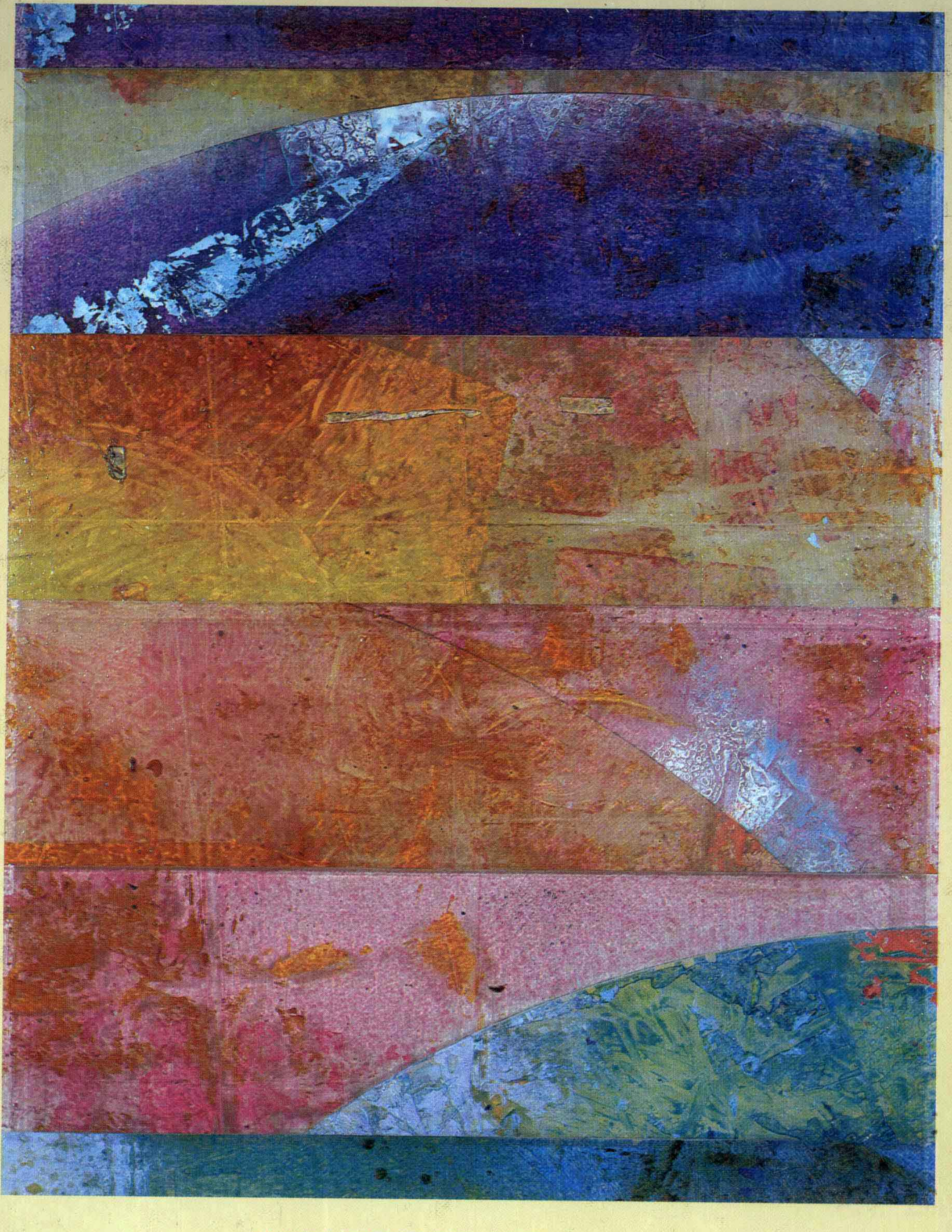
SECOND EDITION

STATISTICAL METHODS FOR THE SOCIAL SCIENCES



ALAN AGRESTI AND BARBARA FINLAY

SECOND EDITION

Statistical Methods for the Social Sciences

Alan Agresti
Department of Statistics
University of Florida

Barbara Finlay
Department of Sociology
Texas A & M University

Dellen Publishing Company
San Francisco

Collier Macmillan Publishers
London

divisions of Macmillan, Inc.

On the cover: The cover, *Floating World*, was executed by San Francisco artist Gregg Renfrow in 1984. The work is on polymer plexiglass and measures 67¼" × 53".

Chapter opening pages: The chapter opening pages present a series of drawings by Sam Richardson. The drawings refer to the wedge shape as a vehicle for surface textures and sculptural elements such as pries and ropes.

