Chapter 1
Statistics, Data, and Statistical Thinking

1.1 Statistics is a science that deals with the collection, classification, analysis, and interpretation of information or data. It is a meaningful, useful science with a broad, almost limitless scope of applications to business, government, and the physical and social sciences.

1.2 Descriptive statistics utilizes numerical and graphical methods to look for patterns, to summarize, and to present the information in a set of data. Inferential statistics utilizes sample data to make estimates, decisions, predictions, or other generalizations about a larger set of data.

1.3 The four elements of a descriptive statistics problem are:

1. The population or sample of interest. This is the collection of all the units upon which the variable is measured.
2. One or more variables that are to be investigated. These are the types of data that are to be collected.
3. Tables, graphs, or numerical summary tools. These are tools used to display the characteristic of the sample or population.
4. Identification of patterns in the data. These are conclusions drawn from what the summary tools revealed about the population or sample.

1.4 The five elements of an inferential statistical analysis are:

1. The population of interest. The population is a set of existing units.
2. One or more variables that are to be investigated. A variable is a characteristic or property of an individual population unit.
3. The sample of population units. A sample is a subset of the units of a population.
4. The inference about the population based on information contained in the sample. A statistical inference is an estimate, prediction, or generalization about a population based on information contained in a sample.
5. A measure of reliability for the inference. The reliability of an inference is how confident one is that the inference is correct.

1.5 The first major method of collecting data is from a published source. These data have already been collected by someone else and are available in a published source. The second method of collecting data is from a designed experiment. These data are collected by a researcher who exerts strict control over the experimental units in a study. These data are measured directly from the experimental units. The final method of collecting data is observational. These data are collected directly from experimental units by simply observing the experimental units in their natural environment and recording the values of the desired characteristics. The most common type of observational study is a survey.

1.6 Quantitative data are measurements that are recorded on a meaningful numerical scale. Qualitative data are measurements that are not numerical in nature; they can only be classified into one of a group of categories.

1.7 A population is a set of existing units such as people, objects, transactions, or events. A variable is a characteristic or property of an individual population unit such as height of a person, time of a reflex, amount of a transaction, etc.

1.8 A population is a set of existing units such as people, objects, transactions, or events. A sample is a subset of the units of a population.

1.9 A representative sample is a sample that exhibits characteristics similar to those possessed by the target population. A representative sample is essential if inferential statistics is to be applied. If a sample does not possess the same characteristics as the target population, then any inferences made using the sample will be unreliable.

1.10 An inference without a measure of reliability is nothing more than a guess. A measure of reliability separates statistical inference from fortune telling or guessing. Reliability gives a measure of how confident one is that the inference is correct.

1.11 A population is a set of existing units such as people, objects, transactions, or events. A process is a series of actions or operations that transform inputs to outputs. A process produces or generates output over time. Examples of processes are assembly lines, oil refineries, and stock prices.

1.12 Statistical thinking involves applying rational thought processes to critically assess data and inferences made from the data. It involves not taking all data and inferences presented at face value, but rather making sure the inferences and data are valid.

1.13 The data consisting of the classifications A, B, C, and D are qualitative. These data are nominal and thus are qualitative. After the data are input as 1, 2, 3, and 4, they are still nominal and thus qualitative. The only differences between the two data sets are the names of the categories. The numbers associated with the four groups are meaningless.

1.14 Answers will vary. First, number the elements of the population from 1 to 200,000. Using MINITAB, generate 10 numbers on the interval from 1 to 200,000, eliminating any duplicates.

The 10 numbers selected for the random sample are:

135075
89127
189226
83899
112367
191496
110021
44853
42091
198461

Elements with the above numbers are selected for the sample.

1.15 a. The experimental unit for this study is a single-family residential property in Arlington, Texas.

b. The variables measured are the sale price and the Zillow estimated value. Both of these variables are quantitative.

c. If these 2,045 properties were all the properties sold in Arlington, Texas in the past 6 months, then this would be considered the population.

d. If these 2,045 properties represent a sample, then the population would be all the single-family residential properties sold in the last 6 months in Arlington, Texas.

e. No. The real estate market across the United States varies greatly. The prices of single-family residential properties in this small area are probably not representative of all properties across the United States.

1.16 a. The experimental unit for this study is an NFL quarterback.

b. The variables measured in this study include draft position, NFL winning ratio, and QB production score. Since the draft position was put into 3 categories, it is a qualitative variable. The NFL winning ratio and the QB production score are both quantitative.

c. Since we want to project the performance of future NFL QBs, this would be an application of inferential statistics.

