Statistics, Data, and Statistical Thinking

Chapter **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 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.

1.16 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 in 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

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

1.20 a. The 500 surgical patients represent a sample. There are many more than 500 surgical patients.

b. Yes, the sample is representative. It says that the surgical patients were randomly selected.

c. The variable measures on each patient was the status of herbal or alternative medicines. These data are qualitative because each response was either "yes" or "no".

1.22 a. The population of interest is the set of all computer security personnel at all United States businesses.

b. The data collection method used was a survey. Surveys were sent to all computer security personnel at all U.S. corporations and government agencies. However, in 2006, only 616 organizations responded to the survey. There could be nonresponse bias. Often, only those subjects with strong opinions will respond to a survey. Thus, the responses may not reflect what the population as a whole thinks.

c. The variable measured in the survey is whether or not there was unauthorized use of computer systems at the firms during the year. Since the responses will be either 'Yes" or 'No', the variable is qualitative.

d. If we assume that the responses were a random sample from the population, we could infer that about 52% of all computer security personnel will admit to unauthorized use of computer systems at their firms during the year.

1.24 a. The sample is the set of 505 teenagers selected at random from all U.S. teenagers.

b. The population from which the sample was selected is the set of all teenagers in the U.S.

c. Since the sample was a random sample, it should be representative of the population.

d. The variable of interest is the topics that teenagers most want to discuss with their parents.

e. The inference is expressed as a percent of the population that want to discuss particular topics with their parents.

f. The "margin of error" is the measure of reliability. This margin of error measures the uncertainty of the inference.

1.26 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.

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

1.28 a. The population of interest is all men and women.

b. The sample of interest is the approximately 300 men and women from Gainesville, Florida, who participated in the study.

c. The study involves inferential statistics. The researcher is not particularly interested in the reponses of just those subjects who participated in the study. She is interested in generalizing her findings to all men and women.

d. One variable is measured for each of the 20 objects placed. For each variable, the 2 possible outcomes were "yes" (place of object was recalled) and "no" (place of object was not recalled). Since the outcomes "yes" and "no" are not measured on a numerical scale, the variables are qualitative.

1.30 a. The experimental units in this study are the 24 new software development projects.

b. The population from which the sample was selected is the set of all new software development projects.

c. The variable of interest in this project is the outcome of reusing previously developed software for the new software development projects. Since the outcomes could either be success or failures, the variable is qualitative.

d. In the sample, 15 of the 24 projects were judged as successfully implemented or 62.5%. This is the success rate of the sample. This would be a good estimate of the population percentage of successfully implemented projects, but it is only an estimate. If we took another sample of size 24, the percentage of successful projects would not necessarily be 62.5%.

1.32 a. The data collection method used was a survey.

b. The target population is the set of all American adults.

c. The sample was not a random sample. Thus, it may not be representative of all American adults. Many people contacted on the telephone refuse to participate in surveys.