Fundamentals of Statistics

Every day, you encounter numerical information that describes or analyzes some aspect of the world you live in. For example, here are some news items that appeared in the pages of *The New York Times* during a one-month period:

- Between 1969 and 2001, the rate of forearm fractures rose 52% for girls and 32% for boys, with the largest increases among children in early puberty, according to a recent Mayo Clinic study.
- Across the New York metropolitan area, the median sales price of a single-family home has risen by 75% since 1998, an increase of more than $140,000.
- A study that explored the relationship between the price of a book and the number of copies of a book sold found that raising prices by 1% reduced sales by 4% at BN.com, but reduced sales by only 0.5% at Amazon.com.

Such stories as these would not be possible to understand without **statistics**, the branch of mathematics that consists of methods of processing and analyzing data to better support rational decision-making processes. Using statistics to better understand the world means more than just producing a new set of numerical information—you must *interpret* the results by reflecting on the significance and the importance of the results to the decision-making

process you face. Interpretation also means knowing when to ignore results, either because they are misleading, are produced by incorrect methods, or just restate the obvious, as this news story "reported" by the comedian David Letterman illustrates:

> *USA Today* has come out with a new survey. Apparently, 3 out of every 4 people make up 75% of the population.

As newer technologies allow people to process and analyze ever-increasing amounts of data, statistics plays an increasingly important part of many decision-making processes today. Reading this chapter will help you understand the fundamentals of statistics and introduce you to concepts that are used throughout this book.

1.1 The Five Basic Words of Statistics

The five words *population*, *sample*, *parameter*, *statistic* (singular), and *variable* form the basic vocabulary of statistics. You cannot learn much about statistics unless you first learn the meanings of these five words.

Population

CONCEPT All the members of a group about which you want to draw a conclusion.

EXAMPLES All U.S. citizens who are currently registered to vote, all patients treated at a particular hospital last year, the entire daily output of a cereal factory's production line.

Sample

CONCEPT The part of the population selected for analysis.

EXAMPLES The registered voters selected to participate in a recent survey concerning their intention to vote in the next election, the patients selected to fill out a patient-satisfaction questionnaire, 100 boxes of cereal selected from a factory's production line.

Parameter

CONCEPT A numerical measure that describes a characteristic of a population.

EXAMPLES The percentage of all registered voters who intend to vote in the next election, the percentage of all patients who are very satisfied with the care they received, the average weight of all the cereal boxes produced on a factory's production line on a particular day.

Statistic

CONCEPT A numerical measure that describes a characteristic of a sample.

EXAMPLES The percentage in a sample of registered voters who intend to vote in the next election, the percentage in a sample of patients who are very satisfied with the care they received, the average weight of a sample of cereal boxes produced on a factory's production line on a particular day.

INTERPRETATION Calculating statistics for a sample is the most common activity, because collecting population data is impractical for most actual decision-making situations.

Variable

CONCEPT A characteristic of an item or an individual that will be analyzed using statistics.

EXAMPLES Gender, the household income of the citizens who voted in the last presidential election, the publishing category (hardcover, trade paperback, mass-market paperback, textbook) of a book, the number of varieties of a brand of cereal.

INTERPRETATION All the variables taken together form the data of an analysis. Although you may have heard people saying that they are analyzing their data, they are, more precisely, analyzing their variables.

You should distinguish between a variable, such as gender, and its **value** for an individual, such as male. An **observation** is all the values for an individual item in the sample. For example, a survey might contain two variables, gender and age. The first observation might be male, 40. The second observation might be female, 45. The third observation might be female, 55. A **variable** is sometimes known as a column of data because of the convention of entering each observation as a unique row in a table of data. (Likewise, you may hear some refer to an observation as a row of data.)

Variables can be divided into the following types:

	Categorical Variables	Numerical Variables
Concept	The values of these variables are selected from an established list of categories.	The values of these variables involve a counted or measured value.
Subtypes	None.	**Discrete values** are counts of things. **Continuous values** are measures, and any value can theoretically occur, limited only by the precision of the measuring process.

(*continues*)

	Categorical Variables	Numerical Variables
Examples	Gender, a variable that has the categories male and female. Academic major, a variable that might have the categories English, Math, Science, and History, among others.	The number of previous presidential elections in which a citizen voted, a discrete numerical variable. The household income of a citizen who voted, a continuous variable.

