**Chapter 2 – Organizing and Summarizing Data**

**OUTLINE**

* 1. Organizing Qualitative Data
  2. Organizing Quantitative Data: The Popular Displays
  3. Additional Displays of Quantitative Data
  4. Graphical Misrepresentations of Data

**Putting It Together**

Chapter 1 discussed how to identify the research objective and collect data. We learned that data can be obtained from either observational studies or designed experiments. When data are obtained, they are referred to as **raw data**.

The purpose of this chapter is to learn how to organize raw data into a meaningful form so that we can understand what the data are telling us. The first step in determining how to organize raw data is to determine whether the data is qualitative or quantitative.

**Section 2.1 Organizing Qualitative Data**

**Objectives**

1. Organize Qualitative Data in Tables
2. Construct Bar Graphs
3. Construct Pie Charts

***Objective 1: Organize Qualitative Data in Tables***

Objective 1, Page 1

What is used to list each category of data and the number of occurrences for each category of data? **When**[**qualitative data**](https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx#xln-lb-lnk_obj2_1_bc0736fe-cbc0-ca31-5c28-041c93316f8b)**are collected, we often first determine the number of occurrences within each category.**

**A frequency distribution lists each category of data and the number of occurrences for each category of data.**

Objective 1, Page 2

**Example 1 *Organizing Qualitative Data into a Frequency Distribution***

A physical therapist wants to determine types of rehabilitation required by her patients. To do so, she obtains a simple random sample of 30 of her patients and records the body part requiring rehabilitation. (See Table 1.) Construct a frequency distribution of location of injury.

**Table 1**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Back | Back | Hand | Wrist | Back | Back |
| Groin | Elbow | Back | Back | Back | Groin |
| Shoulder | Shoulder | Hip | Knee | Hip | Shoulder |
| Neck | Knee | Knee | Shoulder | Shoulder | Neck |
| Back | Back | Back | Back | Knee | Back |

Data fr **STATCRUNCH**

**In this example, we'll learn to organize qualitative data**

**into a frequency distribution.**

A physical therapist wants to determine

types of rehabilitation required by her patients.

To do so, she obtains a sample 30 of her patients

and records the body part requiring rehabilitation.

Construct a frequency distribution

for the following data.

Here are the data.

Let's go ahead and open up StatCrunch.

Here we are in StatCrunch.

I've typed the data in the first column labeled Body Part.

And we'll press Stat, Tables, Frequency.

Next, we'll select the column containing the data-- Body

Part.

And we just want the frequency.

So click on Frequency.

And click Compute.

And there's our frequency distribution.

Let's go over the steps for using StatCrunch.

Type the data, one per line, in one column.

Press the Stat button.

And from Tables, select Frequency.

Select the column containing the data.

Select Frequency.

And click Compute.

And you'll have your result.

STATCRUNCH RESULTS

om Krystal Catton, student at Joliet Junior College

OBJECTIVE 1, PAGE 3

In any frequency distribution, it is a good idea to add up the frequency column. What should the total be equal to? In any frequency distribution, it is a good idea to add up the frequency column to make sure that it equals the number of observations.

In [Example 1](https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx#xln-lb-lnk_obj2_3_3099b78d-80df-c737-015f-998ccd0207ed), the frequency column totals to 30 as it should because there are 30 body parts (observations)

Objective 1, Page 6

Define the relative frequency of a category. STATCRUNCH RESULTS

**Frequency table results for DATA:**

Count = 30

| **DATA** | **Frequency** |
| --- | --- |
| Back | 12 |
| Knee | 5 |
| Shoulder | 4 |
| Wrist | 2 |
| Hip | 2 |
| Hand | 2 |
| Neck | 1 |
| Groin | 1 |
| Elbow | 1 |

What is a relative frequency distribution? Question 1. StatCrunch

**Frequency table results for var1:**

Count = 50

| **var1** | **Frequency** |
| --- | --- |
| O | 24 |
| A | 16 |
| AB | 6 |
| B | 4 |

Ob 1, Part 6.

Often, we want to know the *relative frequency* of the categories rather than the frequency.

**DEFINITION**

**The *relative frequency* is the proportion (or percent) of observations within a category and is found using the formula**

Relative frequency=Frequency /Sum of all frequencies

**A *relative frequency distribution* lists each category of data together with the relative frequency.**

Objective 1, Page 7

**Example 2 *Constructing a Relative Frequency Distribution of Qualitative Data***

Using the summarized data in Table 2, construct a relative frequency distribution.

**Table 2**

| **Body Part** | **Frequency** |
| --- | --- |
| Back | 12 |
| Hand | 2 |
| Wrist | 2 |
| Groin | 1 |
| Elbow | 1 |
| Shoulder | 4 |
| Hip | 2 |
| Knee | 5 |
| Neck | 1 |

Objective 1, Page 8

When working with a relative frequency distribution, what should the total of the relative frequencies be equal to? Why?

5. The Sample Size. Because it is measuring frequency of a piece of data within a full sample size.

OR THAT THE **SUM IS EQUAL TO EXACTLY ONE.**

It is a good idea to add up the relative frequencies to be sure they sum to 1. In fraction form, the sum should be exactly 1. In decimal form, the sum may differ slightly from 1 due to rounding.

***Objective 2: Construct Bar Graphs***

Objective 2, Page 1

Explain how a bar graph is constructed. What do the heights of each rectangle represent? A **bar graph** is constructed by labeling each category of data on either the horizontal or vertical axis and the frequency or relative frequency of the category on the other axis. Rectangles of equal width are drawn for each category.

1. ***A bar graph* is constructed by labeling each category of data on either the horizontal or vertical axis and the frequency or relative frequency of the category on the other axis. The height of each rectangle represents the category's *frequency or relative frequency***

Objective 2, Page 2

**Example 3 *Constructing a Frequency and Relative Frequency Bar Graph***

Use the data summarized in Table 3 to construct a frequency bar graph and relative frequency bar graph.

**Table 3**

| **Body Part** | **Frequency** | **Relative Frequency** |
| --- | --- | --- |
| Back | 12 | 0.4 |
| Hand | 2 | 0.0667 |
| Wrist | 2 | 0.0667 |
| Groin | 1 | 0.0333 |
| Elbow | 1 | 0.0333 |
| Shoulder | 4 | 0.1333 |
| Hip | 2 | 0.0667 |
| Knee | 5 | 0.1667 |
| Neck | 1 | 0.0333 |

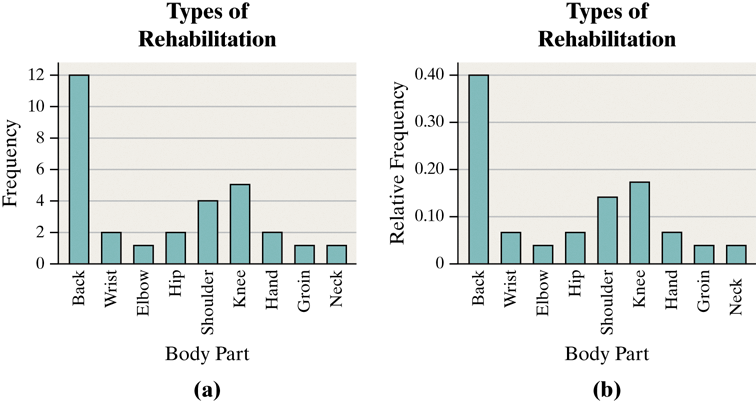
Both bar graphs are labeled “Types of Rehabilitation”. The bar graph with the x-axis labeled “Body Part (a)” and the y-axis labeled “Frequency” is summarized below:

|  |  |
| --- | --- |
| Body part | Frequency |
| Back | 12 |
| Wrist | 2 |
| Elbow | 1 |
| Hip | 2 |
| Shoulder | 4 |
| Knee | 5 |
| Hand | 2 |
| Groin | 1 |
| Neck | 1 |

Bar graph with the x-axis labeled “Body Part (b)” and the y-axis labeled “Relative Frequency” is summarized below:

|  |  |
| --- | --- |
| Body part | Relative frequency |
| Back | 0.40 |
| Wrist | 0.07 |
| Elbow | 0.04 |
| Hip | 0.07 |
| Shoulder | 0.14 |
| Knee | 0.18 |
| Hand | 0.07 |
| Groin | 0.04 |
| Neck | 0.04 |

All data are approximate.



Objective 2, Page 4

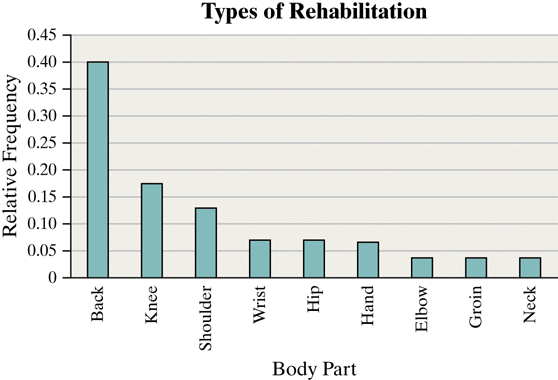
What is a Pareto chart? In bar graphs, the order of the categories does not usually matter. However, bar graphs that have categories arranged in decreasing order of frequency help prioritize information for decision-making purposes.

**DEFINITION**

A **Pareto chart** is a bar graph whose bars are drawn in decreasing order of frequency or relative frequency

Objective 2, Page 5

1. Explain why it is best to use relative frequencies when comparing data sets.



Objective 2, Page 6

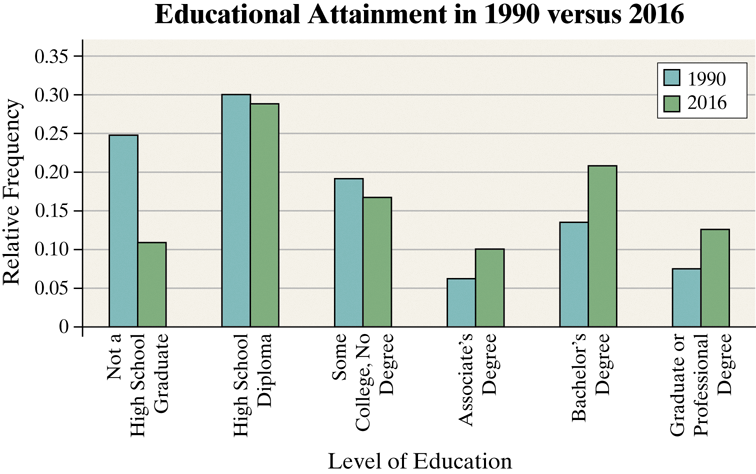
**Example 4 *Comparing Two Data Sets***

The frequency data in Table 4 represent the educational attainment (level of education) in 1990 and 2016 of adults 25 years and older who are U.S. residents. The data are in thousands. So 39,344 represents 39,344,000.

