

PROBLEMS

Note: Student solutions for problems whose numbers appear within a color box are available for purchase at www.cengagebrain.com.

Level A

- 1.** The file P02_01.xlsx indicates the gender and nationality of the MBA incoming class in two successive years at the Kelley School of Business at Indiana University.
 - a. For each year, create tables of counts of gender and of nationality. Then create column charts of these counts. Do they indicate any noticeable change in the composition of the two classes?
 - b. Repeat part a for nationality, but recode this variable so that all nationalities that have counts of 1 or 2 are classified as Other.
- 2.** The file P02_02.xlsx contains information on over 200 movies that were released in 2006 and 2007.
 - a. Create two column charts of counts, one of the different genres and one of the different distributors.
 - b. Recode the Genre column so that all genres with a count of 10 or less are lumped into a category called Other. Then create a column chart of counts for this recoded variable. Repeat similarly for the Distributor variable.
- 3.** The file P02_03.xlsx contains data from a survey of 399 people regarding a government environmental policy.
 - a. Which of the variables in this data set are categorical? Which of these are nominal; which are ordinal?
 - b. For each categorical variable, create a column chart of counts.
 - c. Recode the data into a new data set, making four transformations: (1) change Gender to list "Male"

or "Female"; (2) change Children to list "No children" or "At least one child"; (3) change Salary to be categorical with categories "Less than \$40K," "Between \$40K and \$70K," "Between \$70K and \$100K," and "Greater than \$100K" (where you can treat the breakpoints however you like); and (4) change Opinion to be a numerical code from 1 to 5 for Strongly Disagree to Strongly Agree. Then create a column chart of counts for the new Salary variable.

- 4.** The file P02_04.xlsx contains salary data on all Major League Baseball players for each year from 2002 to 2011. (It is an older version of the data used for examples later in this chapter.) For any three selected years, create a table of counts of the various positions, expressed as percentages of all players for the year. Then create a column chart of these percentages for these years. Do they remain fairly constant from year to year?

Level B

- 5.** The file DJIA Monthly Close.xlsx contains monthly values of the Dow Jones Industrial Average from 1950 through July 2015. It also contains the percentage changes from month to month. (This file is used in an example later in this chapter.) Create a new column for recoding the percentage changes into six categories: Large negative ($<-3\%$), Medium negative ($<-1\%$, $\geq-3\%$), Small negative ($<0\%$, $\geq-1\%$), Small positive ($<1\%$, $\geq0\%$), Medium positive ($<3\%$, $\geq1\%$), and Large positive ($\geq3\%$). Then create a column chart of the counts of this categorical variable. Comment on its shape.

2-4 DESCRIPTIVE MEASURES FOR NUMERICAL VARIABLES

There are many ways to summarize numerical variables, both with numerical summary measures and with charts, and we discuss the most common ways in this section. But before getting into details, it is important to understand the basic goal of this section. We begin with a numerical variable such as Salary, where there is one observation for each person. Our basic goal is to learn how these salaries are distributed across people. To do this, we can ask a number of questions, including the following: (1) What are the most "typical" salaries? (2) How spread out are the salaries? (3) What are the "extreme" salaries on either end? (4) Is a chart of the salaries symmetric about some middle value, or is it skewed in one direction? (5) Does the chart of salaries have any other peculiar features besides possible skewness? In the next chapter, we explore methods for checking whether a variable such as Salary is related to *other* variables, but for now we simply want to explore the distribution of values in the Salary column.

Excel has a number of built-in tools for summarizing numerical variables, and we will discuss these. However, even better tools are available in Excel add-ins, and in this section we will introduce a very powerful add-in from Palisade Corporation called StatTools. There are two important advantages of StatTools over other statistical software. First, it works inside

FUNDAMENTAL INSIGHT

Box Plots Versus Histograms

Box plots and histograms are complementary ways of displaying the distribution of a numerical variable. Although histograms are much more

popular and are arguably more intuitive, box plots are still informative. Besides, side-by-side box plots are very useful for comparing two or more populations, as illustrated in the next chapter.

