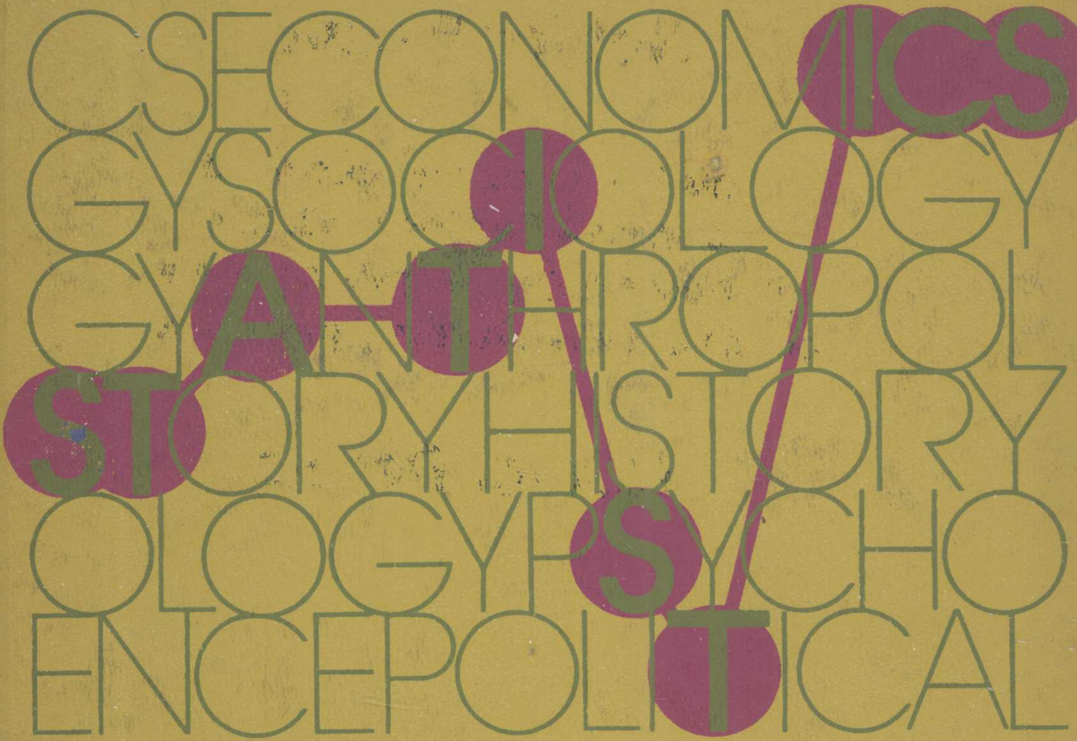

# Statistics: a Tool for the Social Sciences

Mendenhall | Ott | Larson

# STATISTICS:

William Mendenhall
UNIVERSITY OF FLORIDA

Lyman Ott
UNIVERSITY OF FLORIDA

Richard F. Larson
CALIFORNIA STATE UNIVERSITY, HAYWARD

# A
# Tool for the
# SOCIAL
# SCIENCES

DUXBURY PRESS
*North Scituate, Massachusetts*

*Duxbury Press*
*A Division of Wadsworth Publishing Company, Inc.*

L.C. Cat. Card No.: 73-84305
ISBN 0-87872-053-7
Printed in the United States of America

*Statistics: A Tool for the Social Sciences* by William Mendenhall, Lyman Ott, and Richard F. Larson was edited and prepared for composition by Service to Publishers, Inc. Interior design was provided by Mr. David Earle and the cover was designed by Mr. Oliver Kline. The type was set by J.W. Arrowsmith Ltd., and the book was printed and bound by Kingsport Press, Inc.

