Chapter

Statistics, Data, and Statistical Thinking 1

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.

1.4 The first major method of collecting data is from a published source. These data have already been collected by someone else and is 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 third method of collecting data is from a survey. These data are collected by a researcher asking a group of people one or more questions. Again, these data are collected directly from the experimental units or people. The final method of collecting data is observationally. 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.

1.6 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 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 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.

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

b. High school class rank is a number: 1st, 2nd, 3rd, etc. Therefore, it is quantitative.

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.

1.14 a. The experimental unit for this experiment is a drafted NFL quarterback.

b. Draft position is one of three categories. Therefore, it is a qualitative variable. NFL winning ratio is a number. Therefore, it is a quantitative variable. QB production score is a number. Therefore, it is a quantitative variable.

c. Because all quarterbacks drafted over a 38-year period were used, the application of this study is descriptive statistics.

1.16 a. The variable “difference between before and after sprint times” is measured in seconds. Thus, it is quantitative. The variable “improvement” is measured as one of three categories. Thus, it is qualitative.

b. The data set is a sample. It contains observations from only 14 of all high school football players.

1.18 a. The population of interest is all the students in the class. The variable of interest is the GPA of a student in the class.

b. Since GPA is measured on a numerical scale, it is quantitative.

c. Since the population of interest is all the students in the class and you obtained the GPA of every member of the class, this set of data would be a census.

d. Assuming the class had more than 10 students in it, the set of 10 GPAs would represent a sample. The set of ten students is only a subset of the entire class.

e. This average would have 100% reliability as an "estimate" of the class average, since it is the average of interest.

f. The average GPA of 10 members of the class will not necessarily be the same as the average GPA of the entire class. The reliability of the estimate will depend on how large the class is and how representative the sample is of the entire population.

g. In order for the sample to be a random sample, every member of the class must have an equal

1.20 a. Flight capability can have only 2 possible outcomes: volant or flightless. Thus, it is qualitative.

b. Habitat type can have only 3 possible outcomes: aquatic, ground terrestrial, or aerial terrestrial. Thus, it is qualitative.

c. Nesting site can have only 4 possible outcomes: ground, cavity within ground, tree, or cavity above ground. Thus, it is qualitative.

d. Nest density can have only 2 possible outcomes: high or low. Thus, it is qualitative.

e. Diet can have only 4 possible outcomes: fish, vertebrates, vegetables, or invertebrates. Thus, it is qualitative.

f. Body mass is measured in grams, a meaningful number. Thus, it is quantitative.

g. Egg length is measured in millimeters, a meaningful number. Thus, it is quantitative.

h. Extinct status can have only 3 possible outcomes: extinct, absent from island, or present. Thus, it is qualitative.

1.22 a. The population of interest to CSI is all computer security personnel at all U.S. corporations and government agencies.

b. The data collection method is a survey. A survey was sent to 5,412 firms with 351 firms responding.

c. The variable collected was whether or not the respondents admitted unauthorized use of computer systems at their firms during the year. Since the response to the questions was either "yes" or "no", this variable is qualitative.

d. In the sample 41% of the respondents admitted unauthorized use of computer systems at their firms during the year. If there is no nonresponse bias, then we can conclude that 41% of all firms would admit to unauthorized use of computer systems at their firms during the year.

1.24 The following variables would be qualitative because the response would be a category: country of operator/owner, primary use, and class of orbit. The following variables would be quantitative because the response would be a number: Longitudinal position, apogee, launch mass, usable electric power, and expected lifetime.

1.26 a. The population of interest is all senior managers at CPA firms.

b. The data collection method used is a survey.

c. Because only 992 of the 23,500 surveys sent out were returned and useable, there may be a problem with selection bias and/or nonresponse bias.

d. The validity of the inferences drawn from the study would be suspect. The inferences would only be valid if the 992 returned surveys were indeed, representative of the entire population. This is very unlikely.

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

b. The variables measured were the Zillow estimated value and the actual sale price. Both are quantitative variables.

c. If the population was described as all single-family residential properties in Arlington, Texas that sold within a given time period, then these 2,045 single-family residential properties could be the population if these were the only single-family residential property sales in Arlington, Texas in that time period.

d. The population could be all single-family residential properties sold in Arlington, Texas in a given time period and these 2,045 single-family residential properties did not include all the properties sold.

e. No. The single-family residential properties sold in Arlington, Texas probably are not similar to all single-family residential properties sold in the United States. Single-family residential properties sold in Arlington, Texas are not similar to single-family residential properties sold in places like New York City or San Francisco, California.

1.30 a. The population of interest is the set of all adults living in Tennessee. The sample of interest is the set of 575 people selected from Tennessee.

b. The data collection method used was a survey. A random-digit telephone dialing procedure was used to collect the sample. Since some people do not own phones, this would not be a random sample. Everyone in the state of Tennessee would not have an equal chance of being selected. Those without telephones would tend to be the undereducated. Thus, there could be potential biases in the data.

c. The two variables identified in this problem are the number of years of education and the insomnia status of each subject. The number of years of education is quantitative and the insomnia status is qualitative.

d. The researchers inferred that the fewer the years of education, the more likely the person was to have chronic insomnia.

1.32 a. The experimental units of this study were people who used a popular website for engaged couples.

b. The variables of interest are the engagement ring price and the level of appreciation of the recipient.

c. The population of interest were all those people on the popular website for engaged couples with "average" American names.

d. This sample of 33 respondents is probably not representative of the population. Those who decided to respond to the online survey self-selected themselves. They were not randomly selected. Generally speaking, those people who choose to respond to a survey have very strong feelings and are not representative of the entire population.

e. Answers will vary. Enter the numbers 1-50 in the first column of Minitab. Now, apply the random number generator of Minitab, requesting that 25 individuals be selected without replacement. The sample generated include individuals 1, 3, 6, 7, 9, 10, 11, 12, 13, 17, 18, 22, 24, 25, 27, 28, 30, 31, 32, 36, 37, 46, 47, 48, 50. These individuals would be placed in the gift-receiver role and the remaining individuals would be placed in the gift-giver role.

1.34 a. In Method 1, the researchers controlled which hot spots received the new program (through random assignment) and which did not. Therefore, a designed experiment was used to collect data in Method 1.

b. In Method 2, the researchers first divided the 56 hot spots into 4 groups based on the level of drug crimes. The researchers then controlled which hot spots received the new program (through random assignment) and which did not in each of the 4 groups. Therefore, a designed experiment was also used to collect data in Method 2.

c. This would be an application of inferential statistics because not all hot spots in Jersey City were used in the study. Only a sample of 56 hot spots was used.

d. Method 2 would be recommended. By creating 4 groups where the crime rate within each group is similar, we can control for a known source of variation. Within each group, we can then see how the new program compares to no program.

1.36 a. Although eating oat bran can reduce cholesterol, oat bran must be the only thing eaten. Reporting that eating oat bran is an easy and cheap way to lower your cholesterol implies that if you add oat bran to your diet, you can reduce your cholesterol, which may not be the case at all.

b. One would need to collect data on a sample of women who gave birth to babies with birth defects. Then, each woman in the study would also be asked whether she was a victim of domestic violence while pregnant or not. The data collection method for this study would be an observational study. Specifically, one would use a survey to collect the data.

c. In this study, only the results of the most positive response were reported. Not that many high school girls would 'Always' be happy with the way they were. To be more representative of those who are happy with the way they were, one should combine the results of those who responded 'Always true', 'Sort of true', and 'Sometimes true'.