning and Conducting Educational Research
47 David Kline	Selecting the Subjects: Sampling (All Ages)
48 David Kline	Research Method I: Surveys
49 David Kline	Research Method II: Experiments
50 David Kline	Research Method III: Naturalistic Inquiry
51 David Kline	Research Method IV: Evaluation Research
52 David Kline	Research Method V: Cost Evaluation Techniques

53-62 Data Analysis for Planning

53 Gary Lewis	Data Analysis: Objective Science or Abstracted Empiricism
54 Gary Lewis	Categorization of Analysis Techniques
55 Gary Lewis	Visual Data Displays
56 Gary Lewis	Exploratory Analysis of Cross-Classified Continuous Data
57 Gary Lewis	Analysis of Categorical Cross-Classified Tables of Counts
58 Gary Lewis	Preparatory Steps for Computer Data Analysis

59	Gary Lewis	Log-Linear Models
60	Lascelles Anderson	The General Linear Model
61	Gary Lewis	Beyond the General Linear Model
62	Gary Lewis	Data Analysis for Educational Planning: A Case Analysis
63	Gary Lewis	Scheduling

64-68 The Planning of Non-Formal Education (Paper 16 is the Introduction to this series—Planning Non-Formal Education: Background to the Problem)

64	David Kline	Information Needs for Planning Nonformal Education (this could also go in Section II, information section, nos. 25-28)
65	David Kline	Planning Nonformal Education: A General Policy and Planning Assessment Model and Specific Techniques for Needs Assessment
66	David Kline	Policy Issues in and Objectives of Nonformal Education
67	David Harman	Planning Nonformal Education: Content and Curriculum Design
68	David Harman	Planning Nonformal Education: Instructional Methods and Design
69	David Kline	Information Systems to Support the Planning and Evaluation of Nonformal Education
70	David Kline	Planning Nonformal Education: Methods and Techniques

IV. Cases and Research Study Collections to Support Issues Papers, Systematic Papers and Instructional Manuals

71	Harvard Institute for International Development	Upper Secondary Education in Malaysia: Access and Enrollment Policy and Analysis of Effectiveness
72	Muchemwa Murerwa	Nonformal Education and Work: A General Perspective, and the Case of the Botswana Brigades
73	John Beardsley	The Youth Project at Ronkh, Senegal: A Cooperative Development Effort
74	Russell Davis	Planning in the Ministry of Education of El Salvador: The Organization of Planning
75	Noel McGinn Donald Warwick	The Evolution of Educational Planning in El Salvador: A Case Study

76	Noel McGinn Ernesto Schiefelbein	The Use of Educational Statistics for Policy Analysis: A Case Study
77	Russell Davis	Manpower Planning in Chile: A Case on the Basic Method
78	Russell Davis	Manpower Planning with an Alternative Method I: The Dominican Republic Case (Spanish)
79	Andre Daniere Russell Davis	Manpower Planning with an Alternative Method II: The Paraguay Case (Spanish version of Estudio de la Demanda de mano de Obra: Plan y Descripcion del Trabajo. This is accompanied by computer printed tables of results)
80	Noel McGinn Ernesto Schiefelbein	Contribution of Planning to Educational Reform: A Case Study of Chile 1965-1970
81	Russell Davis	Morphological Analysis Applied to a Systems and Organization Study: The Guatemala Case
82	Russell Davis	Small Business and Employment Creation in Calcutta: The CYSEC Case
83	Hugh Snyder	The Teachers College-AID Project with the Peruvian Ministry of Education: A Case Study with a Comparative Dimension
84	Russell Davis Noel McGinn (eds)	A Study of Educational Relevance in El Salvador