**Table 4**

| **Educational Attainment** | **1990** | **2016** |
| --- | --- | --- |
| Not a high school graduate | 39,344 | 23,453 |
| High school diploma | 47,643 | 62,002 |
| Some college, no degree | 29,780 | 36,003 |
| Associate’s degree | 9792 | 21,657 |
| Bachelor’s degree | 20,833 | 44,778 |
| Graduate or professional degree | 11,478 | 27,122 |
| Totals | 158,870 | 215,015 |

1. Draw a side-by-side relative frequency bar graph of the data.



OBJECTIVE 2, PAGE 6 (CONTINUED)

The side-by-side relative frequency bar graph shows additional information that was not easy to identify from the frequency table in Table 4. Comment on the interesting features of the side-by-side relative frequency bar graph.

The frequency data in Table 4 represent the educational attainment (level of education) in 1990 and 2016 of adults 25 years and older who are U.S. residents. The data are in thousands. So 39,344 represents 39,344,000

|  |  |  |
| --- | --- | --- |
| **Educational Attainment** | 1990 | 2016 |
| Not a high school graduate | 39,344 | 23,453 |
| High school diploma | 47,643 | 62,002 |
| Some college, no degree | 29,780 | 36,003 |
| Associate's degree | 9,792 | 21,657 |
| Bachelor's degree | 20,833 | 44,778 |
| Graduate or professional degree | 11,478 | 27,122 |
| **Totals** | 158,870 | 215,015 |

Solution

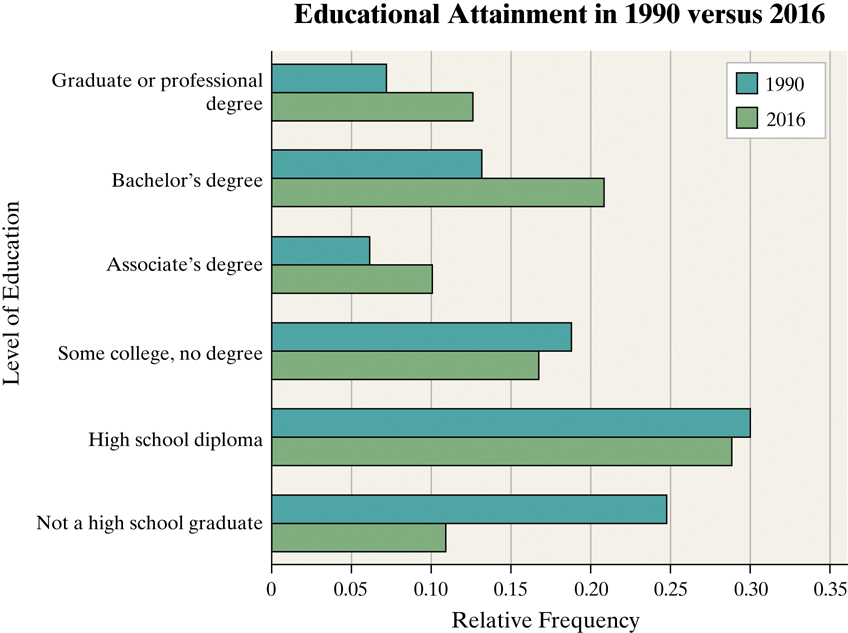
The side-by-side bar graph illustrates that the proportion of Americans 25 years and older who had some college but no degree was higher in 1990. This information is not clear from the frequency table (Table 4) because the total population sizes are different. The increase in the number of Americans who did not complete a degree is due partly to the increases in the size of the population. In addition, the number of individuals with a Bachelor's degree more than doubled (20,833 to 44,778). However, from the side-by-side bar graph, we see that the proportion of Americans 25 years and older who had a Bachelor's degree did not double. It is also clear that adult Americans have more education in 2016 than in 1990 with a much higher percentage of the population having at least a bachelor's degree (20.33% in 1990 versus 33.44% in 2016).

Objective 2, Page 8

#### **Explain when it would be preferable to use horizontal bars rather than vertical bars when constructing a bar graph. 9.**

#### **Question Horizontal Bars**

Bar graphs may also be drawn with horizontal bars. Horizontal bars are preferable when category names are lengthy. For example, Figure 4 uses horizontal bars to display the same data as in [Figure 3](https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx#xln-lb-lnk_obj3_8_0e0bae58-d7c3-047a-d700-cba51c15430d).

1. 

***Objective 3: Construct Pie Charts***

Objective 3, Page 1

1. What is a pie chart?

OBJECTIVE 3, PAGE 2

**Example 5 *Constructing a Pie Chart***

*Pie charts* are typically used to present the relative frequency of qualitative data. In most cases, the data are nominal, but ordinal data can also be displayed in a pie chart.

**DEFINITION**

A **pie chart** is a circle divided into sectors. Each sector represents a category of data. The area of each sector is proportional to the frequency of the category.

Problem

The frequency data presented in Table 6 represent the educational attainment of U.S. residents 25 years and older in 2016. The data are in thousands so 23,453 represents 23,453,000. Construct a pie chart of the data.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | **TABLE 6** | | | **Educational Attainment** | **2016** | | Not a high school graduate | 23,453 | | High school diploma | 62,002 | | Some college, no degree | 36,003 | | Associate's degree | 21,657 | | Bachelor's degree | 44,778 | | Graduate or professional degree | 27,122 | | **Total** | 215,015 | | Data from U.S. Census Bureau | | |

**CONSTRUCT PIE CHART**

In this example, we'll learn to construct a pie

chart using StatCrunch.

The following data represent the educational attainment

of residents of the United States 25 years

and older in 2016 based on data from the U.S. Census Bureau.

The data are in thousands.

Construct a pie chart of the data.

Here are the data.

Let's go to StatCrunch.

I've already typed the data into StatCrunch.

**1. The educational attainment categories**

**are in the first column,**

**2.the frequencies for 2016**

**in the second column.**

**3.To make the pie chart, I click Graph, Pie Chart With Summary.**

We use With Summary whenever we have

a total frequency for each category

rather than the raw data, like high school

graduate, bachelor's degree, bachelor's degree,

some college, et cetera.

**4.The categories are in the first column.**

**5. I label that educational attainment.**

**6. The counts are in 2016.**

And I want to make sure that I display the count

and the percent of total.

**7. Order by-- I'm going to stick with the worksheet order**

**8. and press Compute.**

And there is our pie chart.

**Now we'll go over the StatCrunch steps, type**

**the categories in one column,**

**and the counts in another.**

**Press the Graph button,**

**and from Pie Chart, select With Summary.**

**Select the columns containing the categories and counts,**

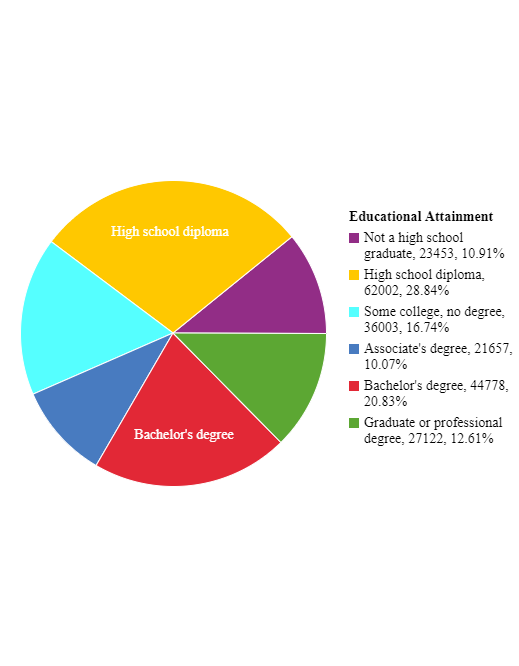
**and click Next.**

**Display the percent of total for each category, click Compute,**

**and you'll have your pie chart.**

1. Which graph, a pie chart or a bar graph, is better at comparing one category to another category?

Which graph, a pie chart or a bar graph, is better at compar

****

11.Which graph, a pie chart or a bar graph, is better at comparing one category to another category?

Bar

12.Which graph, a pie chart or a bar graph, is better at comparing one category to the whole?

**Pie**

In this video, we'll talk about when

should a bar graph or pie chart be used.

Pie charts are useful for showing

the division of all possible values

of a qualitative variable into its parts.

However, because angles are often

hard to judge in pie charts, they

are not as useful in comparing two specific values

of the qualitative variable.

Instead, the emphasis is on comparing the part

to the whole.

Bar graphs are useful when we want

to compare the different parts, not necessarily

the parts to the whole.

For example, to get the big picture regarding

educational attainment, a pie chart is a good visual summary.

Here we can tell what proportion of the total

is represented by adults who had a bachelor's degree.

However, it's difficult to compare the adults who

got a bachelor's degree to the adults who got a high school

diploma only.

To compare bachelor's degrees to high school diplomas,

a bar graph is a good visual summary.

Here we can compare the height of each bar.

And we can determine how much taller the high school diploma

bar is than the bachelor's degree bar.

Since bars are easier to draw and compare,

some practitioners forego pie charts

in favor of Pareto charts when comparing parts to the whole.

**Section 2.2 Organizing Quantitative Data: The Popular Displays**

***Objective 1: Organize Discrete Data in Tables***

Objective 1, Page 1 **This section deals with summarizing**

**quantitative data.**

When summarizing quantitative data,

What do we use to create the classes when the number of distinct data values of a discrete variable is small? **first determine whether the data are discrete or continuous.**

If the data are discrete with relatively few values

of the variable, then the categories of the data,

called classes, will be the observation,

just like it was for qualitative data.