Permissions: Dellen Publishing Company
400 Pacific Avenue
San Francisco, California 94133

Orders: Dellen Publishing Company
℅ Macmillan Publishing Company
Front and Brown Streets
Riverside, New Jersey 08075

Collier Macmillan Canada, Inc.

Library of Congress Cataloging-in-Publication Data

Agresti, Alan.
Statistical methods for the social sciences.

Includes bibliographies and index.
1. Statistics. 2. Social sciences—Statistical
methods. I. Finlay, Barbara. II. Title.
QA276.12.A34 1986 519.5'024'301 85-16031

Printing 4 5 6 7 8 9 Year 7 8 9 0 1

ISBN 0-02-301120-3

PREFACE

This book is designed as a text for introductory courses in statistics at the undergraduate or beginning graduate level. It is intended especially for students in social science disciplines. The social science orientation determined both the choice of techniques to be discussed and the selection of examples and problems. The many examples and problems included in the book are based on research studies in sociology, political science, education, geography, anthropology, psychology, history, journalism, and speech. A small amount of elementary algebra is the only mathematical prerequisite for understanding the material presented and working the problems.

The book is suited for use in either a single-term or a two-term sequence. Chapters 1–9 form a basis for a one-term introductory course. If the instructor has only one term and wishes to go further than Chapter 9 or wishes to cover some material in greater depth, some sections could easily be omitted without disturbing the continuity of presentation. We suggest, for example, the following sections for possible omission: 6.5, 6.7, 7.3, 7.4, 8.3–8.5, 9.4, and 9.5. Also, Chapters 7–9 and Sections 13.1–13.2 are self-contained, and the instructor could move directly into any of these after covering the fundamentals in Chapters 1–6. Four possible paths for a one-term course are as follows:

1. Chapters 1–9 (possibly omitting sections noted above): Standard cross-section of methods, including basic descriptive and inferential statistics, two-sample procedures, categorical data, and linear regression
2. Chapters 1–6, 9–12: Emphasis on bivariate and multiple regression
3. Chapters 1–7, 13: Emphasis on group comparisons and analysis of variance
4. Chapters 1–8, 10, 15: Emphasis on nonparametrics and categorical data

Chapters 10–16 are primarily concerned with regression modeling, and they could be used as the basis of a second course. These chapters can also be naturally linked to an introduction to a computer package such as SPSS^X, the SAS System, BMDP, or OSIRIS, using data sets and printouts provided in the text.

This edition contains some changes and additions in content, compared to the first edition. The technical level has been lowered somewhat in the first nine chapters, to make the book more easily accessible to undergraduate students. The chapters on categorical data have been reorganized, with a separate chapter given to the important concepts of multivariate relationships and the use of statistical control. A new chapter (15) has been added on advanced methods for categorical data, with emphasis on logistic regression and loglinear models—methods that have recently become very popular in social science research. Many homework problems have been added to each chapter, with particular emphasis on elementary problems.

The book is organized first by type of procedure (e.g., comparing two groups in Chapter 7, bivariate association in Chapters 8–9). For each type of procedure, the sections and chapters are also classified according to the levels of measurement of the variables analyzed. Throughout the book, we have attempted to provide understanding through intuition and example, rather than through mathematical derivations. We believe it is of primary importance that the student develop an understanding of the purposes of the various procedures and their interpretations. We feel that this is best accomplished by exposure to realistic but simple examples and to numerous homework problems. Thus, many problems appear at the end of each chapter.

In view of the increasing reliance of users of statistics on sophisticated pocket calculators, home computers, and computer software packages, we have omitted many traditional shortcut computational formulas and formulas for coded and grouped data. In later chapters, computationally complex procedures such as multiple regression and analysis of variance are taught by explaining how to interpret output from a computer package program. Since it is unlikely that students will ever do such analyses by hand, the matrix-based formulas for obtaining parameter estimates, sums of squares, and standard errors have been omitted from this book.

Unlike many other statistics textbooks, we have integrated the presentation of descriptive and inferential methods from a very early point. We believe that artificially dividing these topics into separate presentations leads to confusion for many students. Hence, we present the foundations of inference early in the text (Chapters 5 and 6). In later chapters, we present new descriptive measures together with the corresponding inferential procedures so that the student becomes familiar with a unified process of describing and making inferences for a wide variety of problems.

Some material appears in the text in end-of-chapter Notes or in paragraphs or subsections marked by asterisks. This material is optional, being of lesser importance to beginning students, although it often provides a broader coverage of topics for the interested student. Likewise, nearly every chapter contains a number of optional problems identified by asterisks. These problems either refer to the optional text material, introduce additional material related to the chapter, or pose questions of a more difficult or theoretical nature than the unstarred ones.

