2 Describing Your Data: Frequencies, Cross Tabulations, and Graphs

Learning Objectives:

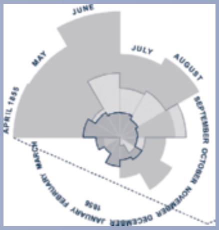
By the end of this chapter you should be able to:

1. Define and describe the terms frequency and frequency distribution.
2. Define and describe the terms relative frequency, percentage frequency, and cumulative percentage frequency.
3. Construct frequency tables for nominal and ordinal data.
4. Construct class intervals for interval and ratio data.
5. Construct frequency tables for interval and ratio data.
6. Create cross tabulations.
7. Calculate percentage change, ratios, and rates.
8. Create and interpret pie and bar charts.
9. Create and interpret frequency polygons and cumulative percentage frequency polygons.
10. Create and interpret histograms, stem-and-leaf plots, and boxplots.

A Brief History of…

Florence Nightingale (1820–1910)

Florence Nightingale, nicknamed "the lady with the lamp," is known for her work in nursing. During the Crimean War (1853–1856), Nightingale sent reports of the conditions and treatment of patients back to Britain. As part of her reports, Nightingale used visual presentations of her data in the form of variations of pie charts. In order to present monthly deaths, she used a more elaborate pie chart that included different forms of death by month. She called these pictures "coxcombs" (see example on the right).



After the war, Nightingale lobbied hard for sanitary reforms in the hospitals. In 1873, while working on the improvement of sanitary conditions in India, Nightingale reported that the mortality rate had dropped from 69 to 18 deaths per 1,000 soldiers.

In 1859, Florence Nightingale became the first female to be elected to the Royal Statistical Society. Shortly after that she became an honorary member of the American Statistical Association.[1]

Introduction

Empirical data is gathered from objects or participants for a research study.

Researchers use a variety of different methods (such as surveys, interviews, experiments, databases, etc.) to gather **empirical data.** For example, a sociologist gathers data on cultural norms, a political scientist on voting behaviour, an economist on stock fluctuation, and an educational psychologist on gender difference in academic performance. All social science disciplines gather some form of empirical data from the real world. Once gathered, the data needs to be organized and presented in a manner that can provide summary information about the phenomena of interest.

In this chapter we will focus on different ways to present summary information about your data using frequency tables, cross tabulations, and graphs. However, before we do that we need to review how data gets from the collection stage to a dataset that we can work with. Figure 2.1 provides four examples of how variables with different levels of measurement can be measured, coded, and entered in a dataset. Here you can see that the variable "Age," measured at the ordinal level, is given a specific coding and entered in a data analysis program (such as SPSS® or Microsoft® Excel). We code variable responses in order to have numeric information to work with. The coding you use is largely based on what makes the most sense to you. At this stage, it is important to keep notes of which codes match which responses (often called a data dictionary) in order to make the correct interpretations.

FIGURE 2.1 **Measurement and Coding**

	Question Example	**Coding**	**Dataset**	
Nominal	**Gender:**		***n***	**Gender**
(Gender)	□ Male	Male = 1	1	1
	□ Female	Female = 2	2	2
			3	1
			4	1
			5	2
			6	2
			7	2
			8	1
			9	1
			10	2
Ordinal	**Please state your age.**		***n***	**Age**
(Age)	□ 20 to 25	20 to 25 = 1	1	1
	□ 26 to 35	26 to 35 = 2	2	2
	□ 36 to 45	36 to 45 = 3	3	2
			4	1
			5	3
			6	2
			7	3
			8	1
			9	1
			10	2
Interval	**I am satisfied with my life:**		***n***	**Life Satisfied**
(Satisfied with life)	□ Strongly Disagree	Strongly disagree = 1	1	4
	□ Disagree	Disagree = 2	2	4
	□ Neutral	Neutral = 3	3	3
	□ Agree	Agree = 4	4	2
	□ Strongly Agree	Strongly Agree = 5	5	4
			6	1
			7	5
			8	3
			9	4
			10	2
Ratio (Internet hours)	**How many hours per *week* do you spend on the Internet?** ________	Record actual number of hours	***n***	**Internet Hours**
			1	3
			2	15
			3	12
			4	6
			5	9
			6	10
			7	11
			8	0
			9	2
			10	8

n = respondent number

FIGURE 2.2
Example of a Dataset

n	Gender	Age	Life Satisfied	Internet Hours
1	1	1	4	3
2	2	2	4	15
3	1	2	3	12
4	1	1	2	6
5	2	3	4	9
6	2	2	1	10
7	2	3	5	11
8	1	1	3	0
9	1	1	4	2
10	2	2	2	8

n = respondent number

FIGURE 2.3
Measurement Level and Type of Data

Variables Measured at the ...		... Create This Type of Data
Nominal Level	→	Nominal Data
Ordinal Level	→	Ordinal Data
Interval Level	→	Interval Data
Ratio Level	→	Ratio Data

Figure 2.2 provides an example of what the combined data for the examples in Figure 2.1 may look like in a dataset.

One final note about the terminology we use regarding variables. We know that variables can be measured at four different levels: nominal, ordinal, interval, and ratio. When we assess a variable with a specific level of measurement we say that it produces a specific type of data. Figure 2.3 captures this idea. For example, when measuring a variable with an interval level of measurement, we say that it produces interval data. Given this, we often shorten the phrase "a variable with an interval level of measurement" to just "an interval variable." So when you see "ordinal variable," this is referring to a variable that has been measured at the ordinal level, which creates ordinal data.

Frequency Distributions and Tables

LO1

Recall from chapter 1 that a variable is a phenomenon of interest that can take on different values. Furthermore, nominal and ordinal variables have qualitative categories (referred to as just categories) as potential values, whereas interval and ratio variables have quantitative values. Given that a variable can have different values, we can count the **frequency** (also referred to as absolute frequency) of the observed values (categories or quantitative values) of a variable. Consider the following two examples.

Frequency refers to the number of observations of a specific value within a variable. Some textbooks call this absolute frequency but the meaning is the same.