1 2 3 4 5 6 7 8 9 10—78 77 76 75 74

# PREFACE

This book is for students in a one-quarter or one-semester statistics course who wish to gain an appreciation of the role of statistics in the social sciences. We use the term "appreciation" deliberately: the intent is to give the reader an overview of the role that statistics plays in social science investigations. This should better equip him to read and interpret articles in social science journals and to utilize statistical methods in his own social science research.

Equal time is given to the two major statistical divisions: *descriptive statistics* and *inferential statistics*; in this, the book differs from many other social science statistics books currently on the market. Included in the text is the standard treatment of descriptive statistics—so essential for the presentation and interpretation of social science data—but also a thorough (but elementary) introduction to the concepts of statistical inference.

To fulfill our teaching objective, we have reduced the dependence of the material on mathematics, particularly probability theory. The mathematical background required for the course is therefore quite elementary in nature, requiring only an understanding of the use of mathematical formulas.

The salient features of this text are its organization, its continuity, and the strong attempt we have made to show the relevance of statistics to social science investigations. Chapter topics are split into two clearly identified groups. One deals with the description of social science data. The other is concerned with statistical inference based on sampling. Definitions, formulas, and important rules for data presentation are set apart from the body of the text in boxes. Terms and discussions of major importance are shown in color. These devices, it is hoped, will make the book very easy to use.

Continuity is maintained by relating each topic to one of the two objectives of statistics—description or inference—as well as to preceding chapters and sections. Chapter introductions and summaries also assist in establishing the continuity and serve to relate the statistical concepts to practical problems encountered in the social sciences. Further relevance to real-world problems is provided by worked examples in the body of the text, many of which have been extracted from current social science literature. And interspersed throughout the chapters and at the ends of chapters are an unusually large number of practical exercises.

The subject matter can be used in a variety of ways. A one-quarter course presently taught for social science majors at the University of Florida provides a balanced offering of topics concerned with both descriptive and inferential statistics. These include Chapters 1 through 6, selected portions of Chapters 7 and 8, and Chapters 9 through 11. A course heavily tilted toward descriptive statistics could include the same chapters but with reduced emphasis on Chapters 6 through 8.

The authors wish to acknowledge the helpful comments of the reviewers, most particularly Norman R. Kurtz, Brandeis University; Susan Schwarz, Dartmouth College; and Robert P. Snow, Arizona State University. Thanks are due also to A. Hald, The Biometrika Trustees, The Chemical Rubber Company, and F. J. Massey, Jr., for permission to reprint tables. We are also indebted to our many typists, including Cathy Kennedy, Ellen Evans, and Carol Storer. And last but not least, we acknowledge the patience, encouragement, and forbearance of our families.

# CONTENTS

Preface xi

1. WHAT IS STATISTICS? 2

1.1 *Prelude* 3
1.2 *Why Study Statistics?* 7
1.3 *The Interplay Between Statistics and the Social Sciences* 8
1.4 *A Note to the Student* 9

2. MEASUREMENT 12

2.1 *Introduction* 13
2.2 *Constants and Variables* 13
2.3 *Scales of Measurement for Qualitative and Quantitative Variables* 17
2.4 *Examples of Social Data* 21
2.5 *Comparisons of Variables Employing Ratios, Proportions, Percentages, and Rates* 24
2.6 *Measurement of a Variable over Time: Time Series* 31
2.7 *Are the Measurements Rounded?* 33
2.8 *Summary* 38

3. GRAPHICAL TECHNIQUES FOR DESCRIBING DATA FROM A SINGLE VARIABLE 42

3.1 *Introduction* 43
3.2 *Organizing Qualitative Data* 44
3.3 *Graphical Description for Qualitative Data: Pie Chart* 47
3.4 *Graphical Description for Qualitative Data: Bar Graph* 50
3.5 *Graphical Description for Qualitative Data: Statistical Map* 53
3.6 *Organizing Quantitative Data* 58
3.7 *Graphical Description for Quantitative Data: Histogram* 63
3.8 *Comments Concerning Histograms* 70
3.9 *Graphical Description for Quantitative Data: Frequency Polygon* 74
3.10 *Cumulative Relative Frequency Polygons* 77
3.11 *Graphical Description for Quantitative Data: Time Graph* 79
3.12 *Different Shapes for Frequency Distributions* 90
3.13 *Graphical Distortions* 91
3.14 *Summary* 92