1.17 a. The population of interest is all citizens of the United States.

b. The variable of interest is the view of each citizen as to whether the president is doing a good or bad job. It is qualitative.

c. The sample is the 2000 individuals selected for the poll.

d. The inference of interest is to estimate the proportion of all U.S. citizens who believe the president is doing a good job.

e. The method of data collection is a survey.

f. It is not very likely that the sample will be representative of the population of all citizens of the United States. By selecting phone numbers at random, the sample will be limited to only those people who have telephones. Also, many people share the same phone number, so each person would not have an equal chance of being contacted. Another possible problem is the time of day the calls are made. If the calls are made in the evening, those people who work in the evening would not be represented.

1.18 a. High school GPA is a number usually between 0.0 and 4.0. Therefore, it is quantitative.

b. Honors/awards would have responses that name things. Therefore, it would be qualitative.

c. The scores on the SAT's are numbers between 200 and 800. Therefore, it is quantitative.

d. Gender is either male or female. Therefore, it is qualitative.

e. Parent's income is a number: $25,000, $45,000, etc. Therefore, it is quantitative.

f. Age is a number: 17, 18, etc. Therefore, it is quantitative.

1.19 I. Qualitative; the possible responses are "yes" or "no," which are non-numerical.

II. Quantitative; age is measured on a numerical scale, such as 15, 32, etc.

III. Qualitative; the possible responses are “yes” or “no,” which are non-numerical.

IV. Qualitative; the possible responses are "laser printer" or "another type of printer," which are non-numerical.

V. Qualitative; the speeds can be classified as "slower," "unchanged," or "faster," which are non-numerical.

VI. Quantitative; the number of people in a household who have used Windows 95 at least once is measured on a numerical scale, such as 0, 1, 2, etc.

1.20 a. For question 1, the data collected would be qualitative. The possible response would be “yes” or “no”.

For question 2, the data collected would be quantitative. The responses would be numbers such as 0, 1, 2, etc.

For question 3, the data collected would be qualitative. The possible responses would be “yes” or “no”.

b. The data collected from the 1,066 adults would be a sample. These adults would only be a part of all adults in the United States.

1.21 a. Whether the data collected on the chief executive officers at the 500 largest U. S. companies is a population or a sample depends on what one is interested in. If one is only interested in the information from the CEO’s of the 500 largest U.S. companies, then these data form a population. If one is interested in the information on CEO’s from all U.S. firms, then these data would form a sample.

b. 1. The industry type of the CEO’s company is a qualitative variable. The industry type is a name.

2. The CEO’s total compensation is a meaningful number. Thus, it is a quantitative variable.

3. The CEO’s total compensation over the previous five years is a meaningful number. Thus, it is a quantitative variable.

4. The number of company stock shares (millions) held is a meaningful number. Thus, it is a quantitative variable.

5. The CEO’s age is a meaningful number. Thus, it is a quantitative variable.

6. The CEO’s efficiency rating is a meaningful number. Thus, it is a quantitative variable.

1.22 a. The population of interest is the status of computer crime at all United States businesses and government agencies.

b. The method of data collection was a survey. Since not all of those who were sent a survey responded, the sample was self-selected. The results are probably not representative of the population. Usually, those who respond to surveys have very strong opinions, either positive or negative.

c. The variable of interest is whether or not the firm or agency had unauthorized use of its computer systems during the year. Since the response would be either yes or no, the variable would be qualitative.

d. If the sample was representative, we could infer that approximately 41% of all U. S. corporations and government agencies experienced unauthorized use of their computer systems during the year.

1.23 Since the data collected consist of the entire population, this would represent a descriptive study. Flaherty used the data to help describe the condition of the U.S. Treasury in 1861.

1.24 This study would be an example of inferential statistics. The researchers collected data over 2 years. Using this information, the researchers are projecting or making inferences about what will happen in the future.

1.25 a. The population of interest is all individuals who earned MBA degrees since January 1990.

b. The method of data collection was a survey.

c. This is probably not a representative sample. The sample was self-selected. Not all of those who were selected for the study responded to all four surveys. Those who did respond to all 4 surveys probably have very strong opinions, either positive or negative, which may not be representative of all of those in the population.

1.26 a. The population of interest is all CPA firms.

b. A survey was used to collect the data.

c. This sample was probably not representative. Not all of those selected to be in the sample responded. In fact, only 992 of the 23,500 people who were sent the survey responded. Generally, those who do respond to surveys have very strong opinions, either positive or negative. These may not be the opinions of all CPA firms.

d. Since the sample may not be representative, the inferences drawn in the study may not be valid.