Objective 1, Page 2

**Example 1 *Constructing Frequency and Relative Frequency Distributions from Discrete Data***

The manager of a Wendy’s® fast-food restaurant wants to know the typical number of customers who arrive during the lunch hour. The data represent the number of customers who arrive at Wendy’s for 40 randomly selected 15-minute intervals of time during lunch. Construct a frequency and relative frequency distribution.

**Number of Arrivals at Wendy’s**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 7 | 6 | 6 | 6 | 4 | 6 | 2 | 6 |
| 5 | 6 | 6 | 11 | 4 | 5 | 7 | 6 |
| 2 | 7 | 1 | 2 | 4 | 8 | 2 | 6 |
| 6 | 5 | 5 | 3 | 7 | 5 | 4 | 6 |
| 2 | 2 | 9 | 7 | 5 | 9 | 8 | 5 |

**However, if the data are discrete**

**but with many different values of the variable,**

or if the data are continuous, then the categories

of the data, the classes, must be

created using intervals of numbers, such as 10 to 19,

20 to 29, and so on.

We first present the techniques for organizing

discrete quantitative data when there

are relatively few different values, and then proceed to organize in continuous quantitative data.

OBJECTIVE 1 Organize Discrete Data in Tables

Use the values of the discrete variable to create the classes when the number of distinct data values is small. The approach to summarizing the data is similar to that of constructing frequency or relative frequency distributions from qualitative data where the categories of data are determined by the actual observations.

Problem

The manager of a Wendy's® fast-food restaurant wants to know the typical number of customers who arrive during the lunch hour. The data in Table 8 represent the number of customers who arrive at Wendy's for 40 randomly selected 15-minute intervals of time during lunch. For example, during one 15-minute interval, seven customers arrived. Construct a frequency and relative frequency distribution.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| | **TABLE 8** | | | | | | | | | --- | --- | --- | --- | --- | --- | --- | --- | | **Number of Arrivals at Wendy's** | | | | | | | | | 7 | 6 | 6 | 6 | 4 | 6 | 2 | 6 | | 5 | 6 | 6 | 11 | 4 | 5 | 7 | 6 | | 2 | 7 | 1 | 2 | 4 | 8 | 2 | 6 | | 6 | 5 | 5 | 3 | 7 | 5 | 4 | 6 | | 2 | 2 | 9 | 7 | 5 | 9 | 8 | 5 | |  |

**Frequency and Relative Frequency Distributions from Discrete Data**

1. Enter the raw data into the spreadsheet. Name the column variable.
2. Select **Stat**, highlight **Tables**, and select **Frequency**.
3. Click on the variable you want to summarize. Click the Type of table you want. If you want both Frequency and Relative Frequency, highlight Frequency; then press Ctrl (or Command on an Apple) and select Relative frequency. Click Compute!.

|  |  |  |
| --- | --- | --- |
| 6 | 11 | 0.275 |
| 5 | 7 | 0.175 |
| 2 | 6 | 0.15 |
| 7 | 5 | 0.125 |
| 4 | 4 | 0.1 |
| 9 | 2 | 0.05 |
| 8 | 2 | 0.05 |
| 11 | 1 | 0.025 |
| 3 | 1 | 0.025 |
| 1 | 1 | 0.025 |

2.2 Organizing Quantitative Data: The Popular Displays

Objective 2, Page 1

Explain how a histogram is constructed?

The *histogram,* a graph used to present quantitative data, is similar to the bar graph.

**DEFINITION**

A **histogram** is constructed by drawing rectangles for each class of data. The height of each rectangle is the frequency or relative frequency of the class. The width of each rectangle is the same, and the rectangles touch each other

Objective 2, Page 2

**Example 2 *Drawing a Histogram of Discrete Data***

Construct a frequency histogram and a relative frequency histogram using the data in Table 9. Recall that this table summarizes the data for the number of customers who arrive at Wendy's for 40 randomly selected 15-minute intervals of time during lunch.

**Table 9**

| **Number of Customers** | **Frequency** | **Relative Frequency** |
| --- | --- | --- |
| 1 | 1 | 0.25 |
| 2 | 6 | 0.15 |
| 3 | 1 | 0.025 |
| 4 | 4 | 0.1 |
| 5 | 7 | 0.175 |
| 6 | 11 | 0.275 |
| 7 | 5 | 0.125 |
| 8 | 2 | 0.05 |
| 9 | 2 | 0.05 |
| 10 | 0 | 0.0 |
| 11 | 1 | 0.025 |

OBJECTIVE 2, PAGE 2 (CONTINUED)

Technology Step-By-Step

**Drawing a Histogram**

1. If necessary, enter the raw data into the spreadsheet. Name the column variable.
2. Select **Graph** and highlight **Histogram**.
3. Click on the variable you want to summarize. Choose the type of histogram (frequency or relative frequency). You have the option of choosing a lower class limit for the first class by entering a value in the cell marked "Bins: Start at:". You have the option of choosing a class width by entering a value in the cell marked "Bins: Width:". Enter labels for the X-axis and Y-axis. Enter a title for the graph. Click Compute!.

OBJECTIVE 3 Organize Continuous Data in Tables

When a data set consists of a large number of different discrete data values or when a data set consists of continuous data, create classes by using intervals of numbers.

Table 10 is a typical frequency distribution created from continuous data. The data represent the number of U.S. residents, ages 25 to 74, who had a bachelor's degree or higher in 2016.

| **TABLE 10** | |
| --- | --- |
| **Age** | **Number (in thousands)** |
| 25–34 | 16,206 |
| 35–44 | 15,102 |
| 45–54 | 14,373 |
| 55–64 | 12,865 |
| 65–74 | 8,775 |
| Data from U.S. Census Bureau | |

**3.lower class limit** (the smallest value within the class)

**4. upper class limit** (the largest value within the class)

5.The **class width** is the difference between consecutive lower class limits

6. Notice that the classes in Table 10 **do not overlap**. This is necessary to avoid confusion as to which class a data value belongs. Notice also that the class widths are equal for all classes

1. Notice that the data are categorized, or grouped, by intervals of numbers. Each interval represents a class. For example, the first class is 25- to 34-year-old U.S. residents who have a bachelor's degree or higher. We read this interval as follows: “The number of U.S. residents, ages 25 to 34, with a bachelor's degree or higher was 16,206,000 in 2016.” There are five classes in the table, each with a **lower class limit** (the smallest value within the class) and an **upper class limit** (the largest value within the class). The lower class limit for the first class in Table 10 is 25; the upper class limit is 34. The **class width** is the difference between consecutive lower class limits. In Table 10, the class width is 35−25=10. The data in Table 10 are continuous. So the class 25−34 actually represents 25−34.999…, or 25 up to every value less than 35.
2. Notice that the classes in Table 10 do not overlap. This is necessary to avoid confusion as to which class a data value belongs. Notice also that the class widths are equal for all classes.

**One exception to the requirement of equal class widths occurs in open-ended tables**. A table is **open-ended** if the **first class has no lower class limit or the last class has no upper class limit**. The data in Table 11 represent the number of births to unmarried mothers in 2015 in the United States. The last class in the table, “40 and older,” is open-ended.

| **TABLE 11** | |
| --- | --- |
| **Age** | **Number of Births (in thousands)** |
| 15–19 | 204 |
| 20–24 | 561 |
| 25–29 | 435 |
| 30–34 | 252 |
| 35–39 | 117 |
| 40 and older | 30 |
| Data from National Vital Statistics Report, Vol. 66, No. 1 | |

 Organize Continuous Data in Tables

2.2 Organizing Quantitative Data: The Popular Displays

OBJECTIVE 3 Organize Continuous Data in Tables

When a data set consists of a large number of different discrete data values or when a data set consists of continuous data, create classes by using intervals of numbers.

Table 10 is a typical frequency distribution created from continuous data. The data represent the number of U.S. residents, ages 25 to 74, who had a bachelor's degree or higher in 2016.

| **TABLE 10** | |
| --- | --- |
| **Age** | **Number (in thousands)** |
| 25–34 | 16,206 |
| 35–44 | 15,102 |
| 45–54 | 14,373 |
| 55–64 | 12,865 |
| 65–74 | 8,775 |
| Data from U.S. Census Bureau | |

Notice that the data are categorized, or grouped, by intervals of numbers. Each interval represents a class. For example, the first class is 25- to 34-year-old U.S. residents who have a bachelor's degree or higher. We read this interval as follows: “The number of U.S. residents, ages 25 to 34, with a bachelor's degree or higher was 16,206,000 in 2016.” There are five classes in the table, each with a **lower class limit** (the smallest value within the class) and an **upper class limit** (the largest value within the class). The lower class limit for the first class in Table 10 is 25; the upper class limit is 34. The **class width** is the difference between consecutive lower class limits. In Table 10, the class width is 35−25=10. The data in Table 10 are continuous. So the class 25−34 actually represents 25−34.999…, or 25 up to every value less than 35.

Notice that the classes in Table 10 do not overlap. This is necessary to avoid confusion as to which class a data value belongs. Notice also that the class widths are equal for all classes.