All variables should have an **operational definition**—that is, a universally-accepted meaning that is clear to all associated with an analysis. Without operational definitions, confusion can occur. A famous example of such confusion was the tallying of votes in Florida during the 2000 U.S. presidential election in which, at various times, nine different definitions of a valid ballot were used. (A later analysis[1] determined that three of these definitions, including one pursued by Al Gore, led to margins of victory for George Bush that ranged from 225 to 493 votes and that the six others, including one pursued by George Bush, led to margins of victory for Al Gore that ranged from 42 to 171 votes.)

1.2 The Branches of Statistics

Two branches, *descriptive statistics* and *inferential statistics*, comprise the field of statistics.

Descriptive Statistics

CONCEPT The branch of statistics that focuses on collecting, summarizing, and presenting a set of data.

EXAMPLES The average age of citizens who voted for the winning candidate in the last presidential election, the average length of all books about statistics, the variation in the weight of 100 boxes of cereal selected from a factory's production line.

INTERPRETATION You are most likely to be familiar with this branch of statistics, because many examples arise in everyday life. Descriptive statistics forms the basis for analysis and discussion in such diverse fields as securities

[1] J. Calmes and E. P. Foldessy, "In Election Review, Bush Wins with No Supreme Court Help," *Wall Street Journal*, November 12, 2001, A1, A14

trading, the social sciences, government, the health sciences, and professional sports. A general familiarity and widespread availability of descriptive methods in many calculating devices and business software can often make using this branch of statistics seem deceptively easy. (Chapters 2 and 3 warn you of the common pitfalls of using descriptive methods.)

Inferential Statistics

CONCEPT The branch of statistics that analyzes sample data to draw conclusions about a population.

EXAMPLE A survey that sampled 2,001 full- or part-time workers ages 50 to 70, conducted by the American Association of Retired Persons (AARP), discovered that 70% of those polled planned to work past the traditional mid-60s retirement age. By using methods discussed in Section 6.4, this statistic could be used to draw conclusions about the population of all workers ages 50 to 70.

INTERPRETATION When you use inferential statistics, you start with a hypothesis and look to see whether the data are consistent with that hypothesis. Inferential statistical methods can be easily misapplied or misconstrued, and many inferential methods require the use of a calculator or computer. (A full explanation of common inferential methods appears in Chapters 6 through 9.)

1.3 Sources of Data

All statistical analysis begins by identifying the source of the data. Among the important sources of data are *published sources*, *experiments*, and *surveys*.

Published Sources

CONCEPT Data available in print or in electronic form, including data found on Internet Web sites. Primary data sources are those published by the individual or group that collected the data. Secondary data sources are those compiled from primary sources.

EXAMPLES Many U.S. federal agencies, including the Census Bureau, publish primary data sources that are available at the Web site www.fedstats.gov. Business news sections of daily newspapers commonly publish secondary source data compiled by business organizations and government agencies.

INTERPRETATION You should always consider the possible bias of the publisher and whether the data contain all the necessary and relevant variables

when using published sources. Remember, too, that *anyone* can publish data on the Internet.

Experiments

CONCEPT A process that studies the effect on a variable of varying the value(s) of another variable or variables, while keeping all other things equal. A typical experiment contains both a treatment group and a control group. The treatment group consists of those individuals or things that receive the treatment(s) being studied. The control group consists of those individuals or things that do not receive the treatment(s) being studied.

EXAMPLE Pharmaceutical companies use experimental studies to determine whether a new drug is effective. A group of patients who have many similar characteristics is divided into two subgroups. Members of one group, the treatment group, receive the new drug. Members of the other group, the control group, receive a **placebo**, a substance that has no medical effect. After a time period, statistics about each group are compared.

INTERPRETATION Proper experiments are either single-blind or double-blind. A study is a single-blind experiment if only the researcher conducting the study knows the identities of the members of the treatment and control groups. If neither the researcher nor study participants know who is in the treatment group and who is in the control group, the study is a double-blind experiment.

When conducting experiments that involve placebos, researchers also have to consider the **placebo effect**—that is, whether people in the control group will improve because they believe that they are getting a real substance that is intended to produce a positive result. When a control group shows as much improvement as the treatment group, a researcher can conclude that the placebo effect is a significant factor in the improvements of both groups.