Example 1: The variable 'Gender' has two possible qualitative categories, male and female. Suppose we collect data on gender by surveying 100 respondents, and observe that 46 are male and 54 are female. In this case, there are two values with differing frequencies.

Example 2: Suppose we measure "Student Grade" with a ratio level of measurement, and assume that grades do not include decimals or negative numbers. There are then 101 potential quantitative values, ranging from 0 to 100. Now imagine that we sampled 10 students and observed that three students received a grade of 76, two students received a grade of 79, one student received a grade of 83, and four students received a grade of 86. In this case we observed four potential values with differing frequencies.

Once the data is entered in a software program, we want to get a sense of what the data looks like at a summary level. One way to summarize data into a form that can be easily reviewed is to create a frequency distribution within a table. A **frequency distribution** is the summary of the values of a variable based on the frequencies with which they occur. It is called a frequency distribution because we are looking at how the values of the variable are distributed across all of the cases in the data. When we display the frequency distribution in a table format, we call it a frequency table. As an example, think back to the survey question in Figure 1.10 in chapter 1, which measured the variable "Political Party Affiliation." If we administered the question to 100 respondents, we could calculate the frequency (f) with which each category was selected by the respondent and create the frequency table in Table 2.1.

Frequency distribution is the summary of the values of a variable based on the frequencies with which they occur.

LO2

Often, it is easier to interpret frequency results when they are converted into **relative frequency** (f/n). Relative frequency is a comparative measure of the proportion of observed values (category or quantitative value) to the total number of responses within a variable. It provides us with the proportion or fraction of one occurance relative to all occurances. You will often hear this referred to as a proportion.

Relative frequency is a comparative measure of the proportion of observed values to the total number of responses within a variable.

The equation for calculating relative frequency (f/n) is:

$$\text{relative frequency} = \frac{f}{n} \qquad \textbf{(2.1)}$$

where: f = frequency of specific responses
n = total number of responses

We use a small "n" to represent the size of a sample and a capital "N" to represent the size of a population.

TABLE 2.1
Frequency of "Political Party Affiliation"

Category	Frequency (f)
Conservative	35
Green Party	10
Le Bloc Québécois	10
Liberal Party	33
NDP	12
Total	100

TABLE 2.2
2009 Statistics Canada Population Estimates of the Northwest Territories

Source: "2009 Statistics Canada Population Estimates of the Northwest Territories," adapted from Statistics Canada website, http://www40.statcan.ca/l01/cst01/demo31a-eng.htm, extracted May 18, 2011.

	Frequency (f)	Relative Frequency (f/n)
Males	22,500	0.517
Females	21,000	0.483
Total	43,500	1.00

Consider Table 2.2, which provides the 2009 Statistics Canada population estimates for the Northwest Territories. The right column provides the relative frequency of males and females. Given that relative frequencies must add to 1.0, it is generally easier to visualize the comparison of relative frequency than raw numbers. For example, it is likely easier to see the magnitude of difference in 0.517 males versus 0.483 females than it is in 22,500 males versus 21,000 females.

A **percentage frequency** is the relative frequency expressed as a percentage value.

Similar to relative frequency, percentage frequency ($\%f$) also provides a useful way of displaying the frequency of data. A **percentage frequency** (commonly referred to as percentage) is the relative frequency expressed as a percentage value (out of 100) and can be calculated as follows:

$$\%f = \frac{f}{n} \times 100 \qquad \textbf{(2.2)}$$

where: f = frequency of responses
n = total number of responses within the variable

The **cumulative percentage frequency** gives the percentage of observations up to the end of a specific value.

Since relative frequencies are written as a decimal (e.g., 0.10), we can convert them to percentage frequencies by multiplying the relative frequency by 100. For example, 0.10 becomes 10 percent (0.10 × 100). It is also useful to show the cumulative percentage frequency ($c.\%f$) (commonly referred to as cumulative precentage). The **cumulative percentage frequency** gives the percentage of observations up to the end of a specific value. Table 2.3 provides the 2009 Statistics Canada population estimates for the Northwest Territories with the percentage frequency added. Again, we can see that it is easier to see the difference in males versus females when we state that 51.7 percent are male and 48.3 percent are female rather than 22,500 are males versus 21,000 are females.

Now that we have covered some of the basics of frequency tables, we need to focus on how to create frequency tables for nominal, ordinal, interval, and ratio variables. In this section, we will focus two types of frequency tables: simple frequency tables and cross-tabulations.

TABLE 2.3
2009 Statistics Canada Population Estimates of the Northwest Territories

Source: "2009 Statistics Canada Population Estimates of the Northwest Territories," adapted from Statistics Canada website, http://www40.statcan.ca/l01/cst01/demo31b-eng.htm, extracted May 18, 2011.

	Frequency (f)	Relative Frequency (f/n)	Percentage Frequency ($\%f$)	Cumulative Percentage Frequency ($c.\%f$)
Males	22,500	0.517	51.7	51.7
Females	21,000	0.483	48.3	100.0
Total	43,500	1.000	100.0	

Take a Closer Look

Example of Population Estimates by Province
To show the value of relative frequency and percentage frequency, Table 2.4 provides the Statistics Canada 2009 population estimates by province and territory for individuals 65 years of age or older. Looking at the relative frequency and percentage frequency, we can see that just over 77 percent of the population age 65 and older live in British Columbia, Ontario, and Quebec.

TABLE 2.4 Statistics Canada 2009 Population Estimates

Source: "Statistics Canada 2009 Population Estimates," adapted from Statistics Canada website, http://www40.statcan.gc.ca/l01/cst01/demo31a-eng.htm, extracted May 18, 2011.