2.2 Organizing Quantitative Data: The Popular Displays

One exception to the requirement of equal class widths occurs in open-ended tables. A table is **open-ended** if the first class has no lower class limit or the last class has no upper class limit. The data in Table 11 represent the number of births to unmarried mothers in 2015 in the United States. The last class in the table, “40 and older,” is open-ended.

| **TABLE 11** | |
| --- | --- |
| **Age** | **Number of Births (in thousands)** |
| 15–19 | 204 |
| 20–24 | 561 |
| 25–29 | 435 |
| 30–34 | 252 |
| 35–39 | 117 |
| 40 and older | 30 |
| Data from National Vital Statistics Report, Vol. 66, No. 1 | |

2.2 Organizing Quantitative Data: The Popular Displays

EXAMPLE 3 Organizing Continuous Data into a Frequency and Relative Frequency Distribution

Problem

Suppose you are considering investing in a Roth IRA. You collect the data in Table 12, which represent the five-year rate of return (in percent, adjusted for sales charges) for a simple random sample of 40 large-blend mutual funds. Construct a frequency and relative frequency distribution of the data.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| | **TABLE 12** | | | | | | | | | --- | --- | --- | --- | --- | --- | --- | --- | | **Five-Year Rate of Return of Mutual Funds (in percent)** | | | | | | | | | 10.94 | 14.60 | 12.80 | 16.00 | 11.93 | 15.68 | 9.03 | 13.40 | | 10.53 | 13.98 | 13.86 | 12.36 | 13.54 | 9.94 | 13.93 | 13.63 | | 14.12 | 14.88 | 14.77 | 13.13 | 8.28 | 19.43 | 12.98 | 13.16 | | 12.26 | 14.20 | 14.80 | 13.26 | 13.67 |  |  |  | |  |  |  |  |  |  |  |  | | Data from Morningstar.com | | | | | | | | |

**Frequency and Relative Frequency Tables-Organize Continuous Data**

1. If necessary, enter the raw data into the spreadsheet. Name the column variable.
2. Select **Data** and highlight **Bin.**
3. Click the variable you want to summarize. Click the “Use fixed width bins” radio button. Enter the lower class limit of the first class in the “Bins: Start at:” cell. Enter the class width in the “Bins: Width:” cell. Leave the ”Include left endpoint” radio button selected. Click Compute!.
4. Select **Stat** and highlight **Tables,** then **Frequency.**
5. Click the Bin(column name) variable. Under Type:, select Frequency and Relative Frequency. Click Compute!

### Frequency table results for Bin(Return):

1. Count = 40

| **Bin(Return)** | **Frequency** | **Relative Frequency** |
| --- | --- | --- |
| 8 to 9 | 2 | 0.05 |
| 9 to 10 | 2 | 0.05 |
| 10 to 11 | 4 | 0.1 |
| 11 to 12 | 1 | 0.025 |
| 12 to 13 | 6 | 0.15 |
| 13 to 14 | 13 | 0.325 |
| 14 to 15 | 7 | 0.175 |
| 15 to 16 | 3 | 0.075 |
| 16 to 17 | 1 | 0.025 |
| 19 to 20 | 1 | 0.025 |

***Objective 4: Construct Histograms of Continuous Data***

Objective 4, Page 2

**Example 4 *Drawing a Histogram of Continuous Data***

Construct a frequency and relative frequency histogram of the five-year rate of return data discussed in Example 3.

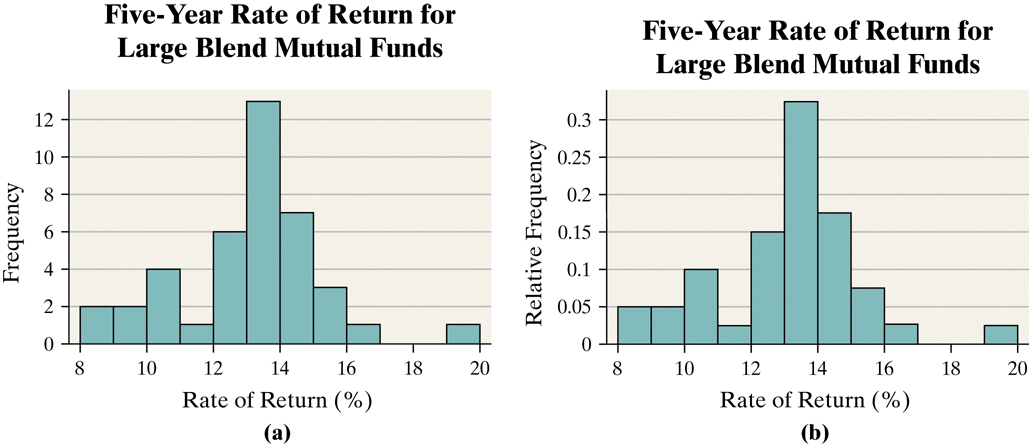
**Table 12 Five-Year Rate of Return of Mutual Funds (in percent)**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 10.94 | 14.60 | 12.80 | 16.00 | 11.93 | 15.68 | 09.03 | 13.40 |
| 10.53 | 13.98 | 13.86 | 12.36 | 13.54 | 09.94 | 13.93 | 13.63 |
| 14.12 | 14.88 | 14.77 | 13.13 | 08.28 | 19.43 | 12.98 | 13.16 |
| 12.26 | 14.20 | 14.80 | 13.26 | 13.67 | 10.08 | 14.86 | 8.71 |
| 12.17 | 10.26 | 15.22 | 13.36 | 13.55 | 13.90 | 15.64 | 12.80 |

Data from [Morningstar.com](http://morningstar.com/)

**Draw a Histogram**

1. If necessary, enter the raw data into the spreadsheet. Name the column variable.
2. Select **Graph** and highlight **Histogram**.
3. Click on the variable you want to summarize. Choose the type of histogram (frequency or relative frequency). You have the option of choosing a lower class limit for the first class by entering a value in the cell marked "Bins: Start at:". You have the option of choosing a class width by entering a value in the cell marked "Bins: Width:". Enter labels for the X-axis and Y-axis. Enter a title for the graph. Click Compute!.



Objective 4, Page 7 Objective 4, Page 4

There is no one correct frequency distribution for a particular set of data. However, some frequency distributions better illustrate patterns within the data than others. So constructing frequency distributions is somewhat of an art form. Use the distribution that seems to provide the best overall summary of the data.

Objective 4, Page 5

*Answer the following after using the applet in Activity 1: Choosing Class Width.*

**Constructing Histograms Is Somewhat of an Art Form**

In Examples 3 and 4, the choices of the lower class limit of the first class and the class width were rather arbitrary. Although formulas and procedures exist for creating frequency distributions from raw data, they do not necessarily provide better summaries.

There is no one correct frequency distribution for a particular set of data. However, some frequency distributions better illustrate patterns within the data than others. So constructing frequency distributions is somewhat of an art form. Use the distribution that seems to provide the best overall summary of the data.

Next, you will use an applet to explore how changing the class width and the lower class limit of the first class affects the appearance of a histogram. As you use the applet, remember: The goal is to design a distribution that is best for revealing the patterns within the data

Part A. 6 classes

Part B. 3 classes

Part C. 10 classes

Part D. 3 classes

What happens to the number of classes as the bin width increases? They shrink as the width increases.

1. The number of classes in a frequency distribution is typically between what two numbers?

5 - 20

1. Explain how to choose the lower class limit of the first class in a frequency distribution.

Choose the smallest observation in the data set or a convenient number slightly smaller than the smallest observation in the data set. For example, in [Table 12](https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx#xln-lb-lnk_obj5_7_ca0e9b2c-0d81-43b1-9f0f-b703e2ae3e3a), the smallest observation in 8.28. A convenient lower class limit of the first class is 8.

1. Once you decide on the number of classes, explain how to determine the class width.

**Guidelines for Determining the Lower Class Limit of the First Class and Class Width**

**Choosing the Lower Class Limit of the First Class**

Choose the smallest observation in the data set or a convenient number slightly smaller than the smallest observation in the data set. For example, in [Table 12](https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx#xln-lb-lnk_obj5_7_ca0e9b2c-0d81-43b1-9f0f-b703e2ae3e3a), the smallest observation in 8.28. A convenient lower class limit of the first class is 8.

**Determining the Class Width**

* Decide on the number of classes. Generally, there should be between 5 and 20 classes. The smaller the data set, the fewer the classes. For example, we might think ten classes would be a good choice for the data in Table 12.
* Determine the class width by computing

Class width≈largest data value − smallest data valuenumber of classes

* Round the value to a convenient number. For example, using the data in Table 12, we obtain class width ≈19.43−8.2810=1.115. Round this down to 1 because this is an easy number to work with.

Rounding up may result in fewer classes than were originally intended, while rounding down may result in more class than originally intended.

***Objective 5: Draw Dot Plots***

Objective 5, Page 1

1. Explain how to draw a dot plot. We draw a **dot plot** by placing each observation horizontally in increasing order and placing a dot above the observation each time it is observed

Objective 5, Page 2

**Example 5 *Drawing a Dot Plot***

Draw a dot plot for the data from Table 8.

**Table 8 Number of Arrivals at Wendy’s**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 7 | 6 | 6 | 6 | 4 | 6 | 2 | 6 |
| 5 | 6 | 6 | 11 | 4 | 5 | 7 | 6 |
| 2 | 7 | 1 | 2 | 4 | 8 | 2 | 6 |
| 6 | 5 | 5 | 3 | 7 | 5 | 4 | 6 |
| 2 | 2 | 9 | 7 | 5 | 9 | 8 | 5 |

**Drawing a Dot Plot**

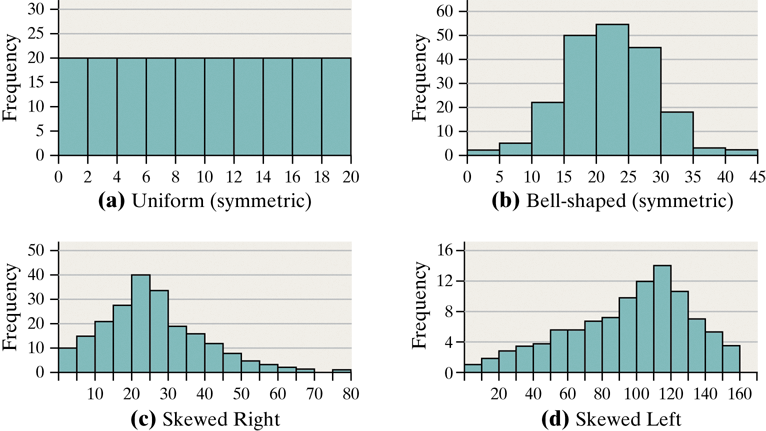
1. If necessary, enter the raw data into the spreadsheet. Name the column variable.
2. Select **Graph** and highlight **Dotplot.**
3. Click on the variable you want to summarize. Enter labels for the X-axis and Y-axis. Enter a title for the graph. Click Compute!.

***Objective 6: Identify the Shape of a Distribution***

One way that a variable is described is through the shape of its distribution. Distribution shapes are typically classified as *symmetric, skewed left,* or *skewed right.* Figure 9 displays various histograms and the shape of the distribution.

