Chapter 1
Introduction

Section 1.1

1.1 1) **Statistics** refers to numerical facts such as the age of a student or the income of a family.
2) Statistics refers to the field or discipline of study. Statistics is a group of methods used to collect, analyze, present, and interpret data and to make decisions.

1.3 **a.** This is an example of inferential statistics because a poll was taken using a sample of adults and based on the results, conclusions are inferred with a certain margin of error.

b. This is an example of descriptive statistics because information was gathered and tabulated, but no inference was made to a larger population.

Section 1.2

1.5 With reference to this table, we have the following definitions:
- Member: Each cause of death included in the table
- Variable: The number of deaths
- Measurement: The number of deaths from each cause of death
- Data set: Collection of the number of deaths from each cause of death listed in the table

Section 1.3

1.7 **a.** A **quantitative variable** is a variable that can be measured numerically.

b. A variable that cannot assume a numeric value but can be classified into two or more nonnumeric categories is called a **qualitative variable**.

c. A **discrete variable** is a variable whose values are countable.

d. A variable that can assume any numerical value over a certain interval or intervals is called a **continuous variable**.

e. Data collected on a quantitative variable is called **quantitative data**.

f. **Qualitative data** is data collected on a qualitative variable.

1.9 **a.** Continuous **b.** Continuous

c. Continuous **d.** Discrete

Section 1.4

1.11 Data collected on different elements at the same point in time or for the same period of time are called **cross-section data**. Total sales for the 2019 Christmas season at 10 stores in a particular mall is an example of cross-section data.

Data collected on the same element for the same variable at different points in time or for different periods of time are called **time-series data**. Total sales for one particular store for the Christmas season for the years 2009 to 2019 is an example of time-series data.

Section 1.5

1.13 A **population** is the collection of all elements whose characteristics are being studied.
A **sample** is a portion of the population selected for study.
A **representative sample** is a sample that represents the characteristics of the population as closely as possible.
Sampling with replacement refers to a sampling procedure in which the item selected at each selection is put back in the population before the next item is drawn.
Sampling without replacement is a sampling procedure in which the item selected at each selection is not replaced in the population.

1.15 A **census** is a survey that includes every member of the population.
A survey based on a portion of the population is called a **sample survey**.
A sample survey is preferred over a census for the following reasons:

1) Conducting a census is very expensive because the size of the population is often very large.
2) Conducting a census is very time consuming.
3) In many cases it is impossible to identify each element of the target population.

1.17 **a.** A sampling technique under which each sample of the same size has the same probability of being selected is called a **simple random sample**.

b. In **systematic random sampling**, we first randomly select one member from the first k units. Then, every k^{th} member, starting with the first selected member, is included in the sample.

c. In a **stratified random sample**, we first divide the population into subpopulations which are called *strata*. Then, one sample is selected from each of these strata. The collection of all samples from all strata gives the stratified random sample.

d. In **cluster sampling**, the whole population is divided into (geographical) groups called *clusters*. Each cluster is representative of the population. Then, a random sample of clusters is selected. Finally, a random sample of elements of each of the selected clusters is selected.

1.19 **a.** Population **b.** Sample

c. Population **d.** Population

e. Sample

1.21 **a.** This is a random sample since it is selected randomly from a complete list of students at the university. Thus, each student in the population has an equal chance of being included in the sample.

b. This is a simple random sample since the software package would give each sample of 150 students an equal chance of being selected.

c. There should be no systematic error since the sampling frame is the entire population, and the use of the software would give each sample of 150 students an equal chance of being selected.

1.23 This is a quota sample since it is composed of 58% males and 42% females, the same proportions found in the population of 1000 employees. It is also a nonrandom and convenience sample because men and women were selected by interviewers as they wished.

1.25 The survey is subject to voluntary response error since it receives responses from only those companies that are willing to take the trouble to complete the questionnaire and mail it in. These respondents may not be representative of all major companies. It also suffers from nonresponse error because many companies did not respond as well as from selection error.

1.27 Since the sample includes only people from one borough of New York City, it is not likely to be representative of the entire city. Therefore, the researcher is not justified in applying the result to the entire city of New York.

Section 1.6

1.29 When an experimenter controls the (random) assignments of elements to different treatment groups, the study is an **experiment**. For an **observational study**, the assignment of elements to different treatments is voluntary, and the experimenter simply observes the results of the study.

1.31 **a.** This is a designed experiment since the doctors controlled the assignment of people to the treatment and control groups.

b. The experiment is not double-blind since the doctors knew who was given aspirin and who was given the placebo.

1.33 This is an observational study since the researchers relied on volunteers to form the treatment and control groups.

1.35 The conclusion is unjustified. The families volunteered; they were not randomly selected from the population of all families on welfare; thus they may not be representative of the entire population.

Section 1.7

1.37

m	f	f^2	mf	m^2f
5	12	144	60	300
10	8	64	80	800
17	6	36	102	1734
20	16	256	320	6400
25	4	16	100	2500
$m = 77$	$\sum f = 46$	$\sum f^2 = 516$	$\sum mf = 662$	$\sum m^2f = 11{,}734$

a. $\sum m = 77$ **b.** $\sum f^2 = 516$ **c.** $\sum mf = 662$ **d.** $\sum m^2f = 11{,}734$

1.39 **a.** $\sum x = 387 + 414 + 404 + 396 + 410 + 422 + 414 = 2847$ miles

b. $(\sum x)^2 = (2847)^2 = 8{,}105{,}409$

c. $\sum x^2 = (387)^2 + (414)^2 + (404)^2 + (396)^2 + (410)^2 + (422)^2 + (414)^2 = 1{,}158{,}777$

Supplementary Exercises

1.41 The data set contains annual revenue for Shake Shack for each year from 2012 to 2018 so it is time-series data.

1.43 **a.** This is an example of sampling without replacement because once a patient is selected, he/she will not be replaced before the next patient is selected.

b. This is an example of sampling with replacement because both times the selection is made from the same group of professors.