	Population 65+ (Frequency)	Relative Frequency (*f*/*n*)	Percentage Frequency (%*f*)	Cumulative Percentage Frequency (*c*.%*f*)
Newfoundland and Labrador	75,200	0.0160	1.60	1.60
Prince Edward Island	21,600	0.0046	0.46	2.07
Nova Scotia	147,900	0.0316	3.16	5.22
New Brunswick	116,400	0.0248	2.48	7.70
Quebec	1,170,400	0.2497	24.97	32.67
Ontario	1,787,900	0.3814	38.14	70.82
Manitoba	168,500	0.3590	3.59	74.41
Saskatchewan	151,900	0.0324	3.24	77.65
Alberta	385,200	0.0822	8.22	85.87
British Columbia	656,300	0.1400	14.00	99.87
Yukon	2,700	0.0006	0.06	99.93
Northwest Territories	2,300	0.0005	0.05	99.98
Nunavut	1,000	0.0002	0.02	100.00
Total	4,687,300	1.0000	100.00	

LO3

Simple Frequency Tables for Nominal and Ordinal Data

A simple frequency table displays the frequency distribution of one variable at a time. These variables can be nominal, ordinal, interval, or ratio. To create a frequency table, list the possible values the variable can have in one column and record the number of times (frequency) that each value occurs in another column. Figures 2.4 and 2.5 provide examples of frequency tables for nominal and ordinal. These figures provide a diagrammatic view of how data goes from the collection stage (in this case by survey question) to data entry and then to the frequency table.

As you can see it is easy to create frequency tables for nominal and ordinal variables as they have a limited **range** of potential values.

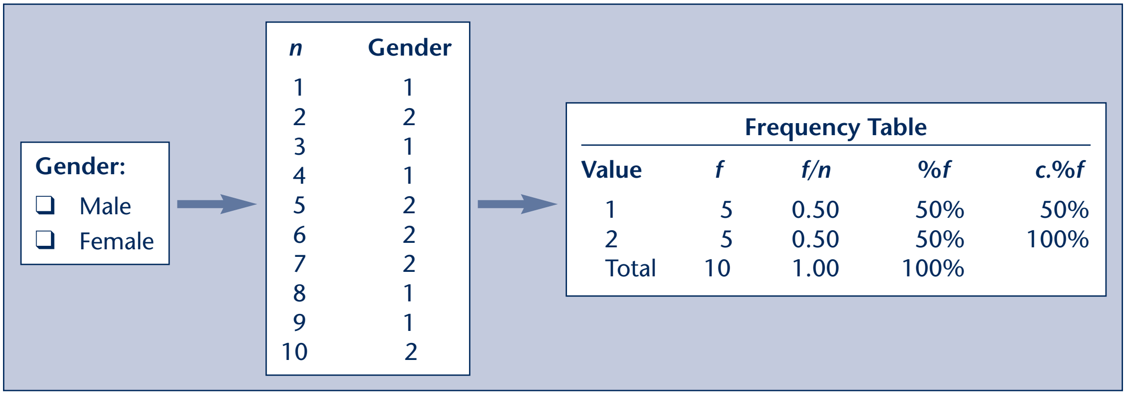
The **range** of the data is the value of the largest observation minus the value of the smallest observation.

LO4 LO5

Simple Frequency Tables for Interval and Ratio Data

Creating frequency tables for interval variables can also be fairly straightforward depending on how the variable is measured. In Figure 2.6, "Satisfied With Life" is measured with a Likert scale, making it an interval variable. Since there are only

FIGURE 2.4 **Nominal Data**



Gender:

□ Male

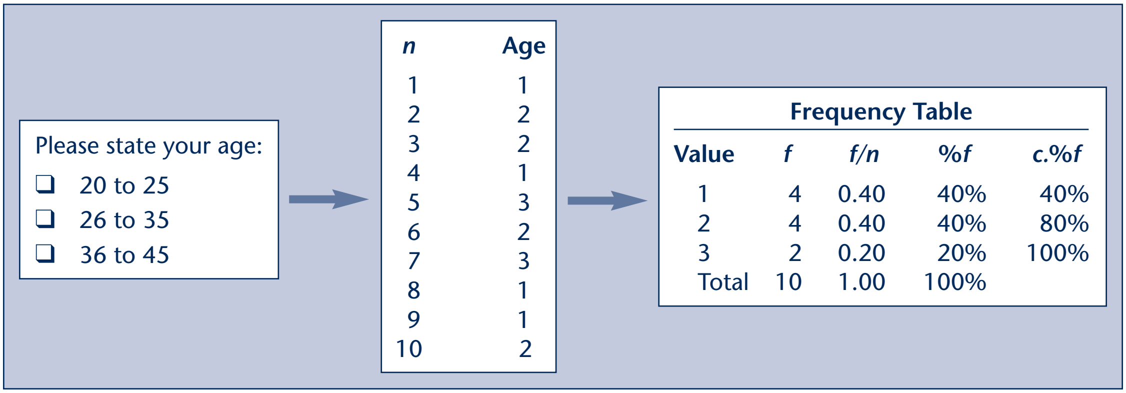
□ Female

n	Gender
1	1
2	2
3	1
4	1
5	2
6	2
7	2
8	1
9	1
10	2

Frequency Table

Value	*f*	*f/n*	%*f*	c.%*f*
1	5	0.50	50%	50%
2	5	0.50	50%	100%
Total	10	1.00	100%	

FIGURE 2.5 **Ordinal Data**



Please state your age:

□ 20 to 25

□ 26 to 35

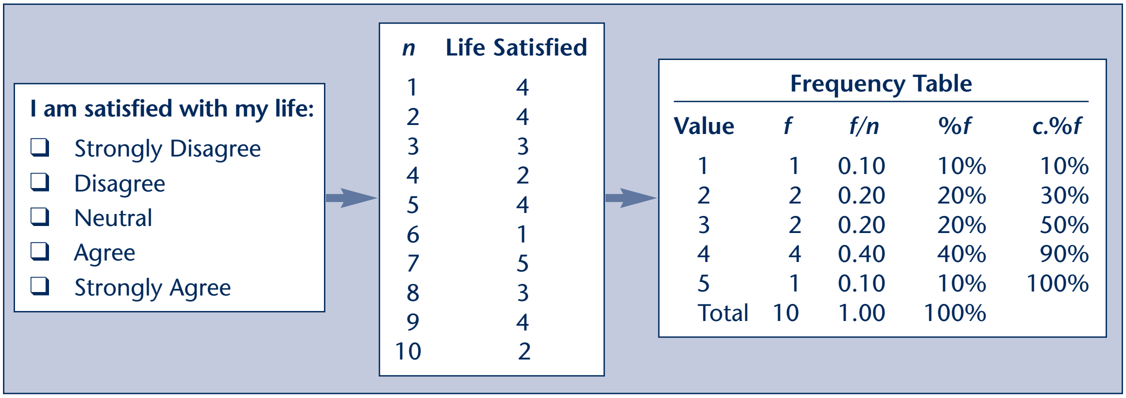
□ 36 to 45

n	Age
1	1
2	2
3	2
4	1
5	3
6	2
7	3
8	1
9	1
10	2

Frequency Table

Value	*f*	*f/n*	%*f*	c.%*f*
1	4	0.40	40%	40%
2	4	0.40	40%	80%
3	2	0.20	20%	100%
Total	10	1.00	100%	