Figures 9(a) and (b) show symmetric distributions. They are symmetric because if we split the histogram down the middle, the right and left sides are mirror images. Figure 9(a) is a **uniform distribution** because the frequency of each value of the variable is evenly spread across the values of the variable. Figure 9(b) displays a **bell-shaped distribution** because the highest frequency occurs in the middle and frequencies tail off to the left and right of the middle. The distribution in Figure 9(c) is **skewed right**. Notice that the tail to the right of the peak is longer than the tail to the left of the peak. Finally, Figure 9(d) illustrates a distribution that is **skewed left** because the tail to the left of the peak is longer than the tail to the right of the peak.

It is important to recognize that data will not always exhibit behavior that perfectly matches the any of the shapes in Figure 9. To identify the shape of a distribution, some flexibility is required. In addition, people may disagree on the shape because identifying shape is subjective. Therefore, you should always support the shape of the distribution.



Objective 6, Page 1

1. Draw an example of a uniform distribution.
2. Draw an example of a bell-shaped distribution.
3. Draw an example of a distribution that is skewed right.

Objective 6, Page 1 (continued)

1. Draw an example of a distribution that is skewed left.

Objective 6, Page 2

**Example 6 *Identifying the Shape of a Distribution***

Figure 10 displays the histogram obtained in Example 4 for the five-year rate of return for large-blended mutual funds. Describe the shape of the distribution.

**Section 2.3 Additional Displays of Quantitative Data**

***Objective 1: Draw Stem-and-Leaf Plots***

Objective 1, Page 1

In a stem-and-leaf plot, how are the stem and leaf identified?

In a stem-and-leaf plot (or *stem plot*), use the digits to the left of the rightmost digit to form the **stem**. Each rightmost digit forms a **leaf**.

For example, a data value of 147 would have 14 as the stem and 7 as the leaf.

STATCRUNCH

**Drawing Stem-and-Leaf Plots**

1. If necessary, enter the raw data into the spreadsheet. Name the column variable.
2. Select **Graph** and highlight **Stem and Leaf.**
3. **Select column data**
4. **Outlier data none**
5. Click on the variable you want to summarize. Select None for outlier trimming. You also have the option of selecting the leaf unit from the drop-down menu. Click Compute!.

**ANSWER** **BELOW**   
Bizarre

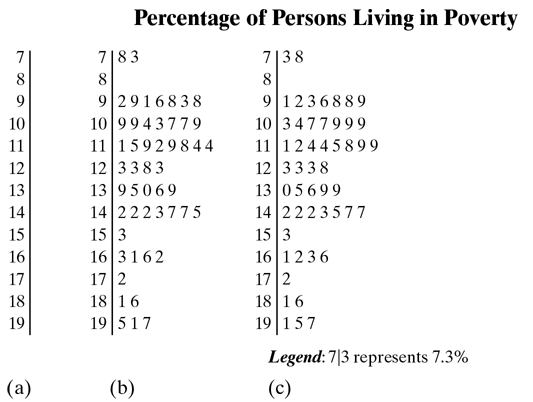
Decimal point is 1 digit(s) to the right of the colon.  
Leaf unit = 1

0 : 78999

1 : 00000011111111122222223344444444

1 : 555566677899

2 : 00

Page 2

Page 3

**Step 1.**Treat the integer portion of the number as the stem and the decimal portion as the leaf. For example, the stem of Alabama will be 16 and the leaf will be 3. The stem of 16 will include all data from 16.0 to 16.9.

2. Write the stems vertically in ascending order and then draw a vertical line to the right of the stems.

3. Write the leaves corresponding to the stem.

4. Within each stem, rearrange the leaves in ascending order. Title the plot and include a legend to indicate what the values represent.

Page 4

1. List an advantage that a stem-and-leaf plot has over frequency distributions and histograms. One advantage of the stem-and-leaf plot over frequency distributions and histograms is that the raw data can be retrieved from the stem-and-leaf plot. So, from a stem-and-leaf plot we can determine the maximum observation. We cannot learn this information immediately from a histogram. Refer to Figure 12, which shows a histogram of the poverty data drawn in StatCrunch. We can see that the largest observation is between 19 and 19.9, but we don't know that the largest value is 19.7
2. Under what conditions do stem-and-leaf plots lose their usefulness? On the other hand, stem-and-leaf plots lose their usefulness when data sets are large or consist of a large range of values.
3. when constructing a stem-and-leaf plot, under what conditions is it advisable to use split stems? Answer: **When The data appear rather bunched**

***Objective 2: Construct Frequency Polygons***

Objective 2, Page 1

1. Explain how to construct a frequency polygon. A **frequency polygon** is a graph that uses points, connected by line segments, to represent the frequencies for the classes. It is constructed by plotting a point above each **class midpoint** (the sum of consecutive lower class limits divided by 2) on a horizontal axis at a height equal to the frequency of the class.

**Construct Frequency and Relative Frequency Polygons**

1. Enter class midpoints in var1 and frequency or relative frequency in var2. Title the columns. Be sure to enter a class midpoint one class below the first class midpoint with a frequency or relative frequency of 0. Also, enter a class midpoint one class above last class midpoint with a frequency or relative frequency of 0.
2. Select **Graph** and highlight **Scatter Plot**.
3. The class midpoints are the X column. The frequency or relative frequency is the Y column. Highlight “Points and Lines” in the Display window. Label the axes and title the graph. Click Compute!.
4. Problem
5. Draw a frequency polygon of the five-year rate of return data summarized in [Table 16](https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx#xln-lb-lnk_obj2_2_d2f5d9fd-1c26-eb86-9e16-355ecdcec804).
6. **Video Solution**

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |

1. **Technology Step-By-Step**
2. [](https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx#xln-lb-lnk_obj2_2_790fc62a-e355-d60f-e892-1f03561a2d05)
3. Approach
4. Begin by calculating the class midpoints of each class. The class midpoint is found by adding consecutive lower class limits and dividing the result by 2. Plot points above each class midpoint at a height equal to the frequency of the class. Next, draw line segments connecting the points. Draw two additional line segments connecting each end of the graph with the horizontal axis. Remember to label your axes and title your graph.
5. Solution
6. The class midpoints of each class are shown in Table 16. Now plot points with the class midpoints as the x-coordinates and the frequencies as the y-coordinates. Connect these points with line segments. Then determine the midpoint of the class preceding the first class (7.5) and the midpoint of the class after the last class (20.5). Finally, connect each end of the graph with the horizontal axis at (7.5,0) and (20.5,0), respectively, to create Figure 14

***Objective 3: Create Cumulative Frequency and Relative Frequency Distributions***

Objective 3, Page 1

1. What does a cumulative frequency distribution display? A **cumulative frequency** distribution displays the aggregate frequency of the category. In other words, it displays the total number of observations less than or equal to the upper class limit of the class.
2. What does a cumulative relative frequency distribution display? A **cumulative relative frequency** distribution displays the proportion (or percentage) of observations less than or equal to the upper class limit of the class.

Explain how to find the cumulative frequency for the fifth class in a cumulative frequency distri Approach

For the cumulative frequency distribution, determine the total number of observations less than or equal to each class. For the cumulative relative frequency distribution, determine the proportion of observations less than or equal to each class.

Solution

Table 17 displays the cumulative frequency and cumulative relative frequency of the data summarized from Table 13. Table 17 shows that 38 of the 40 mutual funds had five-year rates of return of 15.99% or less. The cumulative relative frequency distribution is shown in the fifth column. We see that 95% of the mutual funds had a five-year rate of return of 15.99% or less. Also, a mutual fund with a five-year rate of return of 16% or higher outperformed 95% of its peers. Notice that the last class 19–19.99% has a cumulative relative frequency of 1—this will always be the case.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **TABLE 17** | | | | |
| **Class (Five-year rate of return)** | **Frequency** | **Relative Frequency** | **Cumulative Frequency** | **Cumulative Relative Frequency** |
| 8–8.99 | 2 | 0.05 | 2 | 0.05 |
| 9–9.99 | 2 | 0.05 | 4 | 0.1 |
| 10–10.99 | 4 | 0.1 | 8 | 0.2 |
| 11–11.99 | 1 | 0.025 | 9 | 0.225 |
| 12–12.99 | 6 | 0.15 | 15 | 0.375 |
| 13–13.99 | 13 | 0.325 | 28 | 0.7 |
| 14–14.99 | 7 | 0.175 | 35 | 0.875 |
| 15–15.99 | 3 | 0.075 | 38 | 0.95 |
| 16–16.99 | 1 | 0.025 | 39 | 0.975 |
| 17–17.99 | 0 | 0 | 39 | 0.975 |
| 18–18.99 | 0 | 0 | 39 | 0.975 |
| 19–19.99 | 1 | 0.025 | 40 | 1 |

2.3 Additional Displays of Quantitative Data

**DEFINITION**

An **ogive** (read as “oh jive”) is a graph that represents the cumulative frequency or cumulative relative frequency for the class. It is constructed by plotting points whose x-coordinates are the upper class limits and whose y-coordinates are the cumulative frequencies or cumulative relative frequencies of the class. Then line segments are drawn connecting consecutive points. An additional line segment is drawn connecting the first point to the horizontal axis at a location representing the upper limit of the class that would precede the first class (if it existed).

Approach

A relative frequency ogive is drawn by plotting points whose x-coordinates are the upper class limit of each class and whose y-coordinates are the cumulative relative frequencies of each class. Then connect the points with line segments. Also, an additional line segment is drawn connecting the first point to the horizontal axis at a location representing the upper limit of the class that would precede the first class (if it existed).

Solution

See Figure 15. Notice how 20% of the mutual funds had a five-year rate of return less than or equal to 10.99%. Ogives do not have a line segment drawn from the last point to the horizontal axis because ogives represent the number or proportion of observations less than or equal to the x-coordinate of the point. Note the height of the last point in a relative frequency ogive is always 1.

|  |  |
| --- | --- |
| A line graph shows five-year rate of return. Vertical axis represents cumulative relative frequency and ranges from 0 to 1, with increment of 0.2. Horizontal axis represents five-year rate of return (percent) and ranges from 7.99 to 19.99, with increment of 1. |  |
| Figure 15 | |

2.3 Additional Displays of Quantitative Data

***Objective 5: Draw Time-Series Graphs***

Objective 5, Page 1

1. Define time-series data. If the value of a variable is measured at different points in time, then the data are referred to as **time-series data**. The closing price of Cisco Systems stock at the end of each year for the past 12 years is an example of time-series data.

