STATISTICAL METHODS FOR THE SOCIAL SCIENCES

Fourth Edition

Alan Agresti
University of Florida

Barbara Finlay
Texas A & M University

Upper Saddle River, New Jersey 07458

CHAPTER 1

Introduction

1.1 INTRODUCTION TO STATISTICAL METHODOLOGY

The past quarter-century has seen a dramatic increase in the use of statistical methods in the social sciences. There are many reasons for this. More research in the social sciences has taken on a quantitative orientation. Like research in other sciences, research in the social sciences often studies questions of interest by analyzing evidence provided by empirical data. The growth of the Internet has resulted in an increase in the amount of readily available quantitative information. Finally, with the evolution of evermore powerful computers, software, and statistical methodology, new methods are available that can more realistically address the questions that arise in social science research.

Why Study Statistics?

The increased use of statistics is evident in the changes in the content of articles published in social science research journals and reports prepared in government and private industry. A quick glance through recent issues of journals such as *American Political Science Review* and *American Sociological Review* reveals the fundamental role of statistics in research. For example, to learn about which factors have the greatest impact on student performance in school or to investigate which factors affect people's political beliefs or the quality of their health care or their decision about when to retire, researchers collect information and process it using statistical analyses. Because of the role of statistics in many research studies, more and more academic departments require that their majors take statistics courses.

These days, social scientists work in a wide variety of areas that use statistical methods, such as governmental agencies, business organizations, and health care facilities. For example, social scientists in government agencies dealing with human welfare or environmental issues or public health policy invariably need to use statistical methods or at least read reports that contain statistics. Medical sociologists often must evaluate recommendations from studies that contain quantitative investigations of new therapies or new ways of caring for the elderly. Some social scientists help managers to evaluate employee performance using quantitative benchmarks and to determine factors that help predict sales of products. In fact, increasingly many jobs for social scientists expect a knowledge of statistical methods as a basic work tool. As the joke goes, "What did the sociologist who passed statistics say to the sociologist who failed it? 'I'll have a Big Mac, fries, and a Coke.' "

But an understanding of statistics is important even if you never use statistical methods in your career. Every day you are exposed to an explosion of information, from advertising, news reporting, political campaigning, surveys about opinions on controversial issues, and other communications containing statistical arguments. Statistics helps you make sense of this information and better understand the world. You will find concepts from this text helpful in judging the information you will encounter in your everyday life.

We realize you are not reading this book in hopes of becoming a statistician. In addition, you may suffer from math phobia and feel fear at what lies ahead. Please be assured that you can read this book and learn the primary concepts and methods of statistics with little knowledge of mathematics. Just because you may have had difficulty in math courses before does not mean you will be at a disadvantage here. To understand this book, logical thinking and perseverance are more important than mathematics. In our experience, the most important factor in how well you do in a statistics course is how much time you spend on the course—attending class, doing homework, reading and re-reading this text, studying your class notes, working together with your fellow students, getting help from your professor or teaching assistant—not your mathematical knowledge or your gender or your race or whether you feel fear of statistics at the beginning.

Don't be frustrated if learning comes slowly and you need to read a chapter a few times before it starts to make sense. Just as you would not expect to take a single course in a foreign language and be able to speak that language fluently, the same is true with the language of statistics. Once you have completed even a portion of this text, however, you will better understand how to make sense of statistical information.

Data

Information gathering is at the heart of all sciences, providing the ***observations*** used in statistical analyses. The observations gathered on the characteristics of interest are collectively called ***data***.

For example, a study might conduct a survey of 1000 people to observe characteristics such as opinion about the legalization of marijuana, political party affiliation, political ideology, how often attend religious services, number of years of education, annual income, marital status, race, and gender. The data for a particular person would consist of observations such as (opinion = do not favor legalization, party = Republican, ideology = conservative, religiosity = once a week, education = 14 years, annual income in range 40–60 thousand dollars, marital status = married, race = white, gender = female). Looking at the data in the right way helps us learn about how such characteristics are related. We can then answer questions such as, "Do people who attend church more often tend to be more politically conservative?"