FIGURE 2.6 **Interval Data**



I am satisfied with my life:

□ Strongly Disagree

□ Disagree

□ Neutral

□ Agree

□ Strongly Agree

n	Life Satisfied
1	4
2	4
3	3
4	2
5	4
6	1
7	5
8	3
9	4
10	2

Frequency Table

Value	*f*	*f/n*	%*f*	c.%*f*
1	1	0.10	10%	10%
2	2	0.20	20%	30%
3	2	0.20	20%	50%
4	4	0.40	40%	90%
5	1	0.10	10%	100%
Total	10	1.00	100%	

five potential values, the process for creating a frequency table for this variable is the same as that for nominal and ordinal variables.

Creating a frequency table that provides an easy-to-read summary for many interval variables and ratio variables can be a bit more difficult given the range of potential values to include. Figure 2.7 provides an example of this issue.

Here the values range from 0 to 15, but given a large enough sample there are potentially 168 values (24 hours per day × 7 days in a week). Similarly, consider the interval variables "IQ Scores" or the Yale-Brown "Obsessive Compulsive Disorder (OCD) Score." In these cases the values may range from 0 to 200+ and 0 to 40 respectively. The bottom line is that when there are too many values on which to report frequencies, the frequency table become less useful as a device to communicate summary information about the data.

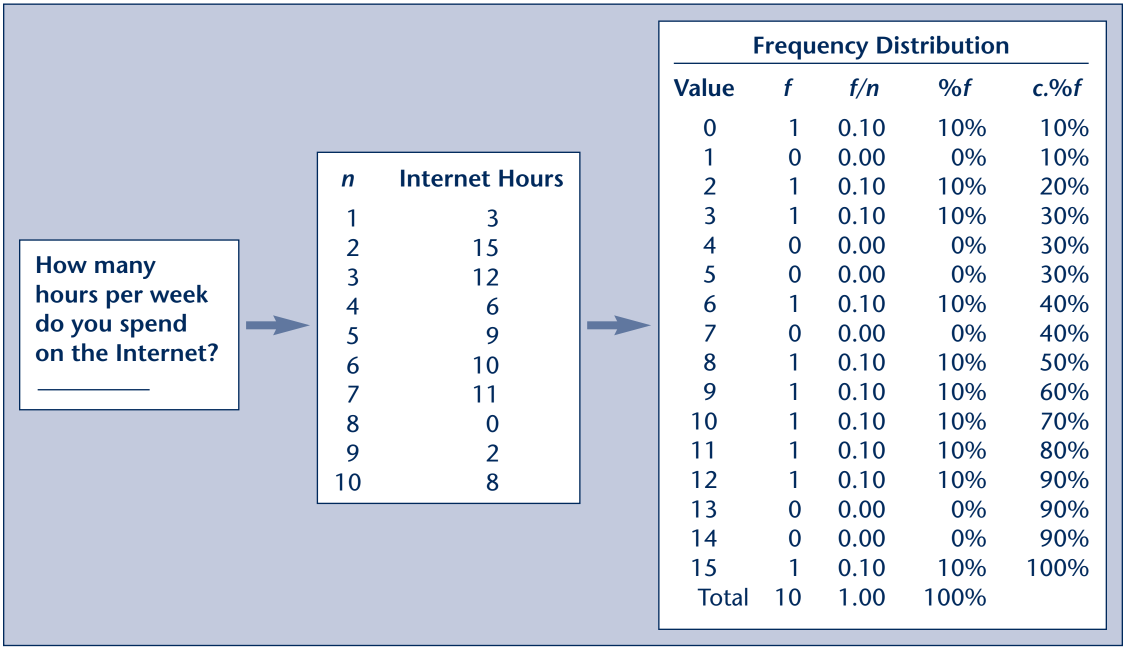
To get around this problem we use class intervals (also called grouped frequencies) to create frequency tables for interval and ratio variables that have a large range of potential values. A **class interval** is a set of values that are combined into a single group for a frequency table. Class intervals have a **class width**, which is the range of each interval, and starting and end values called **class limits**. For class intervals to be meaningful they must following two criteria. First, the class intervals must be exhaustive, meaning that they must include the entire range of the data. Second, class intervals must be mutually exclusive, meaning that the class widths are unique enough that an observed value can only be placed into one class interval.

Class interval is a set of values that are combined into a single group for a frequency table.

Class width is the range of each class interval.

Class limits are actual values in the data that are used as starting and ending values in each class interval

FIGURE 2.7 **Ratio Data**



How many hours per week do you spend on the Internet?

______

n	Internet Hours
1	3
2	15
3	12
4	6
5	9
6	10
7	11
8	0
9	2
10	8

Frequency Distribution

Value	*f*	*f/n*	%*f*	c.%*f*
0	1	0.10	10%	10%
1	0	0.00	0%	10%
2	1	0.10	10%	20%
3	1	0.10	10%	30%
4	0	0.00	0%	30%
5	0	0.00	0%	30%
6	1	0.10	10%	40%
7	0	0.00	0%	40%
8	1	0.10	10%	50%
9	1	0.10	10%	60%
10	1	0.10	10%	70%
11	1	0.10	10%	80%
12	1	0.10	10%	90%
13	0	0.00	0%	90%
14	0	0.00	0%	90%
15	1	0.10	10%	100%
Total	10	1.00	100%	

FIGURE 2.8
Comparing Frequency Tables With and Without Class Intervals

Frequency without Class Intervals

Value	*f*
0	1
1	0
2	1
3	1
4	0
5	0
6	1
7	0
8	1
9	1
10	1
11	1
12	1
13	0
14	0
15	1
Total	10

Frequency with Class Intervals

Value	*f*
0 to 3	3
4 to 7	1
8 to 11	4
12 to 15	2
Total	10

For example, consider Figure 2.7. We could group the hours spent per week on the Internet into four class intervals with widths of four: 0–3, 4–7, 8–11, and 12–15. It is important to note that in doing this, we are not actually changing the data, we are only creating groups for the purpose of presenting a frequency table that summarizes the data. Figure 2.8 compares the frequency table in Figure 2.7 to one where four class intervals have been created. You can see that although you lose some of the detail when using class intervals, it is easier to read.