1.27 a. Length of maximum span can take on values such as 15 feet, 50 feet, 75 feet, etc. Therefore, it is quantitative.

b. The number of vehicle lanes can take on values such as 2, 4, etc. Therefore, it is quantitative.

c. The answer to this item is "yes" or "no," which is not numeric. Therefore, it is qualitative.

d. Average daily traffic could take on values such as 150 vehicles, 3,579 vehicles, 53,295 vehicles, etc. Therefore, it is quantitative.

e. Condition can take on values "good," "fair," or "poor," which are not numeric. Therefore, it is qualitative.

f. The length of the bypass or detour could take on values such as 1 mile, 4 miles, etc. Therefore, it is quantitative.

g. Route type can take on values "interstate," U.S.," "state," "county," or "city," which are not numeric. Therefore, it is qualitative.

1.28 a. The variable of interest to the researchers is the rating of highway bridges.

b. Since the rating of a bridge can be categorized as one of three possible values, it is qualitative.

c. The data set analyzed is a population since all highway bridges in the U.S. were categorized.

d. The data were collected observationally. Each bridge was observed in its natural setting.

1.29 a. The process being studied is the distribution of pipes, valves, and fittings to the refining, chemical, and petrochemical industries by the Wallace Company of Houston.

b. The variables of interest are the speed of the deliveries, the accuracy of the invoices, and the quality of the packaging of the products.

c. The sampling plan was to monitor a subset of current customers by sending out a questionnaire twice a year and asking the customers to rate the speed of the deliveries, the accuracy of the invoices, and the quality of the packaging minutes. The sample is the total numbers of questionnaires received.

d. The Wallace Company's immediate interest is learning about the delivery process of its distribution of pipes, valves, and fittings. To do this, it is measuring the speed of deliveries, the accuracy of the invoices, and the quality of its packaging from the sample of its customers to make an inference about the delivery process to all customers. In particular, it might use the mean speed of its deliveries to the sampled customers to estimate the mean speed of its deliveries to all its customers. It might use the mean accuracy of its invoices from the sampled customers to estimate the mean accuracy of its invoices of all its customers. It might use the mean rating of the quality of its packaging from the sampled customers to estimate the mean rating of the quality of its packaging of all its customers.

e. Several factors might affect the reliability of the inferences. One factor is the set of customers selected to receive the survey. If this set is not representative of all the customers, the wrong inferences could be made. Also, the set of customers returning the surveys may not be representative of all its customers. Again, this could influence the reliability of the inferences made.

1.30 a. The population of interest would be the set of all students. The sample of interest would be the students participating in the experiment. The variable measured in this study is whether the student would spend money on repairing a very old car or not.

b. The data-collection method used was a designed experiment. The students participating in the experiment were randomly assigned to one of three emotional states and then asked a question.

c. The researcher could estimate the proportion of all students in each of the three emotional states who would spend money to repair a very old car.

d. One factor that might affect the reliability of the inference drawn is whether the students in the experiment were representative of all students. It is stated that the sample was made up of volunteer students. Chances are that these volunteer students were not representative of all students. In addition, if these students were all from the same school, they probably would not be representative of the population of students either.

1.31 a. The population of interest would be all accounting alumni of a large southwestern university.

b. Age would produce quantitative data – the responses would be numbers.

Gender would produce qualitative data – the responses would be ‘male’ or ‘female’.

Level of education would produce qualitative data – the responses could be categories such college degree, master’s degree, or PhD degree.

Income would produce quantitative data – the responses would be numbers.

Job satisfaction score would produce quantitative data. We would assume that a satisfaction score would be a number, where the higher the number, the higher the job satisfaction.

Machiavellian rating score would produce quantitative data. We would assume that a rating score

would be a number, where the higher the score, the higher the Machiavellian traits.

c. The sample is the 198 people who returned the useable questionnaires.

d. The data collection method used was a survey.

e. The inference made by the researcher is that Machiavellian behavior is not required to achieve success in the accounting profession.

f. Generally, those who respond to surveys are those with strong feelings (in either direction) toward the subject matter. Those who do not have strong feelings for the subject matter tend not to answer surveys. Those who did not respond might be those who are not real happy with their jobs or those who are not real unhappy with their jobs. Thus, we might have no idea what type of scores these people would have on the Machiavellian rating score.

1.32 a. Give each stock in the NYSE-Composite Transactions table of the Wall Street Journal a number (1 to m). Using a random number table or a computer program, select n different numbers on the interval from 1 to m. The stocks with the same numbers as the n chosen numbers will be selected for the sample.

1.33 a. The experimental units for this study are engaged couples who used a particular website.

b. There are two variables of interest – the price of the engagement ring and the level of appreciation. Price of the engagement ring is a quantitative variable because it is measured on a numerical scale. Level of appreciation is a qualitative variable. There are 7 different categories for this variable that are then assigned numbers.

c. The population of interest would be all engaged couples.

d. No, the sample is probably not representative. Only engaged couples who used a particular web site were eligible to be in the sample. Then, only those with "average" American names were invited to be in the sample.

e. Answers will vary. First, we will number the individuals from 1 to 50. Using MINITAB, 25 random numbers were generated on the interval from 1 to 50. The random numbers are:

1, 4, 5, 8, 12, 13, 17, 18, 19, 20, 22, 26, 27, 30, 31, 33, 34, 35, 38, 39, 40, 42, 43, 46, 49

The individuals who were assigned the numbers corresponding to the above numbers would be assigned to one role and the remaining individuals would be assigned to the other role.

