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.

1.2 **Descriptive statistics** consists of methods that help us organize, display, and describe data using tables, graphs, and summary measures. **Inferential statistics** consists of methods that use sample results to help make decisions or predictions about a population.

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

1.4 An **element** is a specific subject or object about which the information is collected. A **variable** is a characteristic under study that assumes different values for different elements. An **observation** is the value of a variable for a single element. A **data set** is a collection of observations on one or more variables.

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

1.6 **a.** Number of deaths

b. Eight

c. Eight (diseases)

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.

1.8 **a.** Quantitative **b.** Quantitative

c. Qualitative **d.** Qualitative

e. Quantitative

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

c. not applicable **d.** not applicable

1.10 **a.** The qualitative variables are: do they own a house, have they taken a vacation during the past year, are they happy with their financial situation

b. The quantitative variables are: age of oldest person in family, number of family members, number of males in the family, number of females in the family, income of family, and amount of monthly mortgage or rent

c. The discrete variables are number of family members, number of males in the family, number of females in the family, income of family, and amount of monthly mortgage or rent

d. The only continuous variable is: age of oldest family member

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 2011 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 2005 to 2011 is an example of time-series data.

1.12 **a.** Time-series data

b. Time-series data

c. Cross-section data

d. Cross-section data

Section 1.5

1.13 A **population** is the collection of all elements whose characteristics are being studied. A s**ample** is a portion of the population selected for study. A r**epresentative 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.

1.14 Consider a standard deck of 52 cards. Suppose we randomly select one card from the deck and record the value and suit. If we put this card back in the deck before we randomly select a second card, this is an example of **sampling with replacement**. If we lay the first card aside and randomly select the second card from the 51 cards remaining in the deck, this is an example of **sampling without replacement**.

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.

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

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

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

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

e. 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.

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.

1.18 Simple random sample

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

c. Population **d.** Population

e. Sample

1.20 **a.** This is a nonrandom sample since students in the university who were not in her statistics class had no chance of being included in the sample.

b. This is a convenience sample since students in her class were the most accessible members of the population.

c. This sample suffers from selection error. The population consists of all students at the university, but the sampling frame is limited to members of her statistics class.

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 20 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 20 students an equal chance of being selected.

1.22 **a.** This is a random sample since the sampling frame is the entire class.

b. This is a simple random sample since the software package gives each sample of 20 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 20 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 sample because men and women were selected by interviewers as they wished.

1.24 **a.** This is a non-random sample. Only readers of the magazine were able to answer the survey.

b. This sample is subject to voluntary response error, since only those who feel strongly enough about the issues to complete the questionnaire will respond. It also suffers from selection error since only the magazine's readers are included in the sampling frame.

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.

1.26 This survey is subject to response error since some parents may be reluctant to give honest answers to an interviewer's questions about sensitive family matters.

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 New York City.

Section 1.6

1.28 In a **survey**, data are collected without exercising any control over the factors that may affect the characteristics of interest or the results of a survey. In an **experiment**, the researchers exercise control over some or all of these factors.

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.

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

b. There is not enough information to determine if this is a double-blind study. We would need to know if the doctors were aware of which women were assigned to the treatment group and which were assigned to the control (placebo) group.

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.

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

b. The study is double-blind since neither the patients nor the doctors knew who was given the 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.

1.34 This is a designed experiment since the researcher selected participants randomly from the entire population of families on welfare and then controlled which families received the treatment (job training) and which did not.

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.

1.36 If the data showed that the percentage of families who got off welfare was higher in the group that received job training, the conclusion is justified. Since families were randomly assigned to treatment and control groups, the two groups should have been similar, and the difference in outcomes should be due to treatment (job training).

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
$\Sigma m = 77$	$\Sigma f = 46$	$\Sigma f^2 = 516$	$\Sigma mf = 662$	$\Sigma m^2 f = 11{,}734$

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

1.38 **a.** $\Sigma y = 216 + 184 + 35 + 92 + 144 + 175 + 11 + 57 = \$914$

b. $(\Sigma y)^2 = (914)^2 = 835{,}396$

c. $\Sigma y^2 = (216)^2 + (184)^2 + (35)^2 + (92)^2 + (144)^2 + (175)^2 + (11)^2 + (57)^2 = 144{,}932$

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

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

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

Supplementary Exercises

1.40 With reference to this table, we have the following definitions

- Member: Each company included in the table
- Variable: Revenues for 2014
- Measurement: Revenue for 2014 for a specific company
- Data Set: Collection of different 2014 revenues for the companies listed in the table

1.41 The data set contains measurements for different countries for the same period of time, so it is cross-section data.

1.42 **a.** Sample **b.** Population

c. Sample **d.** Population

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.

1.44 **a.** $\Sigma x = 8 + 14 + 3 + 7 + 10 + 5 = 47$ shoe pairs

b. $(\Sigma x)^2 = (47)^2 = 2209$

c. $\Sigma x^2 = (8)^2 + (14)^2 + (3)^2 + (7)^2 + (10)^2 + (5)^2 = 443$

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
$\Sigma x = 72$	$\Sigma y = 65$	$\Sigma x^2 = 1230$	$\Sigma xy = 954$	$\Sigma x^2y = 19{,}328$	$\Sigma y^2 = 841$

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

1.46 **a.** convenience sample

b. judgment sample

c. random sample

1.47 **a.** This is an observational study since each participant 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.

