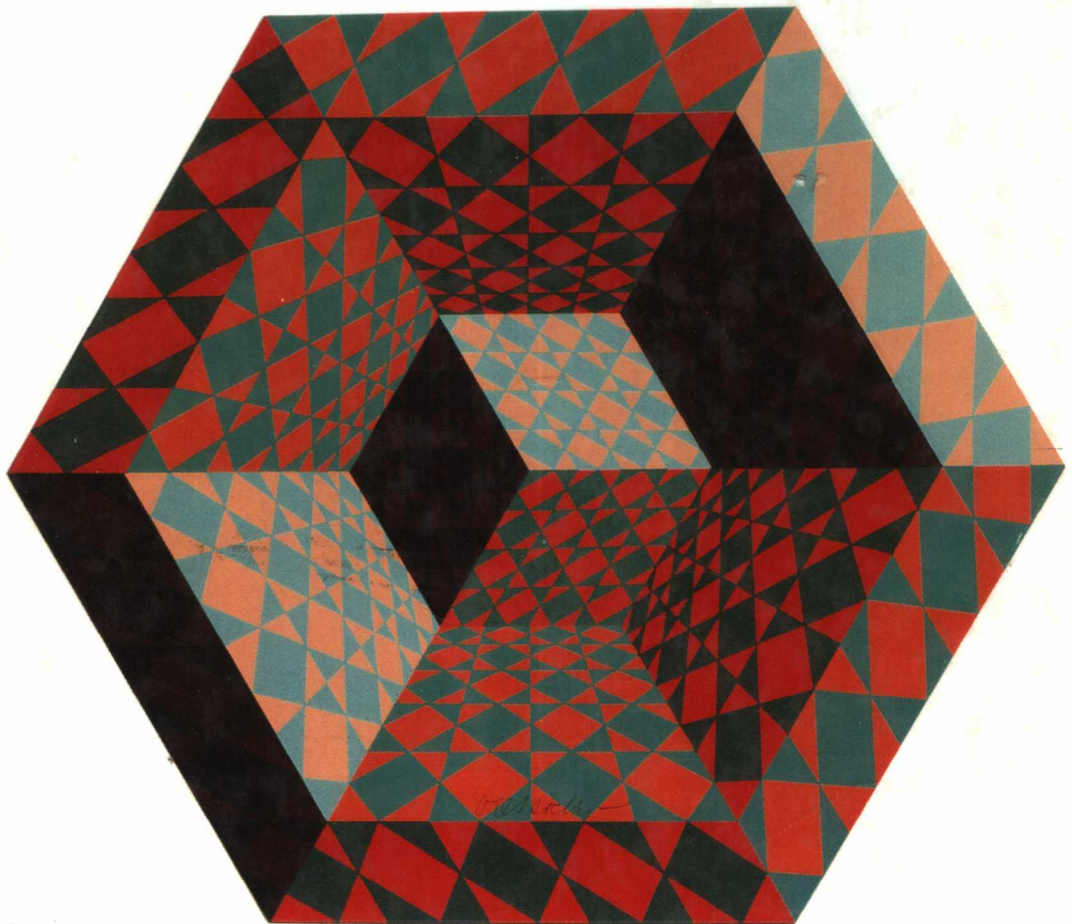


Statistics for Social Data Analysis

SECOND EDITION

George W. Bohrnstedt and David Knoke



Statistics for Social Data Analysis

SECOND EDITION

GEORGE W. BOHRNSTEDT

INDIANA UNIVERSITY

DAVID KNOKE

UNIVERSITY OF MINNESOTA



F. E. PEACOCK PUBLISHERS, INC. ITASCA, ILLINOIS

For our teachers, Edgar F. Borgatta and David R. Segal

*Cover: North Star (wood sculpture) by Victor Vasarely,
courtesy of the Vasarely Center, N.Y.*

Copyright ©1988, 1982
F.E. Peacock Publishers, Inc.
All rights reserved.
Library of Congress
Catalog Card No. 87-062162
ISBN 0-87581-323-2
Printed in the U.S.A.
56 Printing
2 Year

Preface

The second edition of *Statistics for Social Data Analysis* maintains the features that make this book different from other introductory statistics books for social science students. Our approach emphasizes statistics as tools for solving research problems rather than ends in and of themselves. We do not burden the student with excessive proofs and theorems, although the book *is* statistically correct. And consistent with the first edition, we do not include the obligatory chapter on probability; instead, we introduce concepts from probability theory, when needed, in a logical, intuitive manner. For example, we discuss the topics of hypothesis testing and inference in Chapters 4, 5, and 6.