1.45

x	y	x^2	xy	x^2y	y^2
7	5	49	35	245	25
11	15	121	165	1815	225
8	7	64	56	448	49
4	10	16	40	160	100
14	9	196	126	1764	81
28	19	784	532	14,896	361
$\sum x = 72$	$\sum y = 65$	$\sum x^2 = 1230$	$\sum xy = 954$	$\sum x^2y = 19{,}328$	$\sum y^2 = 841$

a. $\sum y = 65$ **b.** $\sum x^2 = 1230$ **c.** $\sum xy = 954$ **d.** $\sum x^2y = 19{,}328$ **e.** $\sum y^2 = 841$

1.47 **a.** This is an observational study because participants are volunteers and they decided how much meat to consume. Thus, the treatment is not controlled by the experimenters.

b. Because this is an observational study, no cause-and-effect relationship between meat consumption and cholesterol level may be inferred. The effect of meat consumption on cholesterol level may be confounded with other variables such as other dietary habits, amount of exercise, and other features of the participants' lifestyles.

1.49 **a.** Since the patients were randomly selected from the population of all people suffering from compulsive behavior and were randomly assigned to treatment and control groups, the two groups should be comparable and representative of the entire population. The patients did not know whether or not they were getting the treatment, so any improvement in their condition should be due to the medicine and not merely to the power of suggestion. Thus, the conclusion is justified.

b. This is a designed experiment since the doctors controlled the assignment of patients to the treatment and control groups.

c. The study is not double-blind since the doctors knew who received the medicine.

Advanced Exercises

1.51 **a.** We would expect \$81,200 to be an invalid estimate of the current mean annual income for all 5432 alumni because only 1240 of the 5432 alumni answered the income question. These 1240 are unlikely to be representative of the entire group of 5432.

b. The following types of bias are likely to be present:
Nonresponse error: Alumni with low incomes may be ashamed to respond. Thus, the 1240 who actually returned their questionnaires and answered the income question would tend to have higher than average incomes.
Response error: Some of those who answered the income question may give a value that is higher than their actual income in order to appear more successful.

c. We would expect the estimate of \$81,200 to be above the current mean annual income of all 5432 alumni, for the given reasons in part b.

Self-Review Test

1. b

3. **a.** Sample without replacement

b. Sample with replacement

5. A sample drawn in such a way that each element of the population has some chance of being included in the sample is called a **random sample**.

A sample in which some members of the population may have no chance of being selected is called a **nonrandom sample**.

A sample that contains the characteristics of the population as closely as possible is called a **representative sample**.

A **convenience sample** is a sample in which the most accessible members of the population are selected.

A **judgment sample** is a sample in which members of a population are selected based on the judgment and prior knowledge of an expert.

A **quota sample** is a sample selected in such a way that each group or subpopulation is represented in the sample in exactly the same proportion as in the target population.

When we select an element from the population and put it back in the population before we select the next element, it is called a **sample with replacement**.

When the selected element is not replaced in the population and each time we select an item, the size of the population is reduced by one element, it is called a **sample without replacement**.

7. A sampling technique under which each sample of the same size has the same probability of being selected is called a **simple random sample**.

In **systematic random sampling**, we first randomly select one member from the first k units. Then, every k^{th} member, starting with the first selected member, is included in the sample.

In a **stratified random sample**, we first divide the population into subpopulations which are called *strata*. Then, one sample is selected from each of these strata. The collection of all samples from all strata gives the stratified random sample.

In **cluster sampling**, the whole population is divided into (geographical) groups called *clusters*. Each cluster is representative of the population. Then, a random sample of clusters is selected. Finally, a random sample of elements of each of the selected clusters is selected.

9. With reference to this table, we have the following definitions:

- Member: Each student included in the table
- Variable: Midterm test score
- Measurement: The midterm test score of a student
- Data Set: Collection of the midterm test scores of the students listed in the table

11.

x	y	x^2	xy	x^2y
3	28.4	9	85.2	255.6
7	17.2	49	120.4	842.8
5	21.6	25	108	540
9	13.9	81	125.1	1125.9
12	6.3	144	75.6	907.2
8	16.8	64	134.4	1075.2
10	9.4	100	94	940
$\sum x = 54$	$\sum y = 113.6$	$\sum x^2 = 472$	$\sum xy = 742.7$	$\sum x^2y = 5686.7$

a. $\sum x = 54$

b. $\sum y = 113.6$

c. $\sum x^2 = 472$

d. $\sum xy = 742.7$

e. $\sum x^2 y = 5686.7$

f. $(\sum y)^2 = 113.6^2 = 12{,}904.96$

13. **a.** No, this method is not likely to produce a random sample.

b. The following types of biases are likely to be present:
Voluntary Response Error: Only readers that have a strong opinion and are willing to pay $1 to respond will do so.
Selection Error: Not all members of the population are included; only those who actually read that newspaper may participate
Response Error: A group may have a financial interest in the casino and place many calls in order to influence the outcome of the poll.

15. observational study