LO4

Creating Class Intervals

Table 2.5 contains the data for the "number of immigrants (in thousands) to Canada" gathered over the course of 20 three-month periods from 2000 to 2004.

Although this example is for a ratio variable, the process that follows is the same for interval variables. Given that the values range from 42 to 73, we need to create class intervals to summarize these numbers into a frequency table. To do so, we need to determine the width and number of class intervals. It is important

TABLE 2.5 **Number of Immigrants (in thousands) to Canada From 2000 to 2004**

Source: "Number of Immigrants (in thousands) to Canada from 2000 to 2004," adapted from Statistics Canada, CANSIM Table 051-0006, http://www5.statcan.gc.ca/cansim/a05?lang=eng&id=0510006, extracted May 18, 2011.

46	58	67	57	58	70	70	52	62	73
53	42	46	59	63	54	55	67	68	48

to choose a reasonable number of intervals. One class interval (42 to 73) that includes 20 observations is of little use, as is 20 intervals that include only one observation each. There is no set number of intervals to include, so it is important to consider the audience you are reporting to when creating the number of class intervals. Generally speaking, five to seven intervals are usually sufficient to give a graphical portrayal of the data. The following steps outline how to create class intervals.

Step 1: Determine the Range of the Data

Since the value of the smallest observation is 46 and the value of the largest is 73, the range is 31 (73 − 42 = 31).

Step 2: Determine the Width and Number of the Class Intervals

Since we want the width and number of class intervals to incorporate all of the data, we divide the range of the data by the number of class intervals we would like to have and, if necessary, round up to the next value. If we opt for four class intervals, the width of each interval is 7.75 (31 ÷ 4), rounded up to 8. With the width of 8 for each class interval, our intervals are 42–49, 50–57, 58–65, and 66–73. To ensure that the intervals are exclusive of one another, we add 1 to the ending class limit value to create the beginning class limit of the next interval.

Ideally, you want the intervals to be of equal width. However, you may have a situation where the number of desired intervals creates a class interval that falls outside of the range of the data. For example, suppose we opted for five class intervals. Our width would then be 7 (31 ÷ 5 = 6.2, rounded to 7) and our intervals would be 42–48, 49–55, 56–62, 63–69, and 70–76. Since the largest value in our data is 73 it would be misleading to keep the last interval at 76. In this case, we may decide to make the width of the last interval unqual to the rest. For example, we may make the final interval 70–73. The same can be true for the first interval. To summarize then, ideally you want class intervals of equal width. However, if that is not possible given your desired number of class intervals, you can adjust the first or last (or both) interval to be unequal to the others so that your frequency table accurately represents the range of the observed data. When doing so, be sure to keep the remaining intervals at equal widths.

Step 3: Determine the Class Boundaries

Even though you have class intervals that are exclusive of one another, you still may have observations on the boundary of two classes intervals. For example, if your class intervals were 42–49 and 50–57 where would you put the value 49.5 if it existed in the data? Would it go in the first class interval or the second? You can see that there is a gap between the limits (between 49 and 50). With continuous data (such as that in Table 2.5) having these gaps causes problems when we have values (such as 49.5) that fall between them. To deal with this we create class boundaries that represent the real limits of the class intervals. Class boundaries are numbers that may not necessarily exist in the data but define where the cut-offs are for each class interval. To calculate the class boundary, you subtract 0.50 from the lower class limit and add 0.50 to the upper class limit for each class interval. The boundaries do not have a value separating them, like class intervals do, since they are continuous. Therefore, our class boundaries are 41.5 to less than (<) 49.5, 49.5 to < 57.5, 57.5 to < 65.5, and 65.5 to < 73.5. Thus, the value 49.5 would go in the second class interval, which has the boundaries 49.5 to < 57.5.

Step 4: Determine Each Class Interval Midpoint

The midpoint is the average value of the class interval. It is often used as a rough estimate of the average case in each interval. It is calculated by adding the lower and upper limits together and dividing by two. Therefore, the midpoints for our intervals are 45.5 [(42 + 49) ÷ 2], 53.5, 61.5, and 69.5.

LO5

Putting the Frequency Table Together

We can now create a frequency table (Table 2.6), using our four class intervals, by recording the number of observations that fall between the class limits of each interval. Usually frequency tables do not include the values for the class limits, boundaries, or midpoints, but we include them here for explanation. Note that the cumulative percentage frequency gives the percentage of observations up to the end of a class. So the cumulative percentage of the third class is 20 + 25 + 25 = 70. We add up all the percentage relative frequencies of the classes up to and including that class.

Based on this frequency table we can report that 70 percent of the time there were fewer than 65.5 thousand (upper class boundary for class interval 58–65) immigrants to Canada, or that 20 percent of the time there were fewer than 49.5 thousand (upper class boundary for class interval 42–49) immigrants.

TABLE 2.6 Sample Frequency Table

Class Interval	Class Limits	Class Boundaries	Midpoints	Frequency (f)	Relative Frequency (f/n)	Percentage Frequency (%f)	Cumulative Percentage Frequency ($c.\%f$)
42–49	42, 49	41.5 to < 49.5	45.5	4	0.20	20	20
50–57	50, 57	49.5 to < 57.5	53.5	5	0.25	25	45
58–65	58, 65	57.5 to < 65.5	61.5	5	0.25	25	70
66–73	66, 73	65.5 to < 73.5	69.5	6	0.30	30	100
Total				20	1.00	100	

Did You Know?