Finally, we have not attempted to provide a catalog of every technique for every situation. This text is meant to be primarily a teaching tool, not an encyclopedic cookbook. We do feel that we have covered the most important procedures for social science research, however, and we have included some methods that are not usually described in introductory statistics books, which are useful to social scientists, for example:

1. Procedures that are more powerful than chi-square when categories in a cross-classification table are ordered
2. Controlling for variables, and testing for causal relationships
3. Models for nonlinear relationships and for statistical interaction
4. Analysis of variance and covariance using dummy variables
5. Loglinear and logit models for categorical data
6. Introductions to path analysis, factor analysis, and LISREL

We hope that the combination of our respective fields of expertise has led to a book that is statistically sound as well as relevant to social science problems. We believe that the student who works through this book successfully will have acquired a good foundation in statistical methodology.

Acknowledgments

We are grateful to the Literary Executor of the late Sir Ronald A. Fisher, F.R.S., to Dr. Frank Yates, F.R.S., and to Longman Group Ltd., London, for permission to reprint Tables IV and V from their book *Statistical Tables for Biological, Agricultural and Medical Research*, 6th ed., 1974.

Many people who read the first edition or earlier versions of this manuscript made suggestions that led to improvements. We would like to give special thanks to Susan Reiland, Maureen Hallinan, Shirley Scritchfield, Sonja Wright, and Douglas Zahn for their help with the first edition, and Jeff Witmer, E. Jacquelin Dietz, Dorothy K. Davidson, and Mary Sue Younger for their comments that helped in the preparation of the second edition. Other individuals who provided helpful advice include John Henretta, Michael Radelet, and Andrew Rosalsky.

Our gratitude extends to many other individuals who helped this book to reach fruition. Our publisher, Don Dellen, was a source of much encouragement and support, and we especially appreciate the free rein he gave us in decisions about the book content and style. Excellent editorial services were provided by Susan Reiland. Brenda Wise, Ruth Byrne, and Edna Larrick worked patiently and conscientiously in typing the manuscript. Finally, we would like to give special thanks to Gerald Leslie and William Mendenhall, who, as former chairmen of the Departments of Sociology and Statistics, respectively, are responsible for the cooperation between these departments at the University of Florida which led to the courses from which this textbook was developed.

CONTENTS

Preface xiii

PART I INTRODUCTION TO UNIVARIATE STATISTICAL ANALYSIS 1

Chapter 1 Introduction 3

1.1 Introduction to Statistical Methodology 4
1.2 Description and Inference 6
1.3 The Role of Computers in Statistics 9
Summary 10
Problems 10
Bibliography 11

Chapter 2 The Measurement Process 13

2.1 Variables and Their Measurement 14
2.2 Types of Sampling Procedures 19
2.3 Sampling and Statistical Inference 25
Summary 26
Problems 26
Bibliography 29

Chapter 3 Descriptive Techniques 31

3.1 Tabular and Graphical Description 32
3.2 Measures of Central Tendency 40
3.3 Measures of Dispersion 49
3.4 Sample Statistics and Population Parameters 54
Summary 55
Notes 56
Problems 58
Bibliography 63

Chapter 4 Probability Distributions 65

4.1 Probability Distributions for Discrete and Continuous Variables 67
4.2 The Normal Probability Distribution 70
4.3 Sampling Distributions 77
4.4 The Central Limit Theorem 79
Summary 87
Notes 87
Problems 88
Bibliography 92

Chapter 5 Statistical Inference: Estimation 93

5.1 Point Estimation 94
5.2 Confidence Interval for a Mean—Large Samples 97
5.3 Confidence Interval for a Proportion—Large Samples 101
5.4 Choice of Sample Size 105
Summary 111
Notes 111
Problems 113
Bibliography 119

Chapter 6 Statistical Inference: Testing Hypotheses 121

6.1 Elements of a Statistical Test 123
6.2 Test for a Mean—Large Samples 125
6.3 Test for a Proportion—Large Samples 132
6.4 Small-Sample Inference for a Mean—The *t* Distribution 136
6.5 Small-Sample Inference for a Proportion—The Binomial Distribution 142
6.6 Making Decisions in Statistical Tests 146
6.7 Some Comments on Tests of Hypotheses 151
Summary 153
Notes 155
Problems 155
Bibliography 162

PART II INTRODUCTION TO BIVARIATE STATISTICAL ANALYSIS 163

Chapter 7 Comparison of Two Groups 165

7.1 Nominal Scales: Difference of Proportions 167
7.2 Interval Scales: Difference of Means 171
7.3 Ordinal Scales: Wilcoxon Test 177
7.4 Comments Concerning Nonparametric Statistics 185
Summary 186
Notes 187
Problems 189
Bibliography 196