1.34 Answers will vary. Using MINITAB, the 5 seven-digit phone numbers generated with area code 373 were:

373-639-0598
373-411-9164
373-502-7699
373-782-2719
373-930-3231

1.35 a. Some possible questions are:

1. In your opinion, why has the banking industry consolidated in the past few years? Check all that apply.

    a. Too many small banks with not enough capital.

b. A result of the Savings and Loan scandals.
c. To eliminate duplicated resources in the upper management positions.
d. To provide more efficient service to the customers.
e. To provide a more complete list of financial opportunities for the customers.
f. Other. Please list.

2. Using a scale from 1 to 5, where 1 means strongly disagree and 5 means strongly agree, indicate your agreement to the following statement: "The trend of consolidation in the banking industry will continue in the next five years."

1 strongly disagree 2 disagree 3 no opinion 4 agree 5 strongly agree

b. The population of interest is the set of all bank presidents in the United States.

c. It would be extremely difficult and costly to obtain information from all bank presidents. Thus, it would be more efficient to sample just 200 bank presidents. However, by sending the questionnaires to only 200 bank presidents, one risks getting the results from a sample which is not representative of the population. The sample must be chosen in such a way that the results will be representative of the entire population of bank presidents in order to be of any use.

1.36 a. The process being studied is the process of filling beverage cans with soft drink at CCSB's Wakefield plant.

b. The variable of interest is the amount of carbon dioxide added to each can of beverage.

c. The sampling plan was to monitor five filled cans every 15 minutes. The sample is the total number of cans selected.

d. The company's immediate interest is learning about the process of filling beverage cans with soft drink at CCSB's Wakefield plant. To do this, they are measuring the amount of carbon dioxide added to a can of beverage to make an inference about the process of filling beverage cans. In particular, they might use the mean amount of carbon dioxide added to the sampled cans of beverage to estimate the mean amount of carbon dioxide added to all the cans on the process line.

e. The technician would then be dealing with a population. The cans of beverage have already been processed. He/she is now interested in the outputs.

1.37 a. The population of interest is the set of all people in the United States over 14 years of age.

b. The variable being measured is the employment status of each person. This variable is qualitative. Each person is either employed or not.

c. The problem of interest to the Census Bureau is inferential. Based on the information contained in the sample, the Census Bureau wants to estimate the percentage of all people in the labor force who are unemployed.

1.38 Suppose we want to select 900 intersections by numbering the intersections from 1 to 500,000. We would then use a random number table or a random number generator from a software program to select 900 distinct intersection points. These would then be the sampled markets.

Now, suppose we want to select the 900 intersections by selecting a row from the 500 and a column from the 1,000. We would first number the rows from 1 to 500 and number the columns from 1 to 1,000. Using a random number generator, we would generate a sample of 900 from the 500 rows. Obviously, many rows will be selected more than once. At the same time, we use a random number generator to select 900

columns from the 1,000 columns. Again, some of the columns could be selected more than once. Placing these two sets of random numbers side-by-side, we would use the row-column combinations to select the intersections. For example, suppose the first row selected was 453 and the first column selected was 731. The first intersection selected would be row 453, column 731. This process would be continued until 900 unique intersections were selected.

1.39 Answers will vary.

a. The results as stated indicate that by eating oat bran, one can improve his/her health. However, the only way to get the stated benefit is to eat only oat bran with limited results. People may change their eating habits expecting an outcome that is almost impossible.

b. To investigate the impact of domestic violence on birth defects, one would need to collect data on all kinds of birth defects and whether the mother suffered any domestic violence or not during her pregnancy. One could use an observational study survey to collect the data.

c. Very few people are *always* happy with the way they are. However, many people are happy with themselves most of the time. One might want to ask a series of questions to measure self-esteem rather than just one. One question might ask what percent of the time the high school girl is happy with the way she is.

d. The results of the study are probably misleading because of the fact that if someone relied on a limited number of foods to feed her children it does not imply that the children are hungry. In addition, one might cut the size of a meal because the children were overweight, not because there was not enough food. One might get better information about the proportion of hungry American children by actually recording what a large, representative sample of children eat in a week.

e. A leading question gives information that seems to be true, but may not be complete. Based on the incomplete information, the respondent may come to a different decision than if the information was not provided.