PROBLEMS

Level A

6. The file P02_06.xlsx lists the average time (in minutes) it takes citizens of 379 metropolitan areas to travel to work and back home each day.
 - a. Create a histogram of the daily commute times.
 - b. Find the most representative average daily commute time across this distribution.
 - c. Find a useful measure of the variability of these average commute times around the mean.
 - d. The empirical rule for standard deviations indicates that approximately 95% of these average travel times will fall between which two values? For this particular data set, is this empirical rule at least approximately correct?
7. The file P02_07.xlsx includes data on 204 employees at the (fictional) company Beta Technologies.
 - a. Indicate the data type for each of the six variables included in this data set.
 - b. Create a histogram of the Age variable. How would you characterize the age distribution for these employees?
 - c. What proportion of these full-time Beta employees are female?
 - d. Find appropriate summary measures for each of the numerical variables in this data set.
 - e. For the Salary variable, explain why the empirical rules for standard deviations do or do not apply.
8. The file P02_08.xlsx contains data on 500 shipments of one of the computer components that a company manufactures. Specifically, the proportion of items that are defective is listed for each shipment.
 - a. Create a histogram that will help a production manager understand the variation of the proportion of defective components in the company's shipments.
 - b. Is the mean or median the most appropriate measure of central location for this data set? Explain your reasoning.
 - c. Discuss whether the empirical rules for standard deviations apply. Can you tell, or at least make an educated guess, by looking at the shape of the histogram? Why?
9. The file P02_09.xlsx lists the times required to service 200 consecutive customers at a (fictional) fast-food restaurant.
 - a. Create a histogram of the customer service times. How would you characterize the distribution of service times?
 - b. Calculate the mean, median, and first and third quartiles of this distribution.
 - c. Which measure of central tendency, the mean or the median, is more appropriate in describing this distribution? Explain your reasoning.
 - d. Find and interpret the variance and standard deviation of these service times.
 - e. Are the empirical rules for standard deviations applicable for these service times? If not, explain why. Can you tell whether they apply, or at least make an educated guess, by looking at the shape of the histogram? Why?
10. The file P02_10.xlsx contains midterm and final exam scores for 96 students in a corporate finance course.
 - a. Create a histogram for each of the two sets of exam scores.
 - b. What are the mean and median scores on each of these exams?
 - c. Explain why the mean and median values are different for these data.
 - d. Based on your previous answers, how would you characterize this group's performance on the midterm and on the final exam?
 - e. Create a new column of differences (final exam score minus midterm score). A positive value means the student improved, and a negative value means the student did the opposite. What are the mean and median of the differences? What does a histogram of the differences indicate?
11. The file P02_11.xlsx contains data on 148 houses that were recently sold in a (fictional) suburban community. The data set includes the selling price of each house,

PROBLEMS

Level A

- 19.** The file P02_19.xlsx lists annual percentage changes in the Consumer Price Index (CPI) from 1914 through 2014. Find and interpret the first and third quartiles and the interquartile range for these annual percentage changes. Discuss whether these are meaningful summary measures for this time series data set. Suppose that the data set listed the actual CPI values, not percentage changes, for each year. Would the quartiles and interquartile range be meaningful in this case? Why or why not?
- 20.** The Consumer Confidence Index (CCI) attempts to measure people's feelings about general business conditions, employment opportunities, and their own income prospects. Monthly average values of the CCI are listed in the file P02_20.xlsx.
- Create a time series graph of the CCI values.
 - Have U.S. consumers become more or less confident through time?
 - How would you explain recent variations in the overall trend of the CCI?
- 21.** The file P02_21.xlsx contains monthly interest rates on 30-year fixed-rate mortgages in the United States from 1977 to 2009. The file also contains rates on 15-year fixed-rate mortgages from late 1991 to 2009. What conclusion(s) can you draw from a time series graph of these mortgage rates? Specifically, what has been happening to mortgage rates in general, and how does the behavior of the 30-year rates compare to the behavior of the 15-year rates?
- 22.** The file P02_22.xlsx contains annual trade balances (exports minus imports) from 1980 to 2010.
- Create a times series graph for each of the three time series in this file.
 - Characterize recent trends in the U.S. balance of trade figures using your time series graphs.
- 23.** What has happened to the total number and average size of farms in the U.S. since the middle of the 20th century? Answer this question by creating a time series graph of the data from the U.S. Department of Agriculture in the file P02_23.xlsx. Is the observed result consistent with your knowledge of the structural changes within the U.S. farming economy?
- 24.** Is educational attainment in the United States on the rise? Explore this question by creating time series graphs for each of the variables in the file P02_24.xlsx. Comment on any observed trends in the annual educational attainment of the general U.S. population over the given period.

- 25.** The monthly averages of the federal funds rate and the bank prime loan rate are listed in the file P02_25.xlsx.
- Describe the time series behavior of these two variables. Can you discern any cyclical or other patterns in the times series graphs of these key interest rates?
 - Discuss whether it would make much sense, especially to a person at the present time, to quote traditional summary measures such as means or percentiles of these series.

