

Statistics, Data, and Statistical Thinking

Chapter

1

- 1.2 Statistical applications may be broadly classified into two types: descriptive statistics and inferential statistics. Descriptive statistics utilizes numerical and graphical methods to look for patterns in a data set, to summarize the information revealed in it, and to present that information in a convenient form. Inferential statistics utilizes sample data to make estimates, decisions, predictions, or other generalizations about a larger set of data.
- 1.4 A population is the collection of all units that a statistician wants to study. A unit may be a person, an object, a transaction, or an event. An observational unit is a selected unit about which a statistician collects data. A sample is a subset of a population.
- 1.6 Statistical inference aims at using the information contained in a smaller sample to learn about a larger population. Any such estimation is associated with an error as the study is not made on the entire population. The measure of reliability is a statement about the degree of uncertainty associated with the inference.
- 1.8 Data for statistical analysis can be obtained in three different ways:
- The first method is from a published source, such as a book, a journal, or a newspaper. Various survey reports or published data sources, such as the Wall Street Journal, may be considered to be a good source of data for related areas. The internet now also provides a medium by which data from published sources can be readily obtained. The second method of collecting data involves conducting a designed experiment, in which the researcher exerts strict control over the units (people, objects, or things) in the study. The third method of obtaining data is through observational study. Here, the researcher observes the experimental units in their natural setting and records the variable(s) of interest. The most common type of observational study is a survey, where a statistician samples a group of people, asks one or more questions, and records the responses.
- 1.10 Any sampling design that one may make to collect data may have two different types of bias: selection bias and non-response bias. Selection bias occurs when all the experimental units in the population do not have an equal chance of being selected in the sample, i.e., when a subset of the experimental units has little or no chance of being selected in the sample. Non-response bias is a type of selection bias that results when data on all experimental units in a sample are not obtained. This happens when a selected respondent is not available, or not willing to provide an answer to the survey query.
- 1.12
- High school GPA is a number usually between 0.0 and 4.0. Therefore, it is quantitative.
 - High school class rank is a number: 1st, 2nd, 3rd, etc. Therefore, it is quantitative.
 - The scores on the SAT's are numbers between 200 and 800. Therefore, it is quantitative.
 - Gender is either male or female. Therefore, it is qualitative.
 - 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 is all the patients who visited the local diabetes center in last 30 days. The variable of interest is their age.
 - b. The variable is quantitative, as the age is represented in terms of years and months.
 - c. It is a census as the data collection involves all the individual units in the population.
 - d. This data would represent only a sample.
 - e. As it is based on the entire population, the measure of reliability will be 100%.
 - f. One can enumerate all the patients who visited over the last 30 days and then select a random sample with the help of a random number generator.
- 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 title of the books, the subject category, and the type classification are qualitative variables. The number of times a book has been borrowed, and the original price of the books are quantitative variables.
- 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. The possible reasons for the hugely wrong prediction may be that the respondents might not have been picked across different locations, and they might be from different levels of social and political backgrounds. Moreover, they might have been so chosen that some preferred results are predicted. The respondents also might not have given correct replies for the survey questions.
 - b. The survey report is unethical because it published only partial truths about the survey results, to suit their interest.
 - c. The survey has not been fair as there have been campaigns in several countries to encourage people to vote for their own country's entry causing a large number of votes to certain items, causing imbalance in the selection of samples. Secondly, there has not been a check on telephone-based voting and one person could vote multiple times making the whole survey unfair.
 - d. The reasons for this misinformation is that the average respondents are not aware of the critical facts about health. Hence, their response has been only from the surface level. To conduct such surveys, the selection of appropriate samples is necessary.