Research can be viewed as the search for relationships between and among key variables of interest. Just as in the first edition, we show how statistics can help determine whether relationships exist, and if so, the strengths of those relationships. In Chapter 1 the nature of a relationship is introduced and throughout the rest of the book, the importance of relationships is stressed. The concept of crosstabulation is discussed earlier than it is in most introductory statistics books (Chapter 4), and statistical inference from samples to populations is emphasized from Chapter 1 on.

In this new edition we continue to avoid the usual format of presenting statistics for nominal, ordinal, interval, and ratio levels of measurement. Our focus is on whether a variable is continuous or discrete. We believe our approach is consistent with current practice as exemplified in the major professional journals in social and behavioral sciences.

Although the second edition retains the philosophy and approach of the first edition, it differs in some important ways. We maintain the hands-on approach to statistics through the use of actual data. The second edition is constructed, with few exceptions, around three data sets; the first edition had two. Many examples and problems in the second edition were designed with the General Social Survey (GSS), but data are from more recent studies than data in the first edition. For small sample analyses, a new data set includes variables on 24 nations. We have kept, however, examples and problems with the 63-cities data set introduced in the first edition. The latter two data sets (24 nations and 63 cities) are included as appendices, and are available from the publisher on diskettes. These two data sets can be used with problems that appear at the end of all chapters (except for Chapter 1) and/or to custom design problems. GSS data are not included as an appendix, but this data set has become widely available in recent years. If the data set is not available at your college or university, consult our *Instructor's Manual* for information on how to obtain it.

One of the strengths of our book is that the problems at the end of the chapter are varied enough to allow instructors to require no more than a handheld calculator as one option. Some of the problems require no more than a PC, and others require a mainframe. While the problems designed to run on a PC are available as a SPSS^{X1} system file, there is no reason why they could not be converted to another statistical package for PCs of the instructor's choice since the number of cases is small.

Other important changes and new features enhance the second edition. First, we have extensively rewritten the materials on hypothesis testing and inference in Chapters 4, 5 and 6. Second, we introduce the notions of odds and odds ratios in Chapter 9 and continue to illustrate their importance for interpreting crosstabulated data in Chapter 10. Third, we have revised and expanded the materials on exploratory data analysis in Chapters 3 and 6. These changes make the book even more useful for instructors who wish to expose students to modern statistical applications in the social sciences at the introductory level.

This text was written primarily for an undergraduate audience, but its orientation and introduction to more advanced topics (e.g., multivariate regression and path analysis) can lead directly into graduate courses. Graduate students who have not had undergraduate training in statistics should find this book a particularly useful introduction.

¹SPSS^X is a trademark of SPSS Inc. of Chicago, IL for its proprietary computer software. No materials describing such software may be produced or distributed without the written permission of SPSS Inc.

We are indebted to A. Hald for reprinting the Area Under the Normal Curve and E. S. Pearson and H. O. Hartley for the reprint of the *F*-Distribution Table. We are also grateful to the literary executor of the late Sir Ronald A. Fisher, F.R.S., to Dr. Frank Yates, F.R.S., and to the Longman Group Ltd. of London, for permission to reprint Tables III and IV from their book *Statistical Tables for Biological, Agricultural and Medical Research* (6th edition, 1974).

We acknowledge permission from SPSS Inc. to use their software package SPSS^X throughout the book to illustrate the analysis of survey data with the computer. We also acknowledge the cooperation of the National Opinion Research Center and the Roper Opinion Research Center for making available the General Social Surveys, and the Inter-University Consortium for Political and Social Research for providing the 63-cities data, both of which are used extensively in the book.