Level B

- 26.** In which months of the calendar year do U.S. gasoline service stations typically have their *lowest* retail sales levels? In which months of the calendar year do the service stations typically have their *highest* retail sales levels? Create time series graphs for the monthly data in the file P02_26.xlsx to respond to these two questions. There are really two series, one of actual values and one of seasonally adjusted values. The latter adjusts for any possible seasonality, such as higher values in June and lower values in January, so that any trends are more apparent.
- 27.** The file P02_27.xlsx contains monthly data for total U.S. electronic and appliance store sales. There are really two series, one of actual values and one of seasonally adjusted values. The latter adjusts for any possible seasonality, such as higher values in June and lower values in January, so that any trends are more apparent.
- Is there an observable trend in these data? That is, do the values of the series tend to increase or decrease over time?
 - Is there a seasonal pattern in these data? If so, what is the seasonal pattern?
- 28.** The file P02_28.xlsx contains total monthly U.S. retail sales data for a number of years. There are really two series, one of actual sales and one of seasonally adjusted sales. The latter adjusts for any possible seasonality, such as higher sales in December and lower sales in February, so that any trends are more apparent.
- Create a graph of both time series and comment on any observable trends, including a possible seasonal pattern, in the data. Does seasonal adjustment make a difference? How?
 - Based on your time series graph of actual sales, make a qualitative projection about the total retail sales levels for the next 12 months. Specifically, in which months of the subsequent year do you expect retail sales levels to be *highest*? In which months of the subsequent year do you expect retail sales levels to be *lowest*?

- d. Identify all full-time employees who are either (1) females between the ages of 30 and 50 (inclusive) with at least 5 years of prior work experience, at least 10 years of prior work experience at Beta, and at least 4 years of post-secondary education; or (2) males between the ages of 40 and 60 (inclusive) with at least 6 years of prior work experience, at least 12 years of prior work experience at Beta, and at least 4 years of post-secondary education.
 - e. For those employees who meet the conditions specified in part d, compare the mean salary of the females with that of the males. Also, compare the median salary of the female employees with that of the male employees.
 - f. What proportion of the full-time employees identified in part d earns less than \$50,000 per year?
35. The file P02_35.xlsx contains data from a survey of 500 randomly selected households. Use Excel filters to answer the following questions.
- a. What are the average monthly home mortgage payment, average monthly utility bill, and average total debt (excluding the home mortgage) of all homeowners residing in the southeast part of the city?
 - b. What are the average monthly home mortgage payment, average monthly utility bill, and average total debt (excluding the home mortgage) of all homeowners residing in the northwest part of the city? How do these results compare to those found in part a?
 - c. What is the average annual income of the first household wage earners who rent their home (house or apartment)? How does this compare to the average annual income of the first household wage earners who own their home?
 - d. What proportion of the surveyed households contains a single person who owns his or her home?
36. Recall that the file Supermarket Transactions.xlsx contains over 14,000 transactions made by supermarket customers over a period of approximately two years. Use Excel filters to answer the following questions.
- a. What proportion of these transactions are made by customers who are married?
 - b. What proportion of these transactions are made by customers who do not own a home?
 - c. What proportion of these transactions are made by customers who have at least one child?
 - d. What proportion of these supermarket customers are single and own a home?

Level B

37. The file P02_35.xlsx contains data from a survey of 500 randomly selected households. Use Excel filters to answer the following questions.
- a. Identify households that own their home and have a monthly home mortgage payment in the top quartile of the monthly payments for all households.
 - b. Identify households with monthly expenditures on utilities that are within two standard deviations of the mean monthly expenditure on utilities for all households.
 - c. Identify households with total indebtedness (excluding home mortgage) less than 10% of the household's primary annual income level.

2-8 CONCLUSION

The summary measures, charts, and tables discussed in this chapter are extremely useful for describing variables in data sets. The methods in this chapter (and the next chapter) are often called *exploratory* methods because they allow you to explore the characteristics of the data and at least tentatively answer interesting questions. Most of these tools have been available for many years, but with the powerful software now accessible to virtually everyone, the tools can be applied quickly and easily to gain insights. We can promise that you will be using many if not all of these tools in your job. Indeed, the knowledge you gain from these early chapters is arguably the most valuable knowledge you will gain from the book.

To help you remember which analyses are appropriate for different questions and different data types, and which tools are useful for performing the various analyses, we have created a taxonomy in the file Data Analysis Taxonomy.xlsx. (It isn't shown here because it doesn't fit nicely on the printed page.) Feel free to refer back to the diagram in this file as you apply the tools in this chapter and the next chapter.