Surveys

CONCEPT A process that uses questionnaires or similar means to gather values for the responses from a set of participants.

EXAMPLES The decennial U.S. census mail-in form, a poll of likely voters, a Web site instant poll or "question of the day."

INTERPRETATION Surveys are either informal, open to anyone who wishes to participate; targeted, directed toward a specific group of individuals; or include people chosen at random. The type of survey affects how the data collected can be used and interpreted.

1.4 Sampling Concepts

Sampling

CONCEPT The process by which members of a population are selected for a sample.

EXAMPLES Choosing every fifth voter who leaves a polling place to interview, drawing playing cards randomly from a deck, polling every tenth visitor who views a certain Web site today.

INTERPRETATION The method by which sampling occurs, the identification of all items in a population, and the techniques used to select individual observations all affect sampling.

Probability Sampling

CONCEPT A sampling process that takes into consideration the chance of occurrence of each item being selected. Probability sampling increases your chances that the sample will be representative of the population.

EXAMPLES The registered voters selected to participate in a recent survey concerning their intention to vote in the next election, the patients selected to fill out a patient-satisfaction questionnaire, 100 boxes of cereal selected from a factory's production line.

INTERPRETATION You should use probability sampling whenever possible, because only this type of sampling allows you to apply inferential statistical methods to the data you collect. In contrast, you should use nonprobability sampling, in which the chance of occurrence of each item being selected is not known, to obtain rough approximations of results at low cost or for small-scale, initial, or pilot studies that will later be followed up by a more rigorous analysis. Surveys and polls that invite the public to call in or answer questions on a Web page are examples of nonprobability sampling.

Simple Random Sampling

CONCEPT The probability sampling process in which every individual or item from a population has the same chance of selection as every other individual or item. Every possible sample of a certain size has the same chance of being selected as every other sample that has that size.

EXAMPLES Selecting a playing card from a shuffled deck, generating a number by throwing a pair of perfect dice, or using a statistical device such as a table of random numbers.

INTERPRETATION Simple random sampling forms the basis for other random sampling techniques. The word random in the phrase *random sampling* may confuse you if you think that random implies the unexpected or the

unanticipated, as the word often does in everyday usage (as in random acts of kindness). However, in statistics, *random* implies no repeating patterns—that is, in a given sequence, a given pattern is equally likely (or unlikely) as another. From this sense of equal chance (and not unexpected or unanticipated) comes the term *random sampling*.

Frame

CONCEPT The list of all items in the population from which samples will be selected.

EXAMPLES Voter registration lists, municipal real estate records, customer or human resource databases, directories.

INTERPRETATION Frames influence the results of an analysis, and using two different frames can lead to different conclusions. You should always be careful to make sure your frame completely represents a population; otherwise any sample selected will be biased, and the results generated by analyses of that sample will be inaccurate.

1.5 Sample Selection Methods

Proper sampling can be done with or without replacement.

Sampling With Replacement

CONCEPT A sampling method in which each selected item is returned to the frame from which it was selected so that it has the same probability of being selected again.

EXAMPLE Selecting entries from a fishbowl and returning each entry to the fishbowl after it is drawn.

Sampling Without Replacement

CONCEPT A sampling method in which each selected item is not returned to the frame from which it was selected. Using this technique, an item can be selected no more than one time.

EXAMPLES Selecting numbers in state lottery games, selecting cards from a deck of cards during games of chance such as Blackjack.

INTERPRETATION Sampling without replacement means that an item can be selected no more than one time. You should choose sampling without

replacement over sampling with replacement, because statisticians generally consider the former to produce more desirable samples.

Other, more complex, sampling methods are also used in survey sampling. In a stratified sample, the items in the frame are first subdivided into separate subpopulations, or **strata**, and a simple random sample is conducted within each of the strata. In a **cluster sample**, the items in the frame are divided into several *clusters* so that each cluster is representative of the entire population. A random sampling of clusters is then taken, and all the items in each selected cluster or a sample from each cluster are then studied.