To generate data, the social sciences use a wide variety of methods, including surveys, experiments, and direct observation of behavior in natural settings. In addition, social scientists often analyze data already recorded for other purposes, such as police records, census materials, and hospital files. Existing archived collections of data are called ***databases***. Many databases are now available on the Internet. A very important database for social scientists contains results since 1972 of the General Social Survey.

EXAMPLE 1.1 The General Social Survey (GSS)

Every other year, the National Opinion Research Center at the University of Chicago conducts the General Social Survey (GSS). This survey of about 2000 adults provides data about opinions and behaviors of the American public. Social scientists use it to investigate how adult Americans answer a wide diversity of questions, such as, "Do you believe in life after death?," "Would you be willing to pay higher prices in order to protect the environment?," and "Do you think a preschool child is likely to suffer if his or her mother works?" Similar surveys occur in other countries, such as the General Social Survey administered by Statistics Canada, the British Social Attitudes Survey, and the Eurobarometer survey and European Social Survey for nations in the European Union.

It is easy to get summaries of data from the GSS database. We'll demonstrate, using a question it asked in one survey, "About how many good friends do you have?"

- Go to the Web site sda.berkeley.edu/GSS/ at the Survey Documentation and Analysis site at the University of California, Berkeley.
- Click on *New SDA*.
- The GSS name for the question about number of good friends is NUMFREND. Type NUMFREND as the *Row* variable name. Click on *Run the table.*

Now you'll see a table that shows the possible values for 'number of good friends' and the number of people and the percentage who made each possible response. The most common responses were 2 and 3 (about 16% made each of these responses). ■

What Is Statistics?

In this text, we use the term "statistics" in the broad sense to refer to methods for obtaining and analyzing data.

Statistics

Statistics consists of a body of methods for obtaining and analyzing data.

Specifically, statistics provides methods for

1. ***Design***: Planning how to gather data for research studies
2. ***Description***: Summarizing the data
3. ***Inference***: Making predictions based on the data

Design refers to planning how to obtain the data. For a survey, for example, the design aspects would specify how to select the people to interview and would construct the questionnaire to administer.

Description refers to summarizing data, to help understand the information they provide. For example, an analysis of the number of good friends based on the GSS data might start with a list of the number reported for each of the people who responded to that question that year. The raw data are a complete listing of observations, person by person. These are not easy to comprehend, however. We get bogged down in numbers. For presentation of results, instead of listing *all* observations, we could summarize the data with a graph or table showing the percentages reporting 1 good friend, 2 good friends, 3, . . ., and so on. Or we could report the average number of good friends, which was 6, or the most common response, which was 2. Graphs, tables and numerical summaries are called ***descriptive statistics***.

Inference refers to making predictions based on data. For instance, for the GSS data on reported number of good friends, 6.2% reported having only 1 good friend. Can we use this information to predict the percentage of the more than 200 million adults in the U.S. at that time who had only 1 good friend? A method presented in this book allows us to predict that that percentage is no greater than 8%. Predictions made using data are called ***statistical inferences***.

Description and ***inference*** are the two types of ***statistical analysis***—ways of analyzing the data. Social scientists use descriptive and inferential statistics to answer questions about social phenomena. For instance, "Is having the death penalty

available for punishment associated with a reduction in violent crime?" "Does student performance in schools depend on the amount of money spent per student, the size of the classes, or the teachers' salaries?"

1.2 DESCRIPTIVE STATISTICS AND INFERENTIAL STATISTICS

Section 1.1 explained that statistics consists of methods for *designing* studies and *analyzing* data collected in the studies. Methods for analyzing data include descriptive methods for summarizing the data and inferential methods for making predictions. A statistical analysis is classified as ***descriptive*** or ***inferential***, according to whether its main purpose is to describe the data or to make predictions. To explain this distinction further, we next define the *population* and the *sample*.