Objective 5, Page 1 (continued)

Explain how to create a time-series plot. If the value of a variable is measured at different points in time, then the data are referred to as **time-series data**. The closing price of Cisco Systems stock at the end of each year for the past 12 years is an example of time-series data.

**DEFINITION**

A **time-series plot** is obtained by plotting the time in which a variable is measured on the horizontal axis and the corresponding value of the variable on the vertical axis. Line segments are then drawn connecting the points.

If the value of a variable is measured at different points in time, then the data are referred to as **time-series data**. The closing price of Cisco Systems stock at the end of each year for the past 12 years is an example of time-series data.

Problem

The Partisan Conflict Index (PCI) tracks the degrees of political disagreement among U.S. politicians in the federal government. It is found by measuring the frequency of newspaper articles reporting disagreement in a given month. Higher values of the index suggest greater conflict among political parties, Congress, and the President. The data in Table 18 represent the PCI in March from 1999 to 2017. Construct a time-series plot of the data. In what year was the index highest? In what year was the index lowest?

**Video Solution**

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |

**Technology Step-By-Step**

[](https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx#xln-lb-lnk_obj5_2_fbb1f99b-6c09-93ea-90e8-b709c9d4a987)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | **TABLE 18** | | | **Year** | **Partisan Conflict Index (PCI)** | | 1999 | 85.87 | | 2000 | 94.67 | | 2001 | 78.23 | | 2002 | 86.67 | | 2003 | 88.49 | | 2004 | 98.55 | | 2005 | 100.07 | | 2006 | 91.49 | | 2007 | 85.44 | | 2008 | 90.87 | | 2009 | 88.04 | | 2010 | 142.42 | | 2011 | 155.83 | | 2012 | 154.18 | | 2013 | 180.56 | | 2014 | 131.4 | | 2015 | 163.54 | | 2016 | 173.88 | | 2017 | 270.72 | | *Source* : Federal Reserve Bank of Philadelphia | | |  |

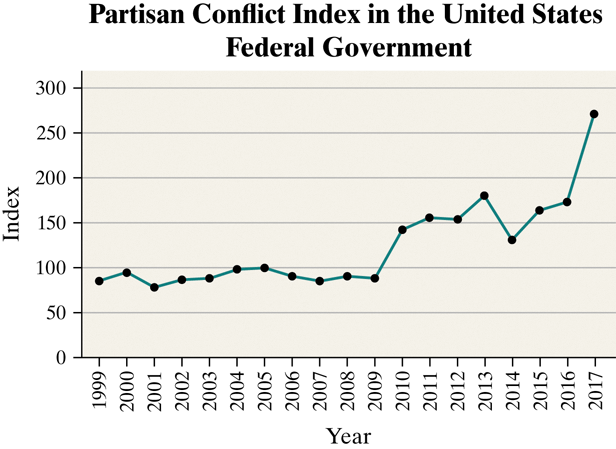
Approach

**Step 1.** Plot points for each year, with the date on the horizontal axis and the Partisan Conflict Index on the vertical axis.

SHOW STEP 2

2.3 Additional Displays of Quantitative Data

2.3 Additional Displays of Quantitative Data

This concludes the Interactive Reading Assignment for Section 2.3. Click A button reads “Save Assignment.” to save your and me back to the navigation page.

**Section 2.4 Graphical Misrepresentations of Data**

**Objective**

1. Describe What Can Make a Graph Misleading or Deceptive

***Objective 1: Describe What Can Make a Graph Misleading or Deceptive***

Objective 1, Page 1

*Answer the following after watching the video.*

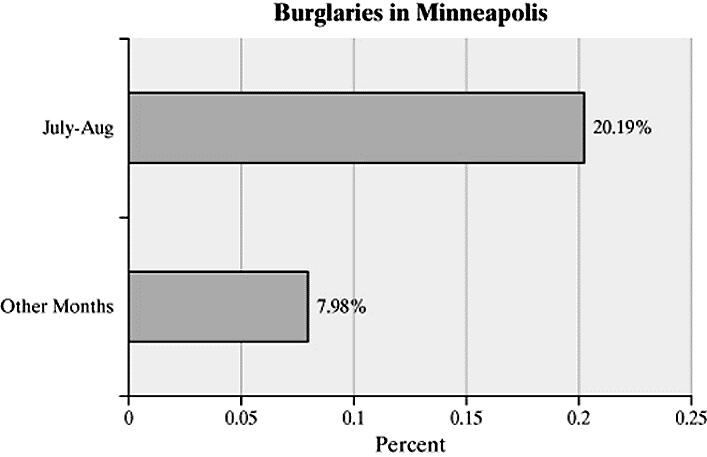
1. Explain the difference between a graph that is misleading and a graph that is deceiving. GRAPH IS MISLEADING IF IT IS UNINTENTIONAL. IT IS DECEIVING IF IT IS INTENTIONAL IMPRESSION
2. List what the most common misrepresentations of data involve.

* Increments between tick marks should be consistent.
* Scales for comparative graphs should be the same.
* The baseline, or zero point, should be at the bottom of the graph.

Objective 1, Page 2

**Example 1 *Misrepresentations of Data***

A home security company located in Minneapolis, Minnesota, develops a summer ad campaign with the slogan "When you leave for vacation, burglars leave for work." According to the city of Minneapolis, roughly 20% of home burglaries occur during the peak vacation months of July and August. The advertisement contains the graphic shown. Explain what is wrong with the graphic.



In this example, we'll take a look

at misrepresentation of data.

A home security company located in Minneapolis, Minnesota

develops a summer ad campaign with the slogan, when

you leave for vacation, burglars leave for work.

According to the city of Minneapolis, roughly 20%

of home burglaries occur during the peak vacation

months of July and August.

The advertisement contains the graphic shown.

Explain what is wrong with the graphic.

So let's consider how the categories of data are defined.

We know for all 12 months, the total has to be 100%.

And if we think about roughly 8% per month times 10 months,

that represents approximately 80% of the burglaries

happen in other months, showing us that this 20.19%

is actually the combined total for July and August,

not per each month.

The unsuspecting reader is misled

into thinking that July and August each have a burglary

rate of 20%.

This graph gives a better picture

of the burglary distribution.

The increased during the month of July

is not as dramatic as the bar graph on the previous screen.

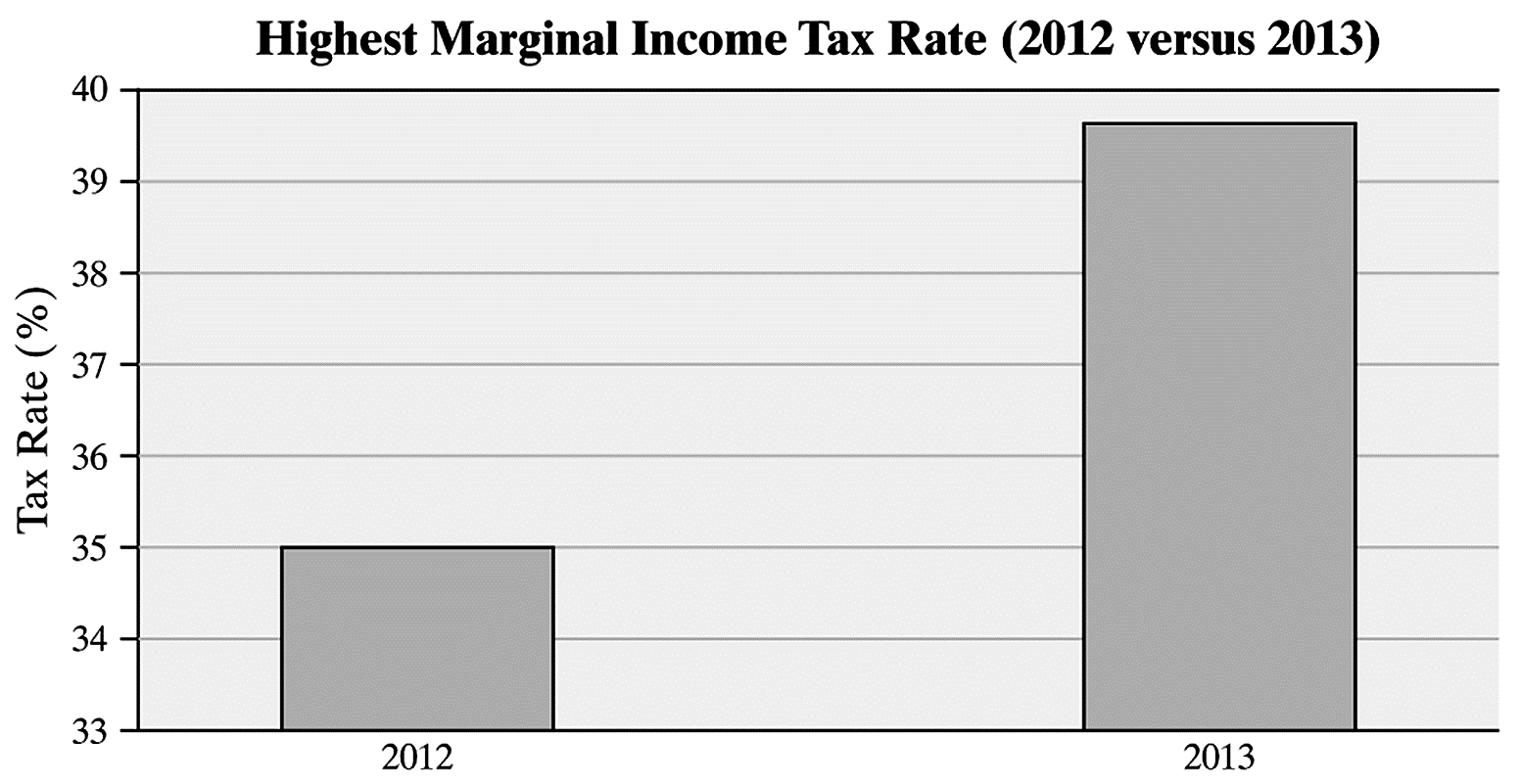
And as it works out, August actually

has fewer burglaries than September or October.