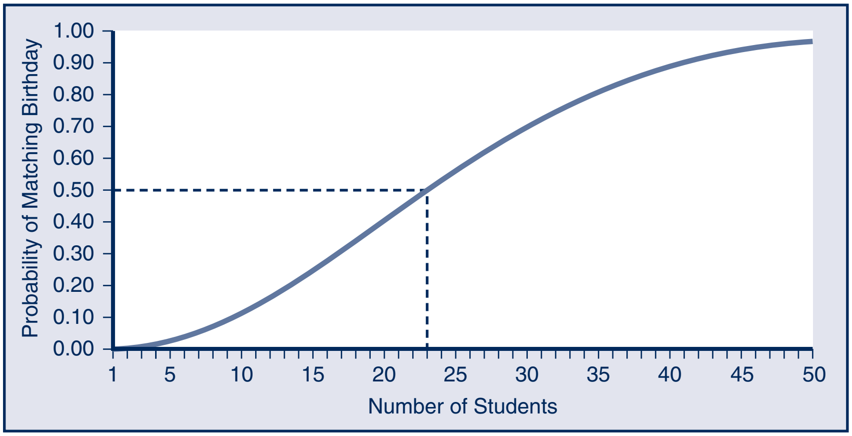
How many students do you need in a classroom to have a 50 percent probability that at least two of them share the same birthday. The answer may surprise you . . . only 23! Ignoring leap-year birthdays, if you randomly select 23 people there is a 50 percent probability that at least two will have the same birthday. In statistics, more precisely probability theory, this is called the birthday problem.

Here's how it works. Imagine you have an empty classroom that you ask students to enter one at a time, after which you estimate the probability that any two students *do not* share a birthday (it's easier to estimate this way). Student 1 enters the classroom. Since the student is alone, he or she has a 100 percent probability of having a unique birthday (365 days ÷ 365 days). Now student 2 enters the classroom. To have a birthday different than student 1, student 2 must have been born on one of the remaining 364 days. The probability of the two students not sharing a birthday is then (365 ÷ 365) × (364 ÷ 365) ≅ 99.73 percent. Student 3 enters the classroom. To have a different birthday than student 1 or 2, student 3's birthday must be any of the 363 days remaining. The probability of the three students not sharing a birthday is then (365 ÷ 365) × (364 ÷ 365) × (363 ÷ 365) ≅ 99.73 percent. Add student 4 and the probability ≅ 98.36, student 5 ≅ 97.28 percent, and so on. After 23 students are in the room, the probability that they do not share a birthday is ≅ 49.27 percent. To estimate the probability that they do share a birthday, we just subtract the probability of not sharing a birthday from 100 percent. For example, the probability of sharing a birthday after five students enter the room is 100 − 97.28 ≅ 2.72 percent, and after 23 students is 100 − 49.27 ≅ 50.73 percent.

The graph in Figure 2.9 shows the how the probability of matching birthdays increases as the number of students in the room increases.

Source: E. H. Mckinney (1966). Generalized Birthday Problem, *The American Mathematical Monthly*, 73 (4), 385–387.

FIGURE 2.9 **Probability of Matching Birthdays**



Cross-Tabulations for Nominal, Ordinal, Interval, and Ratio Data

LO6

Whereas frequency tables display a summary of the distribution of a single variable, cross-tabulations (commonly referred to as cross-tabs) display a summary of the distribution of two or more variables. The difference is that cross-tabs allow you to observe how the frequency distribution of one variable relates to that of one or more other variables. It tabulates the frequencies by the categories or class intervals of the variables being compared.

Cross-tabs can include any combination of nominal, ordinal, interval, and ratio variables. Keep in mind that with interval and ratio variables, it may first be necessary to create class intervals as discussed in the previous section. The steps for creating frequency distributions and class intervals from the previous section are the same for cross-tabs, so we won't repeat them here. Following are some examples using data from Statistics Canada's 2004 Canadian Addiction Survey.[2]

Example 2.1

Two Variables—"Gender" and "Drank Alcohol Before" (Nominal Data)

In this example the nominal variable "Drank Alcohol Before" was measured with the question "Have you ever had a drink?"

TABLE 2.7 **Sample Cross-Tab of Nominal Data**

		Gender		
		Male	Female	Total
Drank Alcohol Before	Yes	806	1,392	2,198
	No	303	703	1,006
Total		1,109	2,095	3,204

With a simple frequency table we would have seen only one variable with the Yes/No or Male/Female frequencies. With a cross-tab we can see the totals from a simple frequency table plus the breakdown of the numbers by the categories Yes/No and Male/Female.

Looking in the Total row at the bottom of the table, we can see that there were 1,109 Males and 2,095 Females, for a total of 3,204 participants. In the Total column on the right, 2,198 said they have had a drink in the past whereas 1,006 said they had not. Within the table itself, we can see that of those that said they have had a drink in the past, 806 were male and 1,392 were female. Furthermore, 303 males and 703 females said they had not had a drink in the past.

Example 2.2

Two Variables—"Gender" and "Drank Alcohol Before" With Percentages (Nominal Data)

To help interpret a cross-tab, percentages are often included, as seen in Table 2.8. Adding percentages to the cross-tab makes it easier to compare categories within the table. However, they can be a bit difficult to read. Look at the shading added to this table. The percentage within "Drank Alcohol Before" highlighted in light shading shows that of those that stated Yes, 36.7 percent were male and 63.3 percent were female. Since these percentages are within the "Drank Alcohol Before" variable, which runs in a row, it totals 100 at the end of the row. Highlighted in dark shading is the percentage in "Gender." Since "Gender" runs in a column, you need to read the percentages down the column. You can see that of male respondents 72.7 percent said they had drank alcohol before, whereas 27.3 percent said they had not. Similarly, of the female respondents, 66.4 percent said they had drank alcohol versus 33.6 percent who said they had not. The totals for the both Male and Female columns add to 100 percent.