Populations and Samples

The entities that a study observes are called the ***subjects*** for the study. Usually the subjects are people, such as in the GSS, but they might instead be families, schools, cities, or companies, for instance.

Population and Sample

The ***population*** is the total set of subjects of interest in a study. A ***sample*** is the subset of the population on which the study collects data.

In the 2004 GSS, the sample was the 2813 adult Americans who participated in the survey. The population was all adult Americans at that time—more than 200 million people.

The ultimate goal of any study is to learn about populations. But it is almost always necessary, and more practical, to observe only samples from those populations. For example, the GSS and polling organizations such as the Gallup poll usually select samples of about 1000–3000 Americans to collect information about opinions and beliefs of the population of *all* Americans.

Descriptive Statistics

Descriptive statistics summarize the information in a collection of data.

Descriptive statistics consist of graphs, tables, and numbers such as averages and percentages. The main purpose of descriptive statistics is to reduce the data to simpler and more understandable forms without distorting or losing much information.

Although data are usually available only for a sample, descriptive statistics are also useful when data are available for the entire population, such as in a census. By contrast, inferential statistics apply when data are available only for a sample but we want to make a prediction about the entire population.

Inferential Statistics

Inferential statistics provide predictions about a population, based on data from a sample of that population.

EXAMPLE 1.2 Belief in Heaven

In two of its surveys, the GSS asked, "Do you believe in heaven?" The population of interest was the collection of all adults in the United States. In the most recent survey

in which this was asked, 86% of the 1158 sampled subjects answered *yes*. We would be interested, however, not only in those 1158 people but in the *entire population* of all adults in the U.S.

Inferential statistics provide a prediction about the larger population using the sample data. An inferential method presented in Chapter 5 predicts that the population percentage that believe in heaven falls between 84% and 88%. That is, the sample value of 86% has a "margin of error" of 2%. Even though the sample size was tiny compared to the population size, we can conclude that a large percentage of the population believed in heaven. ■

Inferential statistical analyses can predict characteristics of entire populations quite well by selecting samples that are small relative to the population size. That's why many polls sample only about a thousand people, even if the population has millions of people. In this book, we'll see why this works.

In the past quarter-century, social scientists have increasingly recognized the power of inferential statistical methods. Presentation of these methods occupies a large portion of this textbook, beginning in Chapter 5.

Parameters and Statistics

Parameters and Statistics

A ***parameter*** is a numerical summary of the population. A ***statistic*** is a numerical summary of the sample data.

Example 1.2 estimated the percentage of Americans who believe in heaven. The parameter was the population percentage who believed in heaven. Its value was unknown. The inference about this parameter was based on a statistic—the percentage of the 1158 subjects interviewed in the survey who answered *yes*, namely, 86%. Since this number *describes* a characteristic of the sample, it is a descriptive statistic.

In practice, the main interest is in the values of the parameters, not the values of the statistics for the particular sample selected. For example, in viewing results of a poll before an election, we're more interested in the *population* percentages favoring the various candidates than in the *sample* percentages for the people interviewed. The sample and statistics describing it are important only insofar as they help us make inferences about unknown population parameters.

An important aspect of statistical inference involves reporting the likely *precision* of the sample statistic that estimates the population parameter. For Example 1.2 on belief in heaven, an inferential statistical method predicted how close the *sample* value of 86% was likely to be to the unknown percentage of the *population* believing in heaven. The reported margin of error was 2%.

When data exist for an entire population, such as in a census, it's possible to find the actual values of the parameters of interest. Then there is no need to use inferential statistical methods.

Defining Populations: Actual and Conceptual

Usually the population to which inferences apply is an actual set of subjects. In Example 1.2, it was adult residents of the U.S. Sometimes, though, the generalizations refer to a *conceptual* population—one that does not actually exist but is hypothetical.

For example, suppose a consumer organization evaluates gas mileage for a new model of an automobile by observing the average number of miles per gallon for five sample autos driven on a standardized 100-mile course. Their inferences refer to the performance on this course for the conceptual population of *all* autos of this model that will be or could hypothetically be manufactured.