Any book with technical materials will have some errors. In spite of our attempts to prevent them, we are certain some will be found. We ask instructors and students to notify us or the publisher if you find errors.

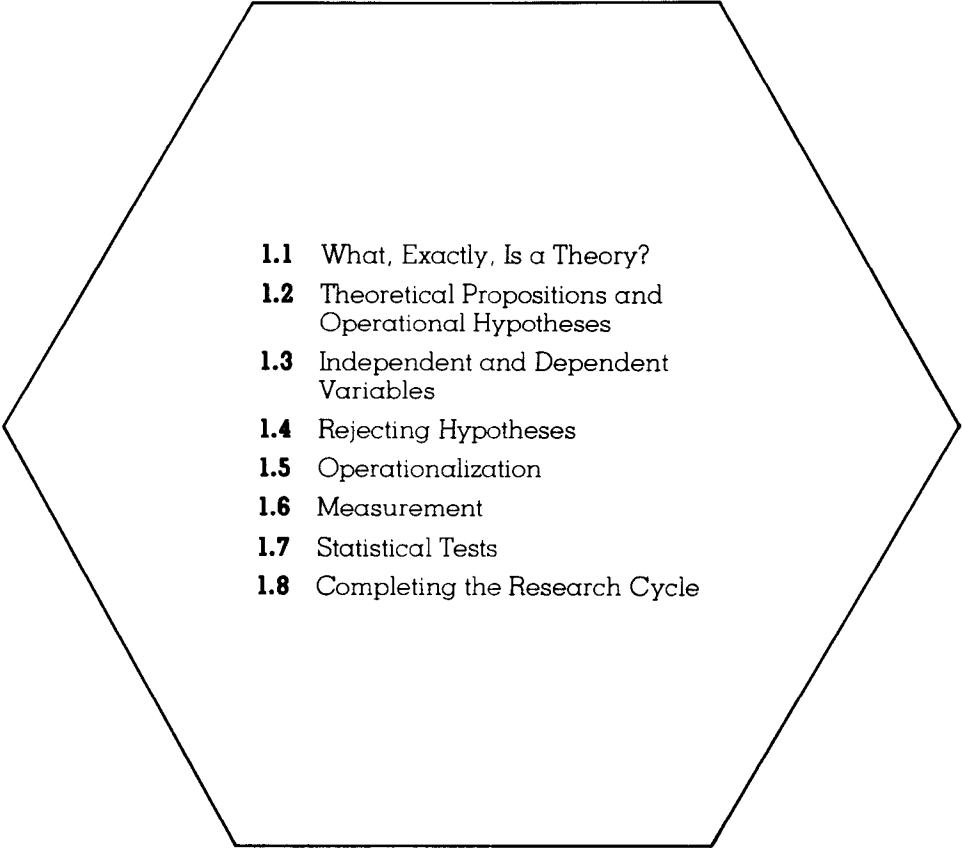
Many people assisted us in this revision. We especially thank Maureen Hallinan, Neil Henry, Herbert Smith, and Elton Jackson for their extensive and useful criticisms of certain sections of the original edition. We thank Frank Burleigh and Marcella DePeters for their invaluable computer assistance. We are grateful for the fine editorial help provided by John B. Goetz and his staff at Design & Production Services Co., Chicago. We thank our publisher F. Edward Peacock for the continual flow of encouragement. The first author is particularly thankful for support, in terms of time and typing, to the Zentrum für Umfragen, Methoden und Analysen (ZUMA) in Mannheim, West Germany and to the Center for Advanced Study in the Behavioral Sciences (CASBS) in Palo Alto, California. He was a guest professor at ZUMA during the summer of 1986 and a fellow at the center during the 1986-87 academic year. Special thanks for typing various parts of the revised manuscript are due Frau Dagmar Haas at ZUMA and to Mss. Deanna Knickerboker and Sharon Ray at Center for Advanced Study in the Behavioral Sciences. We also thank Jane Hilberry for proof-reading.

Finally, we thank our wives and children for their support.