TABLE 2.8 **Sample Cross-Tab of Nominal Data Including Percentages**

			Gender		
			Male	Female	Total
Drank Alcohol Before	Yes	Number	806	1,392	2,198
		% within row (Yes*)	36.7	63.3	100
		% within column (Gender)	72.7	66.4	68.6
	No	Number	303	703	1,006
		% within row (Yes*)	30.1	69.9	100
		% within column (Gender)	27.3	33.6	31.4
Total		Number	1,109	2,095	3,204
		% within row (Yes*)	34.6	65.4	100
		% within column (Gender)	100	100	100

Yes* = Drank Alcohol Before

Example 2.3

Two Variables—"Age" and "Access to Marijuana" (Ordinal Data)

Table 2.9 provides a cross-tab of an ordinal variable "Access to Marijuana" with an ordinal variable "Age." The variable "Access to Marijuana" was measured with the survey question "How difficult or easy would it be to get marijuana if you wanted some?" In this example, age was originally a ratio variable with 88 different age values. Six class intervals were created using the process outlined in the previous section. Note that in this case, the last interval includes those aged 66+. We can see that 55.1 percent of the respondents indicated that gaining access to marijuana was very easy. The age category with the highest percentage indicating accessing marijuana to be very easy was those in the 15 to 25 age group.

TABLE 2.9 Sample Cross-Tab of Ordinal Data

			Age						
			15 to 25	26 to 35	36 to 45	46 to 55	56 to 55	66+	Total
Access to Marijuana	Probably impossible		12	16	22	26	42	85	203
		% of Total	0.3%	0.4%	0.5%	0.6%	1.0%	2.0%	4.8%
	Very difficult		1.7	32	37	52	57	63	258
		% of Total	0.4%	0.8%	0.9%	1.2%	1.3%	1.5%	6.1%
	Fairly difficult		40	57	64	61	39	28	289
		% of Total	0.9%	1.3%	1.5%	1.4%	0.9%	0.7%	6.8%
	Fairly easy		167	222	250	265	153	103	1160
		% of Total	3.9%	5.2%	5.9%	6.2%	3.6%	2.4%	27.3%
	Very easy		592	392	476	448	259	174	2341
		% of Total	**13.9%**	9.2%	11.2%	10.5%	6.1%	4.1%	**55.1%**
	Total		828	719	849	852	550	453	4251
		% of Total	19.5%	16.9%	20.0%	20.0%	12.9%	10.7%	100.0%

Comparing the Distribution of Frequencies

LO7

So far we have looked at how to construct and read frequency tables and cross-tabulations. Both are useful when you want to provide a summary of the distribution of responses for one or more variables. However, sometimes researchers want to compare values within variables or compare different variables. For example, suppose we measured the number of females holding executive positions across 100 organizations over the course of five years. How could we determine the percentage change in the number of females in the positions over time? Or, imagine if we collected data on the number of males and females who contracted H1N1 (also called the swine flu). How could we calculate the ratio of illness in males to females? Finally, suppose we collected data on bicycle thefts in three different cities each with different size populations. How could we compare the theft rates across the cities while taking into account the differing sizes of population? For these questions, we need to use percentage change, ratios, and rates. To demonstrate each, a frequency table or cross-tab is provided to assist you in understanding where the numbers are coming from.

Percentage Change

The **percentage change** is a fraction out of 100 that indicates the relative change in a variable from one time period to another.

In the previous examples, we dealt with situations where we had data from only one time period. However, social scientists are often interested in examining variables across different time periods. When we have data that spans different time periods (e.g., two years), we can calculate the change from one year to another and report this change as a percentage. We call this a **percentage change**.

TABLE 2.10
Number of Marriages

Frequency of Marriages (f)	
2007	148,296
2008	148,831

Source: ****Important: data for 2005 to 2008 is available from Health Statistics Division (hd-ds@statcan.gc.ca) upon request only, as this program is discontinued and data will not be available on CANSIM anymore and neither as Summary Data Table, Marriages by province and territory is discontinued, http://www40.statcan.ca/l01/cst01/famil04-eng.htm.

To calculate percentage change from one time period to another, we use the following equation:

$$p = \frac{f_{time_2} - f_{time_1}}{f_{time_1}} \times 100 \qquad \textbf{(2.3)}$$

where: f_{time_1} = frequency of a specific response at time 1
f_{time_2} = frequency of a specific response at time 2

Table 2.10 provides a frequency table of Statistics Canada's estimated number of marriages for 2007 and 2008. The percentage change from 2007 to 2008 is calculated as:

$$\text{Percentage change} = \frac{\#\text{ in 2008} - \#\text{ in 2007}}{\#\text{ in 2007}} \times 100 \qquad \textbf{(2.4)}$$

$$= \frac{148{,}831 - 148{,}296}{148{,}296} \times 100 = 0.36$$

In this example we can say that there was a slight increase in the number of marriages between 2007 and 2008 of 0.36 percent, which is less than a 1 percent increase. However, according to Statistics Canada the population of Canada was 32,932,000 in 2007 and 33,327,000 in 2008, which means that the percentage change in the population from 2007 to 2008 was 1.19 percent. If the population increased you might ask if the marriage rate really increased. To answer this, we have to calculate the difference using rates. We'll cover this situation in the Rates section of this chapter.

A **ratio** is a comparison of two values of a variable based on their frequency.

Ratios

A **ratio** is a comparison of two values of a variable based on their frequency. You may have heard of or read reports that include ratios. For example, your school

For Your Information

Frequencies, proportions, percentages, percentage change, ratios, and rates can be calculated on variables with nominal, ordinal, interval, and ratio levels of measurement. The main difference is the number of values the variable can possibly have. While "Gender" has two, "Weight" may have many differing values. How you measure the variable determines how many potential values the variable may have.

may have a student-to-faculty ratio of 22:1 (22 students for every one faculty member). Or if you make french toast you might decide to use a ratio of 1/3 cups of milk for every one egg. To calculate a ratio comparing two values of a variable, we divide the first value of interest by the second value of interest. The equation is as follows:

$$Ratio = \frac{f_{v1}}{f_{v2}} \qquad \textbf{(2.5)}$$

where: f_{v1} = frequency of the first value to be compared
f_{v2} = frequency of the second value to be compared

For example, consider Table 2.11, which provides the number of motor vehicle accident deaths for 2004 as reported by Statistics Canada.