1.3 THE ROLE OF COMPUTERS IN STATISTICS

Over time, ever more powerful computers reach the market, and powerful and easy-to-use software is further developed for statistical methods. This software provides an enormous boon to the use of statistics.

Statistical Software

SPSS (Statistical Package for the Social Sciences), SAS, MINITAB, and Stata are the most popular statistical software on college campuses. It is much easier to apply statistical methods using these software than using hand calculation. Moreover, many methods presented in this text are too complex to do by hand or with hand calculators.

Most chapters of this text, including all those that present methods requiring considerable computation, show examples of the output of statistical software. One purpose of this textbook is to teach you what to look for in output and how to interpret it. Knowledge of computer programming is not necessary for using statistical software or for reading this book.

The text appendix explains how to use SPSS and SAS, organized by chapter. You can refer to this appendix as you read each chapter to learn how to use them to perform the analyses of that chapter.

Data Files

Figure 1.1 shows an example of data organized in a ***data file*** for analysis by statistical software. A data file has the form of a spreadsheet:

- Any one row contains the observations for a particular subject in the sample.
- Any one column contains the observations for a particular characteristic.

Figure 1.1 is a window for editing data in SPSS. It shows data for the first ten subjects in a sample, for the characteristics sex, racial group, marital status, age, and annual income (in thousands of dollars). Some of the data are numerical, and some consist of labels. Chapter 2 introduces the types of data for data files.

Uses and Misuses of Statistical Software

A note of caution: The easy access to statistical methods using software has dangers as well as benefits. It is simple to apply inappropriate methods. A computer performs the analysis requested whether or not the assumptions required for its proper use are satisfied.

Incorrect analyses result when researchers take insufficient time to understand the statistical method, the assumptions for its use, or its appropriateness for the specific problem. It is vital to understand the method before using it. Just knowing how to use statistical software does not guarantee a proper analysis. You'll need a good background in statistics to understand which method to select, which options to choose in that method, and how to make valid conclusions from the output. The main purpose of this text is to give you this background.

chapter_1_data.sav [DataSet1] - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help

1 : sex female

	subject	sex	race	married	age	income
1	1	female	white	yes	23	18.3
2	2	female	black	no	37	31.9
3	3	male	white	yes	47	64.0
4	4	female	white	yes	61	46.2
5	5	male	hispanic	yes	30	16.5
6	6	male	white	no	21	14.0
7	7	male	white	yes	55	26.1
8	8	female	white	no	27	59.8
9	9	female	hispanic	yes	61	21.5
10	10	male	black	no	47	50.0

Data View / Variable View

SPSS Processor is ready

FIGURE 1.1: Example of Part of a SPSS Data File

1.4 CHAPTER SUMMARY

The field of statistics includes methods for

- designing research studies,
- describing the data, and
- making inferences (predictions) using the data.

Statistical methods normally are applied to observations in a ***sample*** taken from the ***population*** of interest. ***Statistics*** summarize sample data, while ***parameters*** summarize entire populations. There are two types of statistical analyses:

- ***Descriptive statistics*** summarize sample or population data with numbers, tables, and graphs.
- ***Inferential statistics*** make predictions about population parameters, based on sample data.

A ***data file*** has a separate row of data for each subject and a separate column for each characteristic. Statistical methods are easy to apply to data files using software. This relieves us of computational drudgery and helps us focus on the proper application and interpretation of the methods.

PROBLEMS

Practicing the Basics

1.1. The Environmental Protection Agency (EPA) uses a few new automobiles of each brand every year to collect data on pollution emission and gasoline mileage performance. For the Toyota Prius brand, identify the (a) subject, (b) sample, (c) population.

1.2. In the 2006 gubernatorial election in California, an exit poll sampled 2705 of the 7 million people who voted. The poll stated that 56.5%

reported voting for the Republican candidate, Arnold Schwarzenegger. Of all 7 million voters, 55.9% voted for Schwarzenegger.