George W. Bohrnstedt
Bloomington, Indiana

David Knoke
Minneapolis, Minnesota

Statistics for Social Data Analysis

- 
- 1.1** What, Exactly, Is a Theory?
 - 1.2** Theoretical Propositions and Operational Hypotheses
 - 1.3** Independent and Dependent Variables
 - 1.4** Rejecting Hypotheses
 - 1.5** Operationalization
 - 1.6** Measurement
 - 1.7** Statistical Tests
 - 1.8** Completing the Research Cycle

Abbreviated Contents

Preface	xv
1. The Social Research Process	1
2. Frequency Distributions	27
3. Describing Frequency Distributions	65
4. Crosstabulation	101
5. Statistical Inference and Hypothesis Testing	135
6. Testing for the Difference between Two Means	187
7. Testing for the Difference between Several Means	219
8. Estimating Relations between Two Continuous Variables; Bivariate Regression and Correlation	253
9. Measuring Association with Discrete Variables	305
10. The Logic of Multivariate Contingency Analysis	349
11. Multiple Regression Analysis	381
12. Causal Models and Path Analysis	431
Appendices	467
Glossary of Terms	489
List of Mathematical and Statistical Symbols	503
Answers to Problems	511
Index	547

Contents

PREFACE	xv
1. THE SOCIAL RESEARCH PROCESS	1
1.1 What, Exactly, Is a Theory?	3
1.2 Theoretical Propositions and Operational Hypotheses	6
1.3 Independent and Dependent Variables	8
1.4 Rejecting Hypotheses	10
1.5 Operationalization	12
1.6 Measurement	14
1.7 Statistical Tests	17
1.7.1 Descriptive statistics	17
1.7.2 Measures of association	17
1.7.3 Inference	19
1.8 Completing the Research Cycle	20
Review of Key Concepts	21
Problems	22
2. FREQUENCY DISTRIBUTIONS	27
2.1 Constructing a Frequency Distribution	27
2.1.1 Percentage frequency distributions	28
2.2 Frequency Distributions for Discrete Measures	32
2.2.1 Techniques for displaying data: Tables and bar charts	35
2.3 Frequency Distributions for Orderable Discrete Variables	39
2.3.1 Techniques for displaying data: Histograms and polygons	40

2.4	Frequency Distributions for Continuous Measures	43
2.4.1	Grouping data in measurement classes	44
2.4.2	Rounding	45
2.4.3	True limits and midpoints	46
2.5	Cumulative Distributions	49
2.6	Percentiles	51
2.7	Quantiles	54
2.8	Grouping Error	55
	Review of Key Concepts	56
	Problems	57
3.	DESCRIBING FREQUENCY DISTRIBUTIONS	65
3.1	Measures of Central Tendency	66
3.1.1	Mode	66
3.1.2	Median	67
3.1.3	Mean	70
3.2	Measures of Variation	75
3.2.1	Index of diversity	75
3.2.2	Index of qualitative variation	76
3.2.3	Range	78
3.2.4	Average deviation	80
3.2.5	Variance and standard deviation	81
3.3	Z Scores	83
3.4	Exploratory Data Analysis Methods for Displaying Continuous Data	85
3.5	The Coefficient of Relative Variation	90
	Review of Key Concepts	91
	Problems	92
4.	CROSSTABULATION	101
4.1	Bivariate Crosstabulation	101
4.1.1	An example: Political choices and social characteristics	104
4.2	Population Inference from Samples	109
4.3	Probability and Null Hypotheses	111
4.3.1	Type I and Type II errors	113
4.4	Chi Square: A Significance Test	114
4.5	Sampling Distributions for Chi Square	118
4.5.1	Another example using chi square	123
4.5.2	Chi square as a goodness-of-fit test	124
4.6	Two Meanings of Statistical Inference	126
	Review of Key Concepts	127
	Problems	128