4. NUMERICAL TECHNIQUES FOR DESCRIBING DATA FROM A SINGLE SAMPLE 96

4.1 *Introduction* 97
4.2 *A Measure of Central Tendency: The Mode* 98
4.3 *A Measure of Central Tendency: The Median* 104
4.4 *A Measure of Central Tendency: The Mean* 108
4.5 *Coding to Reduce Computational Labor for the Sample Mean* 113
4.6 *A Measure of Location: Percentiles* 117
4.7 *Data Variation* 120
4.8 *A Measure of Variability for Qualitative Data: The Index of Qualitative Variation* 121
4.9 *A Measure of Variability for Quantitative Data: The Range* 125
4.10 *A Measure of Variability for Quantitative Data: The Average Deviation* 128
4.11 *A Measure of Variability for Quantitative Data: The Variance* 133
4.12 *Coding to Reduce Computational Labor for the Sample Standard Deviation* 140
4.13 *A Check on the Calculation of a Standard Deviation* 142
4.14 *Summary* 145

5. PROBABILITY, INFERENCES, AND SAMPLING DISTRIBUTIONS IN THE SOCIAL SCIENCES 150

5.1 *How Probability Helps in Making Inferences* 151
5.2 *What Is Probability?* 152
5.3 *Finding the Probability of an Event: An Empirical Approach* 155
5.4 *Conditional Probability and Independence* 157
5.5 *Variables: Discrete and Continuous* 158
5.6 *Probability Distributions for Discrete Random Variables* 160
5.7 *A Useful Discrete Random Variable: The Binomial* 163
5.8 *Probability Distributions for Continuous Random Variables* 168
5.9 *A Useful Continuous Random Variable: The Normal Probability Distribution* 171
5.10 *The Normal Approximation to the Binomial Distribution* 173
5.11 *Tabulated Areas Under the Normal Probability Distribution* 177
5.12 *Parameters of Populations* 185
5.13 *Random Sampling* 186
5.14 *Sampling Distributions* 189
5.15 *Summary* 191

6. CONCEPTS OF ESTIMATION: THE ONE-SAMPLE CASE 196

6.1 *Introduction* 197
6.2 *Two Types of Estimators* 198
6.3 *Estimation of a Population Mean: A Point Estimator* 204
6.4 *Estimation of a Population Mean: An Interval Estimator* 208
6.5 *Estimation of the Binomial Parameter p: A Point Estimator* 215
6.6 *Estimation of the Binomial Parameter p: An Interval Estimator* 217
6.7 *Choosing the Sample Size to Estimate a Population Mean, $\mu$, or Proportion, p* 219
6.8 *Summary* 224

7. STATISTICAL TESTS OF HYPOTHESES: THE ONE-SAMPLE CASE 228

7.1 *Introduction: A Statistical Test of an Hypothesis* 229
7.2 *Elements of a Statistical Test* 231
7.3 *Testing an Hypothesis About a Binomial Parameter p* 235
7.4 *Testing an Hypothesis About a Population Mean $\mu$* 238
7.5 *How To Select the Research and Null Hypotheses* 241
7.6 *The z Statistic and the Choice of $\alpha$* 243
7.7 *A One-Sample Goodness-of-Fit Test: The Chi-Square Test* 247
7.8 *The Kolmogorov–Smirnov One-Sample Test: Another Goodness-of-Fit Test* 255
7.9 *Summary* 257