Chapter 8 Measuring Association for Categorical Variables 197

8.1 Cross-Classification Tables 198
8.2 Chi-Square Test of Independence 202
8.3 Measuring Association: 2 × 2 Tables 209
8.4 Measuring Association: $r \times c$ Tables 213
8.5 Measuring Association for Ordinal Variables 219
8.6 Testing for Association for Ordinal Variables 227
Summary 230
Notes 231
Problems 232
Bibliography 241

Chapter 9 Linear Regression and Correlation 243

9.1 The Linear Regression Model 245
9.2 Method of Least Squares and Prediction Equation 251
9.3 Measuring Linear Association—The Correlation Coefficient and Its Square 260
9.4 Inferences for the Slope and Correlation 268
9.5 Some Practical Considerations 274
Summary 278
Notes 279
Problems 280
Bibliography 288

PART III INTRODUCTION TO MULTIVARIATE STATISTICAL ANALYSIS 289

Chapter 10 Introduction to Multivariate Relationships 291

10.1 Criteria for Causality 292
10.2 Controlling for Other Variables 294
10.3 Other Types of Multivariate Relationships 301
10.4 Inferential Issues in Statistical Control 305
Summary 307
Problems 308
Bibliography 313

Chapter 11 Multiple Regression and Correlation Procedures 315

11.1 The Multiple Regression Model 317
11.2 Example Based on Computer Analysis 321
11.3 Multiple and Partial Correlations 325
11.4 Inferences for Partial Coefficients and Correlations 332
11.5 Standard Regression Coefficients 339
Summary 343
Notes 345
Problems 345
Bibliography 356

Chapter 12 Model Building with Multiple Regression 357

12.1 Modeling a Nonlinear Response 358
12.2 Modeling Interaction Among Interval Variables 368
12.3 Comparison of Regression Models 372
12.4 Evaluating the Model 380
Summary 386
Problems 387
Bibliography 395

Chapter 13 Comparison of Several Groups 397

13.1 Interval Scales: One-Way Analysis of Variance 398
13.2 Interval Scales: Multiple Comparison of Means 406
13.3 ANOVA and Regression 411

13.4 Two-Way Analysis of Variance 414
13.5 Two-Way ANOVA and Regression 418
13.6 Nominal Scales: Chi-Square as a Test of Homogeneity 424
13.7 Ordinal Scales: Kruskal–Wallis Test 425
Summary 429
Notes 429
Problems 431
Bibliography 439

Chapter 14 **Comparison of Several Groups While Controlling for a Covariate** **441**

14.1 Comparing Adjusted Means and Comparing Regression Lines 443
14.2 Analysis of Covariance Model Without Interaction Terms 447
14.3 Analysis of Covariance Model with Interaction Terms 452
14.4 Tests for Interaction and for Equivalence of Regression Lines 455
14.5 Comparing Adjusted Means 460
14.6 Comparing Regression Lines 467
Summary 471
Notes 472
Problems 472
Bibliography 480

Chapter 15 **Models for Categorical Variables** **481**

15.1 Logistic Regression 482
15.2 Logit Models for Categorical Variables 486
15.3 Loglinear Models for Categorical Variables 492
15.4 Odds Ratios 497
15.5 Modeling Ordinal Variables 502
Summary 504
Notes 504
Problems 505
Bibliography 508

Chapter 16 **An Introduction to Advanced Methodology** **509**

16.1 Path Analysis 510
16.2 Factor Analysis 510
16.3 LISREL 517
16.4 Markov Chains 520
Bibliography 523

Tables **525**

A *Normal Curve Areas* 527
B t Distribution 528
C χ^2 Distribution 529
D F Distribution 530
E Values of $T(r) = 1.151 \log \left(\frac{1+r}{1-r}\right)$ 533