1.48 **a.** Since the study relies on volunteers, it may not be representative of the entire population of people suffering from compulsive behavior. Furthermore, the doctors used their own judgment to form the treatment and control groups. Thus, subjective factors may have influenced them, and the two groups may not be comparable. As a result, the effect of the medicine on compulsive behavior may be confounded with other variables. Therefore, the conclusion is not justified.

b. Although this study technically satisfies the criteria for a designed experiment (experimenters controlled the assignment of people to treatment groups) it suffers from the weaknesses of an observational study, as pointed out in part a.

c. The study is not double-blind since the physicians knew who received the treatment.

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.

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

b. The study is double-blind since neither patients nor doctors know who was receiving the medicine.

1.51 **a.** We would expect \$61,200 to be a biased 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 \$61,200 to be above the current mean annual income of all 5432 alumni, for the given reasons in part b.

1.52 **a.** Yes, the unvaccinated dogs are the control group.

b. No, the experiment is not double-blind. The owners and the veterinarians who examined the dogs for Lyme Disease know which dogs were vaccinated.

c. The following are potential sources of bias:
Selection error: Dogs whose owners permitted vaccination may not be comparable to other dogs. Their owners may be more concerned about keeping them away from ticks.
Not double-blind: Since owners of vaccinated dogs know their dogs were vaccinated, they may have a different degree of concern about keeping their dogs away from ticks than owners of unvaccinated dogs. Also, the veterinarians know which dogs were vaccinated, which may influence their diagnosis for a dog having symptoms resembling Lyme Disease.

d. The experiment could be improved by making it a randomized double-blind study. Select 200 dogs at random from dogs whose owners will permit vaccination. Randomly assign 100 of these dogs to the treatment group to receive the vaccine. The other 100 dogs would form the control group and would be given a placebo. The veterinarians who examined the dogs later for Lyme Disease would not be told which dogs were vaccinated, to avoid bias in diagnosis.

Self-Review Test

1. b **2.** c

3. **a.** Sample without replacement

b. Sample with replacement

4. **a.** Qualitative

b. Quantitative; continuous

c. Quantitative; discrete

d. Qualitative

5. A sample drawn in such a way that each element of the population has the same 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 **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.

6. A **sampling chance error** is the difference between the result obtained from a sample survey and the result that would have been obtained if the whole population had been included in the survey.

A **nonsampling error** occurs in the collection, recording, or tabulation of data.

A **selection bias** is the error that occurs because the sampling frame is not representative of the population.

A **nonresponse bias** is the error that occurs because many of the people included in the sample do not respond to a survey.

A **response bias** is the error that occurs when people included in the survey do not provide correct answers.

A **voluntary response bias** is the error that occurs when a survey is not conducted on a randomly selected sample, but via a questionnaire in a magazine or newspaper where people are invited to respond to that questionnaire.

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.

8. An **observational study** is one in which data is gathered and observed, but no inference is made. Also, the assignment of elements to different treatments is voluntary.

A **designed experiment** is one in which the experimenter controls the (random) assignment of elements to different treatment groups.

Randomization is the procedure in which elements are assigned to different groups at random.

A **treatment group** is the group of elements that receives a treatment.

A **control group** is the group of elements that does not receive a treatment or receives a placebo.

A **double-blind experiment** is an experiment in which neither patients nor experimenters know who is taking the real medicine and who is taking the placebo.

The **placebo effect** is when patients respond to placebos because they have confidence in their physicians and medicines.

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

10. a. $\Sigma x = 6 + 11 + 3 + 5 + 6 + 2 = 33$ types of cereal

b. $(\Sigma x)^2 = (33)^2 = 1089$

c. $\Sigma x^2 = (6)^2 + (11)^2 + (3)^2 + (5)^2 + (6)^2 + (2)^2 = 231$

11.

x	y	x^2	xy	x^2y	y^2
3	28.4	9	85.2	255.6	806.56
7	17.2	49	120.4	842.8	295.84
5	21.6	25	108	540	466.56
9	13.9	81	125.1	1125.9	193.21
12	6.3	144	75.6	907.2	39.69
8	16.8	64	134.4	1075.2	282.24
10	9.4	100	94	940	88.36
$\Sigma x = 54$	$\Sigma y = 113.6$	$\Sigma x^2 = 472$	$\Sigma xy = 742.7$	$\Sigma x^2y = 5686.7$	$\Sigma(y^2) = 2172.46$

a. $\Sigma x = 54$ **b.** $\Sigma y = 113.6$ **c.** $\Sigma x = 472$ **d.** $\Sigma xy = 742.7$ **e.** $\Sigma x^2y = 5686.7$ **f.** $(\Sigma y)^2 = 2172.46$

12. a. These 10 pigs represent a convenience sample since the first ten (easiest to catch) pigs comprise the sample. Convenience samples are nonrandom samples.

b. From part a we know these 10 pigs comprise a nonrandom sample. Therefore, they are not likely to be representative of the entire population. Faster pigs, for example, are not as likely to be included in the sample.

c. They form a convenience sample.

d. Answers will vary, but one better procedure is as follows: Assign numbers 1 through 40 to the pigs, and write the numbers 1 through 40 on separate pieces of paper, put them in a hat, mix them, and then draw 10 numbers. Pick the pigs whose numbers were drawn.

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.

14. a. This is a designed experiment since the doctors controlled the assignment of dieters to the lower sugar and control groups.

b. Yes, those who received as much as ten percent of their calories from sucrose were the control group.

c. No, this was not a double-blind experiment since both the doctors and dieters knew who was on the low sugar diet and who was not.

15. observational study

16. randomized experiment