5. STATISTICAL INFERENCE AND HYPOTHESIS TESTING	135
5.1 Probability Distributions	135
5.1.1 Discrete probability distributions	136
5.1.2 Continuous probability distributions	138
5.2 Describing Discrete Probability Distributions	139
5.2.1 The expected value and mean of a probability distribution	139
5.2.2 The variance of a probability distribution	140
5.3 Chebycheff's Inequality	141
5.4 Normal Distributions	144
5.4.1 The alpha area	147
5.5 The Central Limit Theorem	149
5.5.1 An example: Occupational prestige	151
5.6 Sample Point Estimates and Confidence Intervals	154
5.7 Constructing Confidence Intervals around Estimators When the Standard Error Is Unknown: The t Distribution	157
5.7.1 An example using the t distribution to construct a confidence interval	161
5.8 Desirable Properties of Estimators	162
5.8.1 Lack of bias	163
5.8.2 Efficiency	164
5.8.3 Consistency	164
5.8.4 The logic of inferential statistics summarized	165
5.9 Hypothesis Testing	165
5.9.1 Statistical hypotheses	165
5.9.2 Evaluating statistical hypotheses using sample data	168
5.9.3 An example of a hypothesis test about a single mean with exact hypotheses	168
5.9.4 An example of hypothesis testing about a single mean with inexact hypotheses	173
5.10 Two-Tailed Hypothesis Tests about a Single Mean	175
5.10.1 One-tailed vs. two-tailed hypothesis tests	177
Review of Key Concepts	179
Problems	179
6. TESTING FOR THE DIFFERENCE BETWEEN TWO MEANS	187
6.1 Testing for the Difference between Two Means When the Standard Error Is Known	187
6.1.1 Stating the operational hypotheses	188
6.1.2 Test procedures	190
6.1.3 Testing the hypotheses	196
6.2 Hypothesis Testing with Proportions	198
6.3 An Example of a Two-Tailed Hypothesis Test with Proportions	200

6.4	Testing for the Difference between Two Means When the Standard Error Is Unknown: The t Test	201
6.5	Comparing Two Distributions with Stem-and-Leaf Diagrams and Boxplots	206
6.6	Confidence Intervals and Point Estimates for Mean Differences and Proportions	208
6.7	Reporting p Values	210
6.8	Comparing Means from the Same Population Across Time	211
	Review of Key Concepts	211
	Problems	212
7.	TESTING FOR THE DIFFERENCE BETWEEN SEVERAL MEANS	219
7.1	The Logic of Analysis of Variance: An Example	219
7.2	Effects of Variables	221
7.3	The ANOVA Model	222
7.4	Sums of Squares	222
	7.4.1 Sums of squares in the problem-solving example	225
7.5	Mean Squares	229
7.6	The F Distribution	232
7.7	Reporting an Analysis of Variance	233
7.8	The relationship of t to F	234
7.9	Determining the Strength of a Relationship: Eta Squared	234
7.10	Testing for Differences between Individual Treatment Means	236
	7.10.1 Multiple means comparison using contrasts	237
	7.10.2 Mean comparison in the problem-solving example	239
7.11	The Use of ANOVA in Nonexperimental Research	240
	7.11.1 An example: Crime rate and population loss	241
	Review of Key Concepts	245
	Problems	245
8.	ESTIMATING RELATIONS BETWEEN TWO CONTINUOUS VARIABLES: BIVARIATE REGRESSION AND CORRELATION	253
8.1	An Example Using Regression and Correlation Techniques: Municipal Scope	253
	8.1.1 Descriptive statistics for municipal scope	255
8.2	Scatterplots and Regression Lines	256
8.3	Linear Regression Equations	259
	8.3.1 Linear regression applied to municipal scope	262
8.4	Measures of Association: The Coefficient of Determination and the Correlation Coefficient	266