8. ESTIMATION AND STATISTICAL TESTS OF HYPOTHESES: TWO SAMPLES 262

8.1 *Introduction* 263
8.2 *A Comparison of Two Population Means* 264
8.3 *A Comparison of Two Binomial Parameters* 270
8.4 *Another Two-Sample Test: The Mann–Whitney U Test* 276
8.5 *The Paired-Difference Experiment: An Example of a Designed Experiment* 282
8.6 *A Simple Comparative Test: The Sign Test* 290
8.7 *A Small-Sample Test Concerning a Population Mean* 297
8.8 *A Small-Sample Test of a Difference in Means* 305
8.9 *Summary* 310

9. CROSS CLASSIFICATION OF BIVARIATE DATA 316

9.1 *Introduction* 317
9.2 *Describing Cross-Classification Data: Percentage Comparisons* 317
9.3 *A Test of Significance for Cross-Classification Data: The Chi-Square Test of Independence* 321
9.4 *A Test of Significance for Cross-Classification Data: Fisher's Exact Test* 333
9.5 *A Test of Significance for Cross-Classification Data: The McNemar Test* 337
9.6 *Summary* 339

10. MEASURES OF ASSOCIATION: NOMINAL AND ORDINAL DATA 344

10.1 *Introduction* 345
10.2 *A Nominal Measure of Association: The Phi Coefficient* 345
10.3 *A Nominal Measure of Association: The Contingency Coefficient* 349
10.4 *Reduction in Error for Measures of Association* 352
10.5 *A Proportional Reduction in Error Measure of Association for Nominal Scales: Lambda* 359
10.6 *A Proportional Reduction in Error Measure of Association for Ordinal Scales: Gamma* 362
10.7 *A Proportional Reduction in Error Measure of Association for Ordinal Scales: Spearman's Rho* 372
10.8 *A Measure of Association for Ordinal Data: Kendall's W* 376
10.9 *Summary* 382

11. REGRESSION AND CORRELATION 388

11.1 *Introduction* 389
11.2 *Scatter Diagrams and the Freehand Regression Line* 390
11.3 *Method of Least Squares* 396
11.4 *The Coefficient of Determination* $r^2$ 404
11.5 *Computing Formulas for* $r$ 409
11.6 *A Test of Significance Concerning* $r$ 414
11.7 *Summary* 417

12. DIFFERENCES AMONG MORE THAN TWO POPULATION MEANS 420

12.1 *Introduction* 421
12.2 *The Logic Behind an Analysis of Variance* 421

12.3 *A Test of an Hypothesis Concerning More Than Two Population Means: An Example of an Analysis of Variance* 423
12.4 *Kruskal–Wallis One-Way Analysis of Variance by Ranks* 433
12.5 *Summary* 437

13. SUMMARY 440

APPENDIX 444

*Table 1* *Normal-Curve Areas* 445
*Table 2* *Percentage Points of the t Distribution* 446
*Table 3* *Percentage Points of the Chi-Square Distribution* 447
*Table 4* *Percentage Points of the F Distribution.* $a = .05$ 449
*Table 5* *Percentage Points of the F Distribution.* $a = .01$ 451
*Table 6* *Squares and Square Roots* 453
*Table 7* *Critical Values for* $D = \max |F_0(x) - S_n(x)|$ *in the Kolmogorov–Smirnov Test for Goodness of Fit: One Sample* 468
*Table 8* *Random Numbers* 469