(a) For this exit poll, what was the population and what was the sample?

(b) Identify a statistic and a parameter.

1.3. The student government at the University of Wisconsin conducts a study about alcohol abuse among students. One hundred of the 40,858 members of the student body are sampled and asked to complete a questionnaire. One question asked is, "On how many days in the past week did you consume at least one alcoholic drink?"

(a) Identify the population of interest.

(b) For the 40,858 students, one characteristic of interest was the percentage who would respond *zero* to this question. This value is computed for the 100 students sampled. Is it a parameter or a statistic? Why?

1.4. The Institute for Public Opinion Research at Florida International University has conducted the FIU/Florida Poll (www.fiu.edu/orgs/ipor/ffp) of about 1200 Floridians annually since 1988 to track opinions on a wide variety of issues. The poll reported in 2006 that 67% of Floridians believe that state government should not make laws restricting access to abortion. Is 67% the value of a statistic, or of a parameter? Why?

1.5. A GSS asked subjects whether astrology—the study of star signs—has some scientific truth (GSS question SCITEST3). Of 1245 sampled subjects, 651 responded *definitely or probably true*, and 594 responded *definitely or probably not true*. The proportion responding *definitely or probably true* was $651/1245 = 0.523$.

(a) Describe the population of interest.

(b) For what population parameter might we want to make an inference?

(c) What sample statistic could be used in making this inference?

(d) Does the value of the statistic in (c) necessarily equal the parameter in (b)? Explain.

1.6. Go to the GSS Web site, sda.berkeley.edu/GSS/. By entering TVHOURS as the *Row variable*, find a summary of responses to the question, "On a typical day, about how many hours do you personally watch television?"

(a) What was the most common response?

(b) Is your answer in (a) a descriptive statistic, or an inferential statistic?

1.7. Go to the GSS Web site, sda.berkeley.edu/GSS/. By entering HEAVEN as the *Row variable*, you can find the percentages of people who said *definitely yes*, *probably yes*, *probably not*, and *definitely not* when asked whether they believed in heaven.

(a) Report the percentage who gave one of the *yes* responses.

(b) To obtain data for a particular year such as 1998, enter YEAR(1998) in the *Selection filter* option box before you click on *Run the Table*. Do this for HEAVEN in 1998, and report the percentage who gave one of the *yes* responses. (This question was asked only in 1991 and 1998.)

(c) Summarize opinions in 1998 about belief in hell (variable HELL in the GSS). Was the percentage of *yes* responses higher for HEAVEN or HELL?

1.8. The Current Population Survey (CPS) is a monthly survey of households conducted by the U.S. Census Bureau. A CPS of 60,000 households indicated that of those households, 8.1% of the whites, 22.3% of the blacks, 20.9% of the Hispanics, and 10.2% of the Asians had annual income below the poverty level (*Statistical Abstract of the United States, 2006*).

(a) Are these numbers statistics, or parameters? Explain.

(b) A method from this text predicts that the percentage of *all* black households in the United States having income below the poverty level is at least 21% but no greater than 24%. What type of statistical method does this illustrate—descriptive or inferential? Why?

1.9. A BBC story (September 9, 2004) about a poll in 35 countries concerning whether people favored George W. Bush or John Kerry in the 2004 U.S. Presidential election stated that Kerry was clearly preferred. Of the sample from Germany, 74% preferred Kerry, 10% preferred Bush, with the rest undecided or not responding. Multiple choice: The results for Germany are an example of

(a) descriptive statistics for a sample.

(b) inferential statistics about a population.

(c) a data file.

(d) a population.

1.10. Construct a data file describing the criminal behavior of five inmates in a local prison. The characteristics measured were race (with observations for the five subjects: white, black, white, Hispanic, white), age (19, 23, 38, 20, 41), length of sentence in years (2, 1, 10, 2, 5), whether convicted on a felony (no, no, yes, no, yes), number of prior arrests (values 2, 0, 8, 1, 5), number of prior convictions (1, 0, 3, 1, 4).