Answers to Selected Problems 535

Index 549

PART I

INTRODUCTION TO UNIVARIATE STATISTICAL ANALYSIS

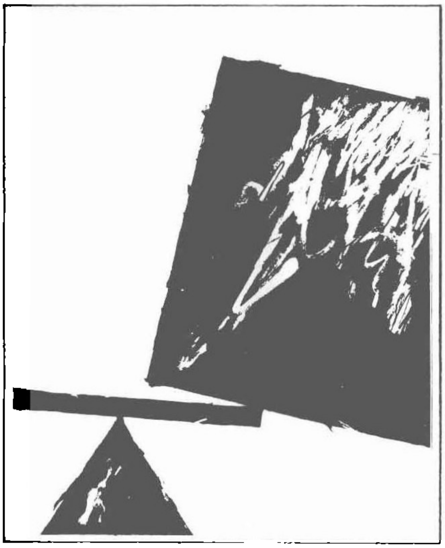
CHAPTER 1

Introduction

Contents

1.1 Introduction to Statistical Methodology
1.2 Description and Inference
1.3 The Role of Computers in Statistics

Summary
Problems
Bibliography



1.1 Introduction to Statistical Methodology

Within the past 20 to 30 years, almost all social science disciplines have seen a rapid increase in the use of statistics. The increase is evident in the changes in the content of articles published in major journals, in the styles of textbooks, and in the increasingly common requirement of academic departments that their majors take courses in statistics. A quick glance through recent issues of *American Political Science Review*, *American Sociological Review*, *Social Forces*, or other leading social science journals is enough to convince one of the central place of statistics in social science research. Job advertisements for social scientists commonly list a knowledge of statistics as an important work tool. Almost any student preparing for a career as a social scientist must become familiar with basic statistical methodology.

The study of statistics is also useful to those who do not expect to become professional social scientists. In fact, statistics is an important part of a general education for living in today's world. Almost daily we are confronted with advertising, news reporting, political campaigning, and other communications containing statistical arguments. The study of statistics helps us to understand and evaluate these arguments.

Science and Observation

Much of science involves collecting and organizing information about the world around us. Although different sciences have different subject matters and therefore different methods, all sciences engage in some form of information-gathering through observation. The social sciences use a wide variety of information-gathering techniques. Among these are questionnaire surveys, telephone surveys, content analysis of literary materials, planned experiments, and direct observation of behavior in natural settings. In addition, social scientists often utilize information already observed and recorded for other purposes, such as police records, census materials, and hospital files. The information gathered through such processes is collectively called ***data***. These data consist of measurements taken on the various characteristics being observed. The measurements are obtained by classifying observations according to specific rules. Thus, we can measure sex by classifying individuals according to gender; we can measure age by classifying people according to the number of years since birth; and we can measure the populations of cities by obtaining census counts for their numbers of residents.

What Is Statistics?

The general field of statistics involves methods for (a) designing and carrying out research studies; (b) describing collected data; and (c) making decisions, predictions, or inferences about phenomena represented by the data. In this book, we deal primarily with the latter two aspects. This is not to imply that the other aspects of the research process are unimportant. If a study is poorly designed or if the data are improperly collected or recorded, then the conclusions may be worthless or misleading, no matter how good a statistical analysis is performed. Methods for designing and carrying out research studies are covered in detail in textbooks on research methods (e.g., Bailey, 1982).

The descriptive aspect of statistics allows researchers to summarize large quantities of data using measures that are easily understood by an observer. It would always be possible, of course, simply to present a long list of measurements for each characteristic being observed. In a study of ages at first marriage, for example, we might just present the reader with a listing of the ages of all persons marrying for the first time within the past year in a particular county. This kind of detail, however, is not easy to assess—the reader simply gets bogged down in numbers. Instead of presenting *all* observations, we could use one of several statistical measures that would summarize the *typical* age at marriage in the collection of data. This would be much more meaningful to most people than the complete listing.

The other aspect of statistics that we will study is making decisions or inferences about characteristics by interpreting data patterns. This process often takes the form of noting whether an *expected* or *predicted* pattern of data values is actually found in the observations. We might expect, for example, that Catholics would be more likely to believe in life after death than would Unitarians. In order to decide whether this expectation is true, we might survey a number of Catholics and Unitarians, questioning them about their views concerning an afterlife. We could then use the statistical methodology described in this textbook to make comparisons between the two groups and to make a decision about whether the prediction is true for Catholics and Unitarians in general. In addition, we could estimate how great the difference is between the beliefs of the two groups.

Inherent in both the descriptive and decision-making aspects of statistics is the development of explanations for complex social processes. Much of social science research deals with sorting out relationships among such factors as political party preference, income, degree of education, religion, race, sex, and so forth. ***Statistical models***, mathematical representations of such relationships, provide a mechanism for doing so. We shall see that many relatively simple statistical procedures are based on the assumption of a particular model for a real-world situation.

1.2 Description and Inference

A statistical procedure is usually classified as ***descriptive*** or ***inferential***. In order to see the distinction between these purposes, it is necessary to understand the terms *population* and *sample*.

Population and Sample

The ***population*** is the total set of individual objects or persons of interest in a study.

A ***sample*** is a subset of the population that is actually observed.

Although research is directed toward learning about populations, it is often necessary to study only samples from those populations. For example, the Gallup and