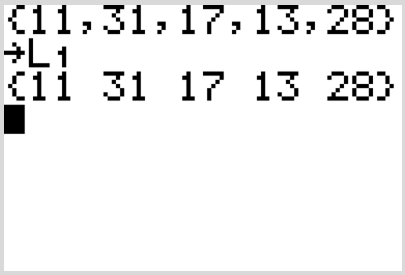
calculator keys

Entering Data

You can choose one of two ways to enter data values for a variable.

When entering one short list of values for a single variable:

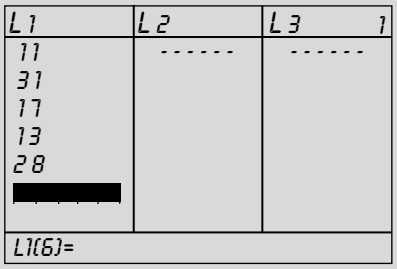
Press **[2nd]**[(] and enter the values separated by commas. (Press [,] to type a comma.) When you finish entering values, press **[2nd]**[)][**STO** ▸] and enter the name of the variable in which to store the values. For example, to store values in variable L1, press **[2nd]**[**1**]. Press **[ENTER]** to complete the data entry. Your calculator will display the values separated by spaces and your screen will look similar to this:



When entering the values for several variables, or many values for a single variable:

Press **[STAT]**. Select **1:Edit** and press **[ENTER]**. Use the cursor keys to move the cursor to the column of the variable for which you want to enter data. (If you have just cleared your RAM memory, the cursor will be in the column for variable L1.) Enter the first data value and press **[ENTER]**. Repeat until all values have been entered. Your screen will look similar to this:

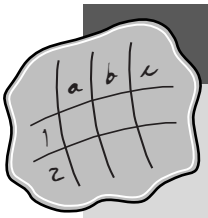
(*continues*)



You can enter the data values for a second variable by using the cursor keys to move to the column of another variable. To delete values previously entered into a column, move the cursor to the name of variable and press **[CLEAR][ENTER]**.

When you have finished entering all values, press **[2nd][MODE]** to quit and return to the main display.

If you have a connection cable and the TI Connect software, you can also enter values for a variable using the TI Data Editor application.



spreadsheet solution

Entering Data

Select **File → New**. Select **Blank Workbook** from the task pane. (If using an older version of Excel, select the **Workbook** icon in the New dialog box.) Click cell **A1**. Enter a name for variable in this cell and press **[ENTER]**. Type the first data value and press **[ENTER]**. Repeat until all values have been entered. Notice that every time you press **[ENTER]** the worksheet entry automatically advances down one row.

When you have finished entering data, select **File → Save As**, type a filename, and click the **Save** button to save your data.

One-Minute Summary

To understand statistics, you must first master the basic vocabulary presented in this chapter. You have also been introduced to data collection, the various sources of data, sampling methods, as well as the types of variables used in statistical analysis. The remaining chapters of this book focus on four important reasons for learning statistics:

- To present and describe information (Chapters 2 and 3)
- To draw conclusions about populations based only on sample results (Chapters 4 through 9)
- To obtain reliable forecasts (Chapter 10)
- To improve processes (Chapter 11)

Test Yourself

1. The portion of the population that is selected for analysis is called:
   (a) a sample
   (b) a frame
   (c) a parameter
   (d) a statistic
2. A summary measure that is computed from only a sample of the population is called:
   (a) a parameter
   (b) a population
   (c) a discrete variable
   (d) a statistic
3. The height of an individual is an example of a:
   (a) discrete variable
   (b) continuous variable
   (c) categorical variable
   (d) constant
4. The body style of an automobile (sedan, coupe, wagon, etc.) is an example of a:
   (a) discrete variable
   (b) continuous variable
   (c) categorical variable
   (d) constant
5. The number of credit cards in a person's wallet is an example of a:
   (a) discrete variable
   (b) continuous variable
   (c) categorical variable
   (d) constant
6. Statistical inference occurs when you:
   (a) compute descriptive statistics from a sample
   (b) take a complete census of a population
   (c) present a graph of data
   (d) take the results of a sample and draw conclusions about a population