Concepts and Applications

1.11. The "Student survey" data file at *www.stat.ufl.edu/~aa/social/data.html* shows responses of a class of social science graduate students at the University of Florida to a questionnaire that asked about GE = gender, AG = age in years, HI = high school GPA (on a four-point scale), CO = college GPA, DH = distance (in miles) of the campus from your home town, DR = distance (in miles) of the classroom from your current residence, NE = number of times a week you read a newspaper, TV = average number of hours per week that you watch TV, SP = average number of hours per week that you participate in sports or have other physical exercise, VE = whether you are a vegetarian (yes, no), AB = opinion about whether abortion should be legal in the first three months of pregnancy (yes, no), PI = political ideology (1 = very liberal, 2 = liberal, 3 = slightly liberal, 4 = moderate, 5 = slightly conservative, 6 = conservative, 7 = very conservative), PA = political affiliation (D = Democrat, R = Republican, I = independent), RE = how often you attend religious services (never, occasionally, most weeks, every week), LD = belief in life after death (yes, no), AA = support affirmative action (yes, no), AH = number of people you know who have died from AIDS or who are HIV+. You will use this data file for exercises in future chapters.

(a) Practice accessing a data file for statistical analysis with your software by going to this Web site and copying this data file. Print a copy of the data file. How many observations (rows) are in the data file?

(b) Give an example of a question that could be addressed using these data with (i) descriptive statistics, (ii) inferential statistics.

1.12. Using a spreadsheet program (such as Microsoft Office Excel) or statistical software, your instructor will help the class create a data file consisting of the values for class members of characteristics such as those in the previous exercise. One exercise in each chapter will use this data file.

(a) Copy the data file to your computer and print a copy.

(b) Give an example of a question that you could address by analyzing these data with (i) descriptive statistics, (ii) inferential statistics.

1.13. For the statistical software your instructor has chosen for your course, find out how to access the software, enter data, and print any data files that you create. Create a data file using the data in Figure 1.1 in Section 1.3, and print it.

1.14. Illustrating with an example, explain the difference between

(a) a *statistic* and a *parameter*.

(b) *description* and *inference* as two purposes for using statistical methods.

1.15. You have data for a population, from a census. Explain why descriptive statistics are helpful but inferential statistics are not needed.

1.16. A sociologist wants to estimate the average age at marriage for women in New England in the early eighteenth century. She finds within her state archives marriage records for a large Puritan village for the years 1700–1730. She then takes a sample of those records, noting the age of the bride for each. The average age in the sample is 24.1 years. Using a statistical method from Chapter 5, the sociologist estimates the average age of brides at marriage for the population to be between 23.5 and 24.7 years.

(a) What part of this example is descriptive?

(b) What part of this example is inferential?

(c) To what population does the inference refer?

1.17. In a recent survey by Eurobarometer of Europeans about energy issues and global warming,[1] one question asked, "Would you be willing to pay more for energy produced from renewable sources than for energy produced from other sources?" The percentage of *yes* responses varied among countries between 10% (in Bulgaria) to 60% (in Luxembourg). Of the 631 subjects interviewed in the UK, 45% said *yes*. It was predicted that for all 48 million adults in the UK, that percentage who would answer *yes* falls between 41% and 49%. Identify in this discussion (a) a statistic, (b) a parameter, (c) a descriptive statistical analysis, (d) an inferential statistical analysis.

1.18. Go to the Web site for the Gallup poll, www.galluppoll.com. From information listed on or linked from the homepage, give an example of a (a) descriptive statistical analysis, (b) inferential statistical analysis.

1.19. Check whether you have access to JSTOR (Journal Storage) at your school by visiting www.jstor.org. If so, click on *Browse* and then *Sociology* or another discipline of interest to you. Select a journal and a particular issue, and browse through some of the articles. Find an article that uses statistical methods. In a paragraph of 100–200 words, explain how descriptive statistics were used.

[1] *Attitudes towards Energy*, published January 2006 at ec.europa.eu/public_opinion