8.4.1	The coefficient of determination	269
8.4.2	The correlation coefficient	271
8.4.3	Correlating Z scores	272
8.4.4	The relation of regression and correlation coefficients	273
8.5	Standardized Regression, or Beta, Coefficients	274
8.5.1	Regression toward the mean	275
8.6	Significance Tests for Regression and Correlation	277
8.6.1	Testing the significance of the coefficient of determination	278
8.6.2	Testing the significance of b and a	280
8.6.3	The relationship between F and t^2	283
8.6.4	Confidence intervals	284
8.6.5	Testing the significance of the correlation coefficient	284
8.7	A Problem with an Outlier: Testing the Second Hypothesis on Municipal Scope	286
8.8	Nonlinear Regression	289
8.8.1	Testing for curvilinearity	290
8.9	Special Cases of Inference Involving Correlations	294
8.9.1	Testing the difference between two means in nonindependent populations	294
8.9.2	Testing the difference between two correlations in independent populations	296
	Review of Key Concepts	297
	Problems	298
9.	MEASURING ASSOCIATION WITH DISCRETE VARIABLES	305
9.1	Measures of Association for Nonorderable Discrete Variables	305
9.1.1	An example: Religion and school prayer attitudes	306
9.1.2	Lambda	307
9.2	Measures of Association for Orderable Discrete Variables	309
9.2.1	An example: ERA support and sex role attitude	311
9.2.2	Gamma	313
9.2.3	Tau b	318
9.2.4	Tau c	322
9.2.5	Somers's d'_{yx}	323
9.2.6	Comparing orderable measures of association	325
9.3	The Association of Ranked Data: Spearman's Rho	326
9.3.1	An example: Female labor force participation and suicide	327
9.4	The 2×2 Table	328
9.4.1	An example: Religious intensity and prayer	329
9.4.2	Yule's Q	330
9.4.3	Phi	333
9.4.4	Odds and the odds ratio	334

9.4.5	Relation to chi-square-based measures	337
	Review of Key Concepts	339
	Problems	339
10.	THE LOGIC OF MULTIVARIATE CONTINGENCY ANALYSIS	349
10.1	Controlling Additional Variables	350
10.1.1	Spuriousness	352
10.1.2	Explanation	353
10.1.3	Multiple causes	355
10.2	Controlling for a Third Variable in 2×2 Tables	355
10.2.1	A hypothetical example: Family religiosity and teenagers' sex activity	356
10.2.2	No effect of third variable	357
10.2.3	Partial effect of third variable	359
10.2.4	Complete explanation by third variable	360
10.2.5	Interaction effect of third variable	362
10.2.6	Summary of conditional effects	363
10.3	The Partial Correlation Coefficient	366
10.3.1	An example: Relationships among three variables	368
10.3.2	Testing the partial correlation for significance	370
10.4	Multivariate Contingency Analysis in Larger Tables	371
	Review of Key Concepts	371
	Problems	372
11.	MULTIPLE REGRESSION ANALYSIS	381
11.1	An Example: Explaining Sexual Permissiveness	382
11.2	The Three-Variable Regression Model	386
11.2.1	Interpretation of b_1 and b_2	389
11.2.2	Standardized regression coefficients (beta weights)	390
11.2.3	The coefficient of determination in the three-variable case	392
11.2.4	Testing the significance of the coefficient of determination with two independent variables	396
11.2.5	Testing b_1 and b_2 for significance	398
11.2.6	Confidence intervals for b_1 and b_2	402
11.2.7	Partial correlation in the three-variable case	402
11.3	Multiple Regression with Several Independent Variables	403
11.3.1	Testing the coefficient of determination for several independent variables	405
11.3.2	Testing regression coefficients for several independent variables	405
11.3.3	An example: Examining the effects of gender on sexual permissiveness	407