Objective 1, Page 3

**Example 2 *Misrepresentations of Data by Manipulating the Vertical Scale***

A national news organization developed the following graphic to illustrate the change in the highest marginal tax rate effective January 1, 2013. Why might this graph be considered misleading?



In this example, we'll take a look

at misrepresentation of data by manipulating

the vertical scale.

A national news organization developed the following graphic

to illustrate the change in the highest marginal tax

rate effective January 1, 2013.

Why might this graph be considered misleading?

When we take a look at this graph,

that may lead you to believe that marginal tax rates are

more than tripling since the height of the bar for 2013

is more than three times the height of the bar for 2012.

In fact, the difference is only 4.6 percentage points

because the tax rate in 2012 is 35% and in 2013 it's 39.6%.

The reason for this incorrect conclusion

is that the vertical axis does not begin at 0

but, instead, it begins at 33%.

Let's take a look at this graph that does not

distort the difference in tax rates.

The increase in tax rate is still

apparent in the graph-- this bar is definitely

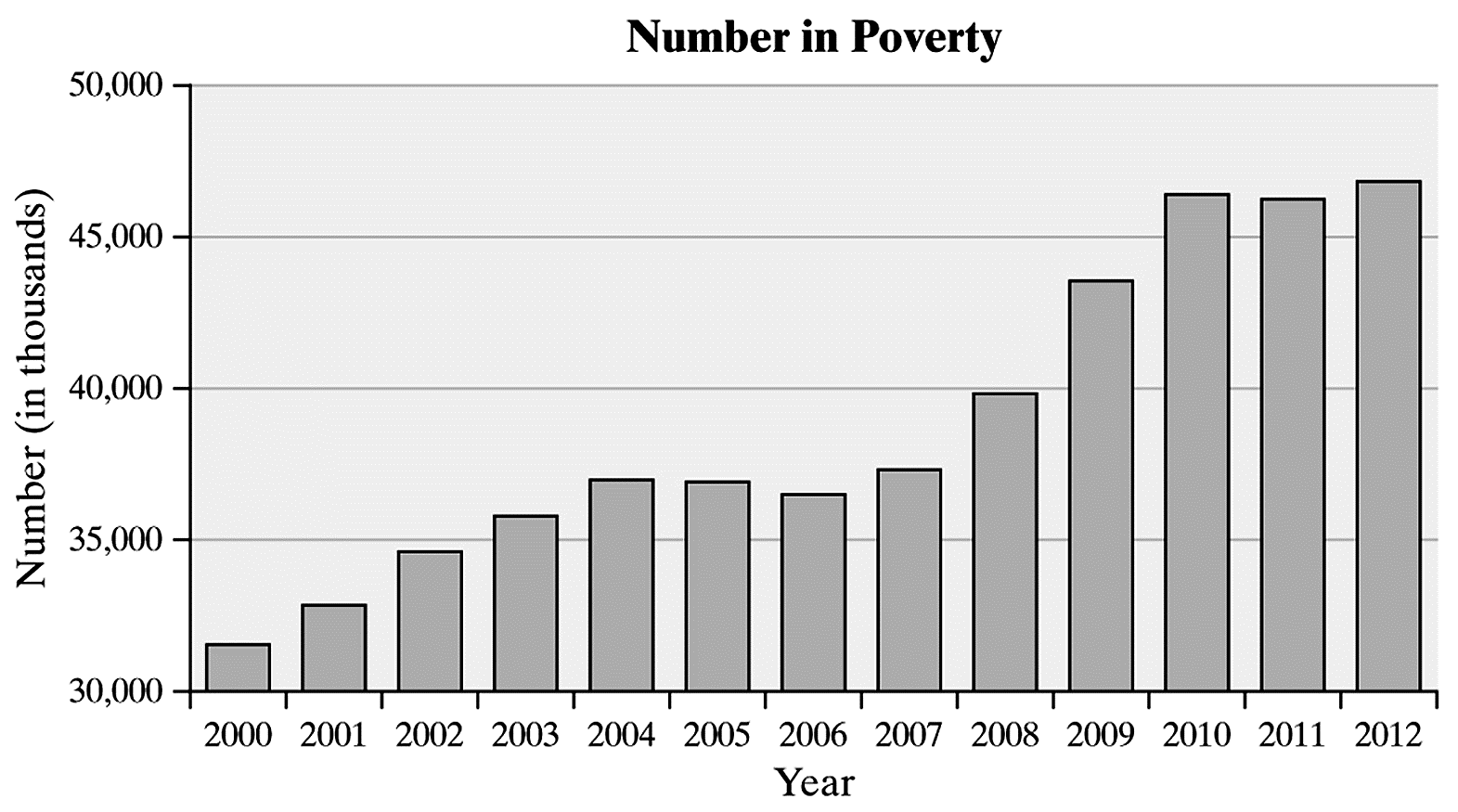
higher than that bar-- and without distorting

the size of the actual increase.

Objective 1, Page 5

**Example 3 *Misrepresentations of Data by Manipulating the Vertical Scale***

The graph depicts the number of residents in the United States living in poverty. Why might this graph be considered misrepresentative?



In this example, we'll talk about misrepresentation of data

by manipulating the vertical scale.

The following graph depicts the number

of residents in the United States living in poverty.

Why might this graph be considered "misrepresentative"?

Here's the graph.

The graph may mislead the reader into believing

that the number in poverty has more than doubled since 2005

because the bar for 2012 is more than twice the height

of the bar in 2005.

However, notice that the vertical axis

begins at 30,000 instead of 0.

This type of scaling is common when the smallest observed

data value is a rather large number.

It's not necessarily done intentionally

to confuse or mislead the reader.

Often, the main purpose in graphs-- particularly time

series graphs-- is to discover a trend rather than

the actual differences in the data.

However, the author of the graph should clearly

indicate that the graph does not begin at 0 by including

the following symbol in the vertical scale.

This symbol indicates that the scale has been truncated

and the graph has a gap.

Although the data does stand out,

it's better to use a time series plot

when displaying time series data rather than a bar plot.

In addition, it's better to use the percent of the population

in poverty rather than the actual number

living in poverty.

This is due to the fact that increases in poverty

may be due to increases in population

as well as a deterioration of the economy.

Here's a time series plot of the percentage of US

residents living in poverty.

The lack of bars allow us to focus on the trend in the data

rather than the relative size or area of the bars.

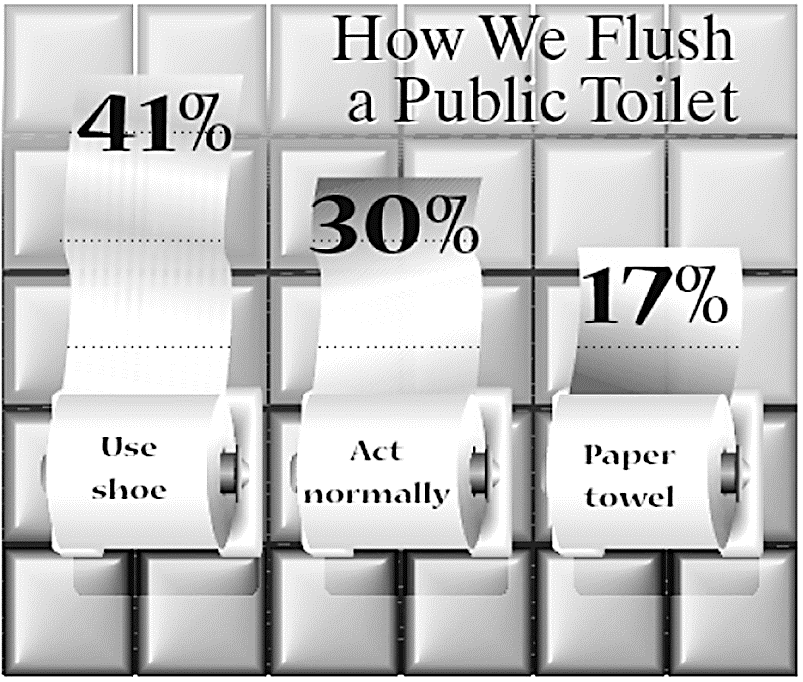
Also, it's clear that the percent in poverty

was highest in 2010, and this is not clear from the bar graphs.

Objective 1, Page 7

**Example 4 *Misrepresentations of Data***

The bar graph shown is a *USA Today*-type graph. A survey was conducted by Impulse Research in which individuals were asked how they would flush a toilet when the facilities are not sanitary. What is wrong with the graphic?



In this example, we'll take a look

at misrepresentation of data.

The following bar graph is a USA Today type graph.

The survey was conducted by Impulse Research

for Quilted Northern Confidential, in which

individuals were asked how they would flush a toilet when

the facilities are not sanitary.

What's wrong with the graph there?

A simple bar chart would have got

the idea across in a clear and concise way.

They used the toilet paper to tie it into the research

question, but it's very misleading.

The first thing that we're unsure of

is whether the roll itself is supposed

to be part of the graph or not.

When we compare 41% to 17%, that's

supposed to be 2.4 times bigger.

If we compare the height including the roll of toilet

paper, then this height is less than twice that height.

Meaning that it's not accurate.

So maybe they intended us to only

use the part up to but not including the roll.

Well, if we did that, it's still about twice that height,

not 2.4 times.

So the vertical scale is off.

Objective 1, Page 9

1. Why is the use of 3-D effects strongly discouraged?
2. Why do we emphasize that the bars or classes should have the same width?