7. The human resources director of a large corporation wants to develop a dental benefits package and decides to select 100 employees from a list of all 5,000 workers in order to study their preferences for the various components of a potential package. All the employees in the corporation constitute the _____.
   (a) sample
   (b) population
   (c) statistic
   (d) parameter
8. The human resources director of a large corporation wants to develop a dental benefits package and decides to select 100 employees from a list of all 5,000 workers in order to study their preferences for the various components of a potential package. The 100 employees who will participate in this study constitute the _____.
   (a) sample
   (b) population
   (c) statistic
   (d) parameter
9. Those methods involving the collection, presentation, and characterization of a set of data in order to properly describe the various features of that set of data are called:
   (a) statistical inference
   (b) the scientific method
   (c) sampling
   (d) descriptive statistics
10. Based on the results of a poll of 500 registered voters, the conclusion that the Republican candidate for U.S. president will win the upcoming election is an example of:
   (a) inferential statistics
   (b) descriptive statistics
   (c) a parameter
   (d) a statistic
11. A summary measure that is computed to describe a characteristic of an entire population is called:
   (a) a parameter
   (b) a population
   (c) a discrete variable
   (d) a statistic

12. You were working on a project to look at the value of the American dollar as compared to the English pound. You accessed an Internet site where you obtained this information for the past 50 years. Which method of data collection were you using?
    (a) Published sources
    (b) Experimentation
    (c) Surveying
13. Which of the following is a discrete variable?
    (a) The favorite flavor of ice cream of students at your local elementary school
    (b) The time it takes for a certain student to walk to your local elementary school
    (c) The distance between the home of a certain student and the local elementary school
    (d) The number of teachers employed at your local elementary school
14. Which of the following is a continuous variable?
    (a) The eye color of children eating at a fast-food chain
    (b) The number of employees of a branch of a fast-food chain
    (c) The temperature at which a hamburger is cooked at a branch of a fast-food chain
    (d) The number of hamburgers sold in a day at a branch of a fast-food chain
15. The number of cars that arrive per hour at a parking lot is an example of:
    (a) a categorical variable
    (b) a discrete variable
    (c) a continuous variable
    (d) a statistic
16. The possible responses to the question "How long have you been living at your current residence?" are values from a continuous variable.
    (a) True
    (b) False
17. The possible responses to the question "How many times in the past three months have you visited a museum?" are values from a discrete variable.
    (a) True
    (b) False
18. An insurance company evaluates many variables about a person before deciding on an appropriate rate for automobile insurance. The number of accidents a person has had in the past three years is an example of a _____ variable.

19. An insurance company evaluates many variables about a person before deciding on an appropriate rate for automobile insurance. The distance a person drives in a day is an example of a _____ variable.
20. An insurance company evaluates many variables about a person before deciding on an appropriate rate for automobile insurance. A person's marital status is an example of a _____ variable.

Answers to Test Yourself Questions

1. a
2. d
3. b
4. c
5. a
6. d
7. b
8. a
9. d
10. a
11. a
12. a
13. d
14. c
15. b
16. a
17. a
18. discrete
19. continuous
20. categorical

References

1. Berenson, M. L., D. M. Levine, and T. C. Krehbiel. *Basic Business Statistics: Concepts and Applications, Ninth Edition*. Upper Saddle River, NJ: Prentice Hall, 2004.
2. Cochran, W. G. *Sampling Techniques, Third Edition*. New York: Wiley, 1977.

3. Gitlow, H. S., and D. M. Levine. *Six Sigma for Green Belts and Champions*. Upper Saddle River, NJ: Financial Times – Prentice Hall, 2005.
4. Levine, D. M., T. C. Krehbiel, and M. L. Berenson. *Business Statistics: A First Course, Third Edition*. Upper Saddle River, NJ: Prentice Hall, 2003.
5. Levine, D. M., D. Stephan, T. C. Krehbiel, and M. L. Berenson. *Statistics for Managers Using Microsoft Excel, Fourth Edition*. Upper Saddle River, NJ: Prentice Hall, 2005.
6. Levine, D. M., P. P. Ramsey, and R. K. Smidt, *Applied Statistics for Engineers and Scientists Using Microsoft Excel and Minitab*. Upper Saddle River, NJ: Prentice Hall, 2001.
7. Microsoft Excel 2002. Redmond, WA: Microsoft Corporation, 2001.
8. Sincich, T., D. M. Levine, and D. Stephan, *Practical Statistics by Example Using Microsoft Excel and Minitab, Second Edition*. Upper Saddle River, NJ: Prentice Hall, 2002.