Glossary 470
Answers to Exercises 475
Index 503

# 1

# WHAT IS STATISTICS?

1.1 PRELUDE

1.2 WHY STUDY STATISTICS

1.3 THE INTERPLAY BETWEEN STATISTICS AND THE SOCIAL SCIENCES

1.4 A NOTE TO THE STUDENT

QUESTIONS AND PROBLEMS

REFERENCE

## 1.1 PRELUDE

What is statistics? Is it the addition of numbers? Is it graphs, means, medians, and modes, batting averages, percentages of passes completed, percent unemployed—in general, tedious descriptions of society and nature by means of a set of figures? Or is statistics a modern and scientific method for penetrating the unknown, that is, an exciting way of acquiring new information about the world in which we live? True, it involves numbers, and an understanding of the theory of statistics requires a basic knowledge of mathematics. But the concepts behind statistics—what it is attempting to do and how we go about it—are reasonable, easily explained, and understandable. The purpose of our text is to answer the questions: What is statistics? How does it work? How can it be applied in the social sciences? In general, we seek to provide the student with an understanding of the basic concepts of statistics and to tell him how he can use statistics as a tool. We shall do so while keeping in mind that many students in the social sciences possess relatively little mathematical training and often lack self-confidence in this area.

What is statistics? Statistics can be viewed from several perspectives. Because this text is written for students in the social sciences, we will write about statistics from a social science perspective. In so doing we need to consider two types of statistics, descriptive statistics and inferential statistics. Both are important tools in the social sciences.

What is descriptive statistics? The social scientist is often confronted with a mass of data that needs to be described. He might, for example, have specific bits of information, such as the total family income of each student in the university. If he wishes to describe the university in terms of the total family income of each student, the listing of 25,000 family incomes would be cumbersome indeed. Descriptive statistics provides us with the techniques that enable us to describe the university *concisely* in terms of the total family income of its students. We might also want to describe the drug habits of the residents of New York State, occupational status of the residents of Chicago, the religious beliefs of the residents of Tieton, or the preschool preparation of first graders currently attending Lincoln Grade School. In every instance, the researcher (it is assumed) has information on each subject in the population and merely wishes to describe a characteristic of New York State, the city of Chicago, the village of Tieton, or the first grade at Lincoln.

What is inferential statistics? Consider this story: The presidential elections of 1960 and earlier were laced with suspense. The polls closed at 7:00 or 8:00 P.M., but television viewers were treated to an exciting evening of uncertainty that, in most cases, did not end until early morning, when all the votes had been counted. Well-known political analysts exhibited a great lack of respect for the electronic monster—the computer—that had begun invading their areas of expertise, and they left no doubt in the minds of television viewers that man's judgment could never be replaced by the calculations of a machine. Yet in 1964 a Johnson victory over Goldwater was confidently forecast before midnight; and even the skeptical news analysts seemed convinced of the "monster's" predictions. An era ended in 1968 when, as viewers sat before their television sets, Nixon's victory was predicted early in the evening. The same dubious news analysts who had slighted computer calculations in 1960 and 1964 were forecasting the outcome of the 1968 election with great certainty. Conclusions in some states were forecast as early as a half-hour after the closing of the polls. Many viewers turned away from their television sets in disgust. The great race, the exciting evening they had anticipated, was over before it had begun, terminated by the great success of the electronic computer.

Our view of the evening was different. Knowing that computers only follow human instructions, we recognized the rapid and successful prediction of the election victor as an achievement of statistics. So what is inferential statistics?

The best way to answer this question is to consider a few examples. Suppose that you wish to estimate the proportion of all migrant workers in a particular state who have completed six years of public school education. Suppose, further, that you have obtained a fairly complete list of these workers from the state records for workmen's accident compensation and you have to decide how you will collect the pertinent information. One method, extremely costly and time consuming, would be to interview each of the migrant workers. But an easier and more efficient approach would be to randomly sample a thousand workers from the large body of all migrant workers and to contact each of these persons individually. You could then use the proportion of workers in the sample who possess six years of education to *estimate* the proportion of all migrant workers in the state who possess the same characteristic. We shall show later that the sample proportion will be quite close to the proportion for all migrant workers in the state. In addition, we shall be able to place a limit on the difference between the estimate and the true value of the proportion.