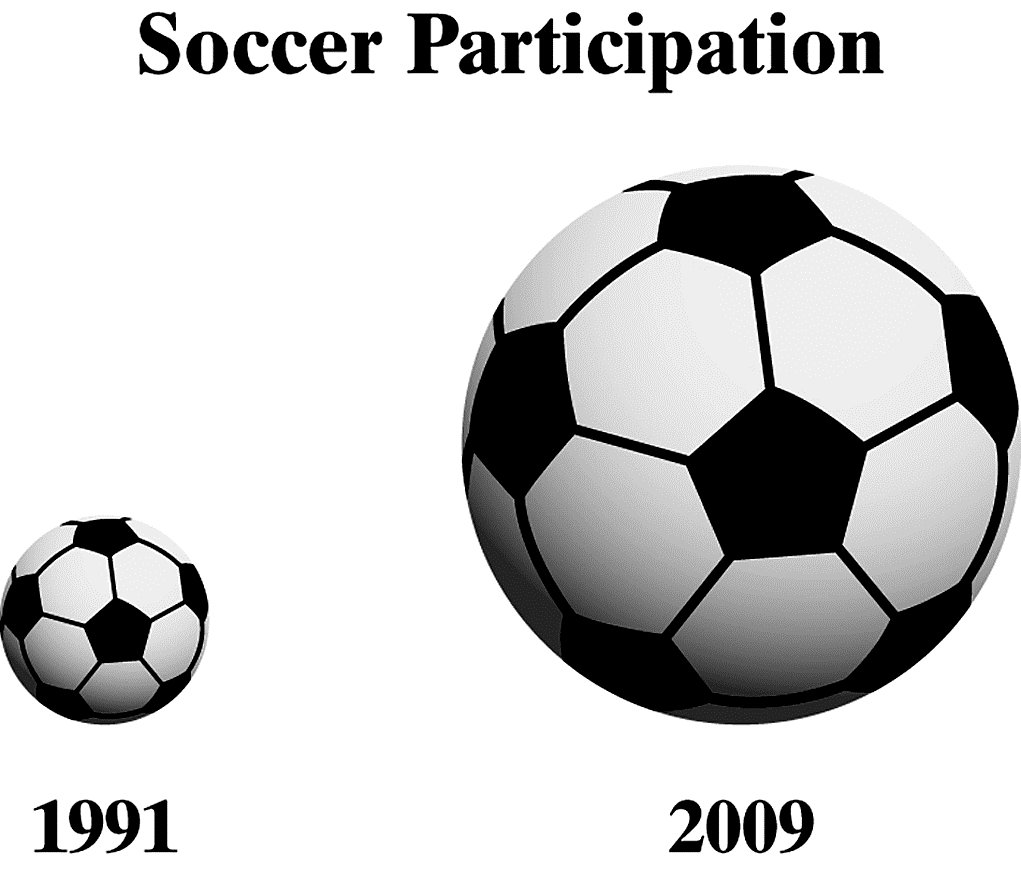
Newspapers, magazines, and Internet sites often go for a "wow" factor when displaying graphs. The graph designer may be more interested in catching the reader's eye than making the data stand out. The two most commonly used tactics are 3-D graphs and *pictograms* (graphs that use pictures to represent the data). The use of 3-D effects is strongly discouraged because such graphs are often difficult to read, add little value to the graph, and distract the reader from the data.

When comparing bars, our eyes are really comparing the *areas* of the bars. That is why we emphasize that the bars or classes should have the same width. Uniform width ensures that the area of the bar is proportional to its height so that we can simply compare the heights of the bars. However, when we use two-dimensional pictures in place of bars, as with pictograms, it is not possible to obtain a uniform width. To avoid distorting the picture when values increase or decrease, both the height and width of the picture must be adjusted. This often leads to misleading graphs.

Objective 1, Page 10

**Example 5 *Misrepresentations of Data by Manipulating Dimension***

Soccer continues to grow in popularity as a sport in the United States. In 1991, there were approximately 10 million participants in the United States aged 7 years and older. By 2009, this number had climbed to 14 million. To illustrate this increase, we could create a graphic like the one shown below. Describe how the graph may be misleading. *Source:* U.S. Census Bureau; National Sporting Goods Association



In this example, we'll take a look at misleading graphs.

Soccer continues to grow in popularity

as a sport in the United States.

In 1991, there were approximately

10 million participants in the United States aged 7 or older.

By 2009, this number had climbed to 14 million.

To illustrate this increase, we could create a graphic

like the following.

Describe how the graph may be misleading.

The actual increase from 10 million to 14 million

participants is an increase of 40%.

Now, I ask you-- does the soccer ball on the right

seem 40% larger than the ball on the left?

Actually, it's four times the size,

which represents an increase of over 300%.

One problem here is that the actual data are not labeled.

If it said "10 million" and "14 million"

next to the two soccer balls, you

could understand the increase just

from the values-- ignoring the size of the pictures.

But since those are missing, there's

not much else you can do, except draw the wrong conclusion as

compared to the right conclusion.

A simple bar graph would have shown the difference

between the two years-- 1991 and 2009-- in a much better way.

We can see the increase-- from 10 million to 14 million--

without distorting the facts.

The bar on the right is definitely

40% larger than the bar on the left.

Now, if you wanted to do this and still

make it look appealing, you could try a graph

like the following where one soccer ball represents

1 million participants.

We can see, from 2009, that there were 4 million

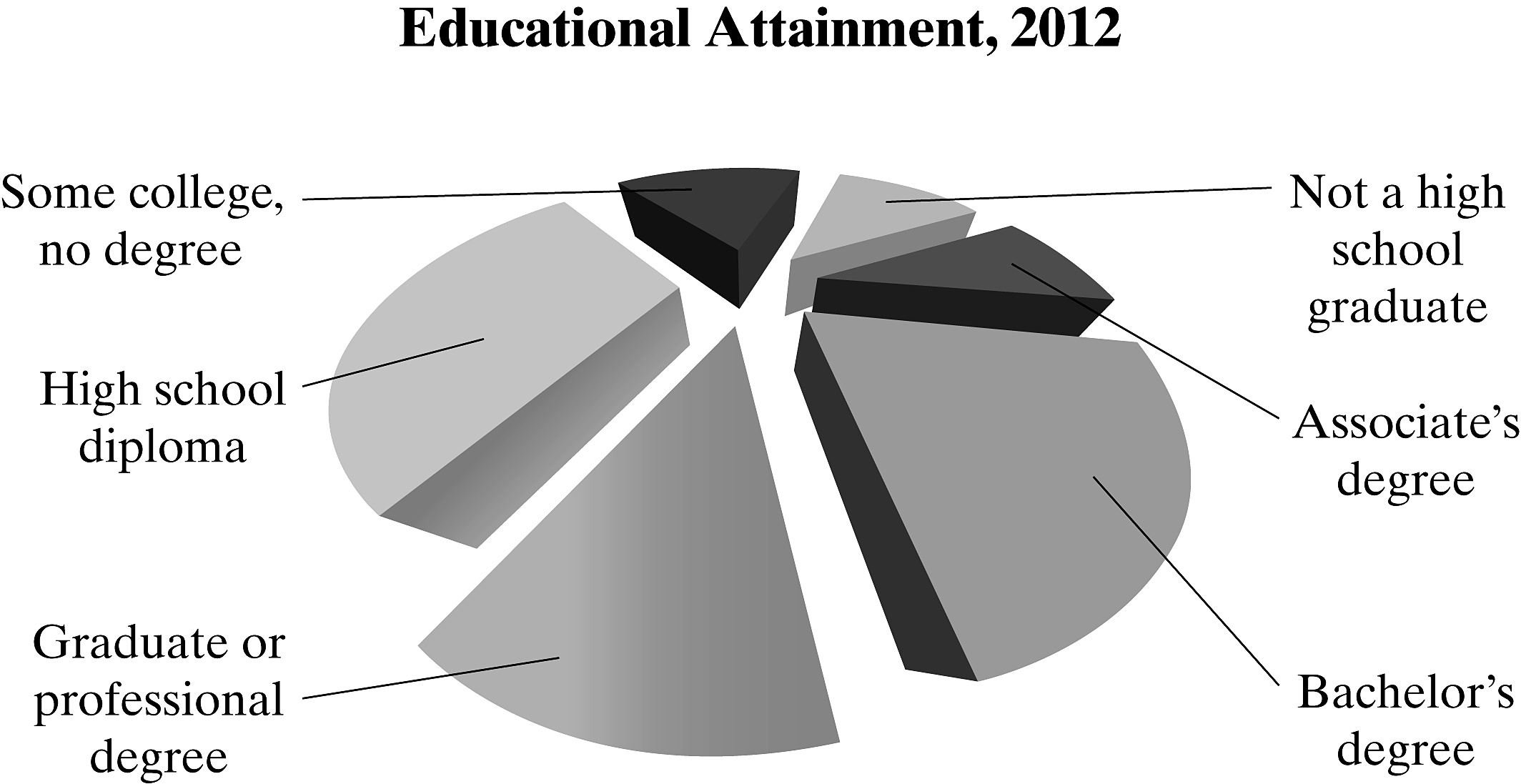
more participants because of the 4 extra soccer balls

than there were in 1991.

Objective 1, Page 11

**Example 6 *Misrepresentations of Data: Three-Dimensional Scale***

The figure represents the educational attainment (level of education) in 2016 of adults 25 years and older who are U.S. residents. Why might this graph be considered misrepresentative?



n this example, we'll discuss misleading graphs.

The figure represents the educational attainment level

of education in 2016 of adults 25 years and older

who are US residents.

Why might this graph be considered misrepresentative?

Here's the graph.

When working with a 3D style graph,

one problem that we run into is that the category

hat's closest to us appears to be overrepresented.

Here, the blue wedge facing us, bachelors degree, is only 21%

of the entire pie.

However, it appears to be much larger than that.

Here are the data that the results are based upon.

And 44,778 out of the total is roughly 21%.

n this case, a standard pie chart

would display the data in a much fairer representation.

Again, if we look at the bachelor's degree here,

we can clearly see that its percentage is less than that

of, say, high school diploma.

But if we go back to the 3D pie chart,

t appears to be close in size to that.

Let's take one quick look at what

happens as we start to rotate these pie slices around

n a three dimensional way.

Here's that pie chart, and we're going

o take a look at what happens when

we rotate our perspective or our view of this pie chart.

Here we go.

Notice that, for instance, the yellow slice

changes as we move it around in terms

of how large that proportion seems to be

or that percentage seems to be.

And that's why these 3D pie charts are not a great way

o represent a set of data.