11.4	Dummy Variable Regression Analysis	409
11.4.1	Testing for interaction effects	413
11.4.2	An example with three dummy variables and a continuous variable	416
11.4.3	Analysis of variance with dummy variable regression	419
	Review of Key Concepts	422
	Problems	423
12.	CAUSAL MODELS AND PATH ANALYSIS	431
12.1	Causal Assumptions	431
12.1.1	Covariation	433
12.1.2	Time order	434
12.1.3	Nonspuriousness	434
12.2	Causal Diagrams	435
12.3	Path Analysis	439
12.3.1	An example: The drinking behavior model	440
12.3.2	Structural equations	441
12.3.3	Estimating path coefficients	442
12.3.4	Decomposing implied correlations into causal parameters	443
12.3.5	Decomposing implied correlations by tracing paths	448
12.3.6	An example using path analysis: Estimating the drinking behavior model	451
12.3.7	An example of a chain path model: School busing attitude	455
	Review of Key Concepts	459
	Problems	459
	APPENDICES	467
A.	The Use of Summations	467
B.	Critical Values of Chi Square (table)	475
C.	Area under the Normal Curve (table)	477
D.	Student's <i>t</i> Distribution (table)	478
E.	<i>F</i> Distribution (table)	479
F.	Fisher's <i>r</i> -to- <i>Z</i> Transformation (table)	483
G.	SPSS ^X File for the 24-Nations Data Set	484
H.	SPSS ^X File for the 63-Cities Data Set	486
	Glossary of Terms	489
	List of Mathematical and Statistical Symbols	503
	Answers to Problems	511
	Index	547

The Social Research Process

The social research process often begins with questions about the world: What kind of people vote for Democrats? Do lower-income people have more children than middle-income people do? What incomes can be earned in various occupations? Why do Protestants have higher suicide rates than Catholics or Jews? Do blacks achieve less education than whites because they have lower IQs or because of other differences? Each of these questions is phrased in terms of the *relationship* between two or more observable characteristics of people or groups, such as income and occupation. We will have much to say about various relationships throughout this text, since they represent the *central concept* in social science.

If social research is to answer questions like these, it naturally must ask where the questions come from. Personal experience, hunch, intuition, friends' suggestions, or a variety of stimuli such as newspaper and TV accounts clearly provide relevant clues. But social scientists also have an inheritance from the past from which ideas for social research can be drawn. This inheritance is a steady accumulation of social knowledge which has been painstakingly assembled by several generations of psychologists, political scientists, sociologists, anthropologists, and economists, as well as applied researchers in education, business, and law. Together, their writings contain many theories and empirical findings about social phenomena. A student's training in the social sciences begins with an introduction to the theoretical ideas of the great masters. The

thoughts of Aristotle, Emile Durkheim, Karl Marx, Max Weber, John M. Keynes, Alfred Marshall, Charles Merriam, Arthur Bentley, Bronislaw Malinowski, Sir James Frazer, and other founding fathers of social science are a source of continuing inspiration for researchers. The importance of more contemporary if less renowned social scientists in providing ideas to be tested also must be recognized.

At its best, social science is firmly grounded in the real world. It seeks to explain social behavior, but it is distinct from related fields like social philosophy and theology which deal with idealizations that have few empirical referents. The more comprehensive social theories present distinctive views of reality which are sometimes labeled *paradigms*, or examples or patterns.¹ Another term which is used frequently is *model*. Paradigms are usually seen as broader and more encompassing than models, but both are abstractions and simplifications of the complex real world. Partitions of the seamless totality are essential if a theory is to be of any use in guiding social inquiry. No theory can seek to account meaningfully for all the significant aspects of social life. Instead, selective attention must be given to a few aspects of the phenomena to be explained. As a result, theory deals with only a part of the world and takes the rest for granted or, at least, assumes it to be sufficiently unobtrusive so it can be safely ignored while concentrating on the topic of interest.

Examples of such theoretical abstractions abound. One of the most popular in psychology is the stimulus-response, or S-R, paradigm. In B. F. Skinner's operant psychology theory, behavior is seen as purely reactive to external stimuli.² It posits that all behavior is a response to external stimuli, and there is no need to consider the mediating mental processes. In contrast, psychoanalytic explanations of social behavior rely on extensive, elaborate mechanisms of internal processes like Freud's trinity of id, ego, and superego. Many of these processes are unconscious, and their existence is inferred by observing patients' behaviors in dreams, slips of the tongue, and neurotic compulsions. Though the S-R and the psychoanalytic theories of behavior have markedly divergent elements, both concentrate their explanations on a few key aspects of reality and leave other features aside.

1. Thomas Kuhn, *The Structure of Scientific Revolutions* (Chicago: University of Chicago Press, 1962).

2. B. F. Skinner, *Science and Human Behavior* (New York: Macmillan Co., 1953).