A second example is that of the Gallup, Harris, and other public opinion polls. How can these pollsters presume to know the opinions of more than 100 million Americans? They certainly cannot reach their conclusions by contacting every person in the United States. Instead, they randomly sample the opinions of a small number of citizens, often as few as 900, to *estimate* the reaction of every person in the country. The amazing result of this process is that the proportion in the sample who hold a particular opinion will match very closely the proportion of voters holding that opinion in the complete population from which the sample was drawn. Some students may find this assertion difficult to believe; convincing supportive evidence will be supplied in subsequent chapters.

As a third example, suppose that we wish to estimate the average number of children in families that live in the urban area of a large city. One way to do this would be to interview every family in the city, but this method has obvious disadvantages. It would be very costly to conduct so many interviews, and it would be incomplete and subject to inaccuracies because some families would not be at home when the interviewer made his visit. A second way, and one that employs the concepts of statistics, would be to select a random sample of 400 families from the city and interview each family. Families not at home would be revisited until they were interviewed. Then we would *estimate* the true average number of children in the family for the entire city by calculating the average for the sample of 400 families. This method of determining the average number of children per family is much less costly than sampling the entire city and, as we shall show later, it is very accurate. The difference between the sample estimate and the true average number of children per family is usually quite small.

Let us now try to identify from the examples the characteristics common to all inferential statistical problems. First, each example involved making an observation or measurement that could not be predicted with certainty in advance.

an unpredictable (random) manner. We cannot say in advance whether a particular migrant worker selected from the state list will possess six years of education, whether a randomly selected voter will vote for Jones, or exactly how many children a particular family will have.

Second, each example involved sampling. A sample of migrant workers was selected from the state list, a sample of people was taken from the entire voting population of the United States, and a sample of 400 families was obtained from the total number of families in the city.

Third, although not obvious, each example involved the collection of data or measurements, one measurement corresponding to each element of the sample. Each migrant worker is interviewed and is assigned a score: $X = 1$ if he has had as many as six years of education and $X = 0$ if he has not. The total number of migrant workers with at least six years of education in the sample is the sum of these measurements. Scores, representing ratings, are measurements. Similar measurements are obtained when we sample voter intentions; and the sampling of the number of children in the 400 households yields 400 measurements, one for each family.

Finally, each example exhibits a common objective. That is, the purpose of sampling is to obtain information and to make an inference about a much larger set of measurements called the population. For the estimation problem concerning migrant workers, the population of interest is the large set of 1s and 0s corresponding to those who have and those who have not had at least six years of education. Similarly, the population for the voter problem would be the set of 1s and 0s corresponding to the 100 million or more voters in the United States. Each voter would be assigned a 1 if he intended to vote for candidate Jones and a 0 if he did not. The objective of sampling is to estimate the proportion of eligible voters who favor Jones, that is, the proportion of 1s in the population. The population associated with the household survey is the number of children per household recorded for all families in the city. The objective is to infer the mean number of children per family. In other words, the researcher wishes to make an inference about a large set of families based on information contained in the sample of 400 families. Note again that populations are collections of measurements or scores and are not collections of people (in accordance with the usual connotation of the term). Note also that populations may exist in fact or may be imaginary. The populations for the three examples exist even though we do not actually possess the complete collection of 1s and 0s corresponding to the two educational categories for migrant workers, or the entire set of voters favoring or opposing Jones. In contrast, if we sample the pulse rates of a set of heart patients, the population of pulse rates measured on all heart patients in the future is a product of the physician's imagination and thus is an imaginary population.

When we group the four characteristics—random observations, sampling, numerical data, and a common inferential objective—we might think to define *inferential statistics* as a theory of information. Information is obtained by experimentation or, equivalently, by *sampling*. The data are employed to make an inference about a larger set of measurements, existing or imaginary, called a *population*. Inferences about populations are usually expressed as estimates of or as decisions about one or more characteristics of the population.

Relevant definitions are as follows:

**Definition 1.1**

*A* population *is the set representing all measurements of interest to the researcher.*

**Definition 1.2**

*A* sample *is a subset of measurements selected from the population of interest.*