TABLE 2.11
2004 Motor Vehicle Accident Deaths

Source: "2004 Motor Vehicle Accident Deaths," adapted from Statistics Canada publication *Health Reports,* Catalogue 82-003-XIE2008003, Vol. 19, No. 3, http://www.statcan.gc.ca/pub/82-003-x/2008003/article/10648-eng.pdf.

	Frequency (f)	Relative Frequency (f/n)
Males	2,035	0.708
Females	840	0.292
Total	2,875	1.00

Here we can see that more males are killed in motor vehicle accidents than females. However, how many males does that represent per female? Using equation 2.5 we see that:

$$\begin{aligned} Ratio &= \frac{f_{v1}}{f_{v2}} \qquad \textbf{(2.6)} \\ &= \frac{\text{\# of male motor vehicle deaths}}{\text{\# of female motor vehicle deaths}} \\ &= \frac{2{,}035}{840} = 2.42 \end{aligned}$$

We can now say that in 2004 there were 2.42 males killed in a motor vehicle accident for every one female killed in a motor vehicle accident (2.42:1). Conversely, if we wanted to know the ratio of females to males, we would just switch the numerator and denominator:

$$\frac{840}{2{,}035} = 0.41{:}1$$

Meaning that there were approximately 0.41 females killed in motor vehicle accidents per one male killed in motor vehicle accidents.

EXERCISE 2.1

Previously we looked at the cross-tab of the variables "Gender" and "Drank Alcohol Before" from the 2004 Canadian Addiction Survey. This cross-tab is reproduced in Table 2.12. Using equation 2.5 try to determine the following ratios.

TABLE 2.12
Cross Tabulation of "Gender" and "Drank Alcohol Before"

		Gender		
		Male	**Female**	**Total**
Drank Alcohol Before	Yes	806	1,392	2,198
	No	303	703	1,006
Total		1,109	2,095	3,204

Question 1: What is the ratio of those respondents who have previously drank alcohol to those that stated they have not?

Answer 1: There are 2.18 respondents who have previously drank alcohol to every one respondent who has not. (2,198 ÷ 1,006 = 2.18).

Question 2: What is the ratio of females who state they have not previously drank alcohol to males who state they have not previously drank alcohol?

Answer 2: There are 2.32 females who state they have not previously drank alcohol to every one male who stated they have not previously drank alcohol.

Rates

A **rate** is the frequency with which a phenomenon occurs relative to a population size or time unit.

Ratios are useful when the values being compared are in the same units. For example, the number of cars sold by Salesperson A versus Salesperson B; the average heart rate for participants in Group A versus Group B; or the number of speeding tickets for different age groups of drivers. However, if you need to compare values where other factors affect those values, then **rates** are more useful. For example, it wouldn't make a lot of sense to compare the actual number of new housing starts (construction of new homes) in British Columbia to New Brunswick because the population sizes are different. So while these numbers may be useful to a researcher by themselves, comparisons become distorted when we compare regions that have different poulations sizes. To adjust for poulation size we would report the new housing starts per person. That is, we use the ratio of total new housing starts in a region to the population size of that region. In doing so we obtain the rates of new housing starts that can then be used to compare regions of different population sizes. To calculate rates, we use equation 2.7.

Note that "rate" is often referred to as "crude rate" because the rate does not take into consideration the structure of the population (such as age or gender differences in the population). For simplicity, we will just use the term "rate."

$$Rate = \frac{\textit{Number of events for the population of interest}}{\textit{Total population of the population of interest}} \times 10{,}000 \quad \textbf{(2.7)}$$

TABLE 2.13
2009 Housing Starts and Populations for British Columbia and New Brunswick

	Housing Starts	Population
British Columbia	16,077	4,455,200
New Brunswick	3,521	749,500

Sources: Adapted from Statistics Canada website, http://www40.statcan.gc.ca/l01/cst01/demo02a-eng.htm, extracted May 18, 2011, and Canada Mortgage and Housing Corporation (CMHC), Starts and Completions Survey, 2009.

Notice in the equation that we multiple by 10,000. We do this in order to avoid small decimals. For example, a rate of 0.025 car sales per person is a more difficult to interpret than 250 car sales per 10,000 people. You can multiply by whatever number makes sense as long as you report it properly. So mulitplying by 1,000 instead of 10,000 gives you 25 car sales per 1,000 people. Taking the rate without mulitplying by a number (or multiplying by 1) gives you the rate per person. When the rate is per person, you'll often hear it called "per capita."

Continuing with our housing starts example, Table 2.13 provides the 2009 housing starts and populations for British Columbia and New Brunswick, according to Statistics Canada. Looking at the actual numbers it appears that British Columbia is growing more (based on new housing starts) than New Brunswick. However, if we take population into account, and create a rate of new housing starts per 10,000 people, we can equitably compare the two provinces.

British Columbia:

$$\begin{aligned} rate &= \frac{16{,}077}{4{,}455{,}200} \times 10{,}000 \\ &= 0.003609 \times 10{,}000 \\ &= 36.09 \end{aligned} \qquad \textbf{(2.8)}$$

New Brunswick:

$$\begin{aligned} rate &= \frac{3{,}521}{749{,}500} \times 10{,}000 \\ &= 0.004698 \times 10{,}000 \\ &= 46.98 \end{aligned}$$

Based on our calculation in equation 2.8, in 2009, New Brunswick had more new housing starts per person than British Columbia. In fact, New Brunswick had 46.98 new housing starts per 10,000 people whereas British Columbia had 36.09 per 10,000 people.

In our previous example, we compared one phenomenon (new housing starts) in two different populations (British Columbia and New Brunswick). However, with ratios we can compare a phenomonon across a number of different populations. Suppose we want to compare provinces and territories on the volume of wine purchased per person. If we compare only total volume of wine purchased per province and territory we would see that the top three purchasers (by volume) of wine would be Ontario (137,737,000 litres), Quebec (128,614,000 litres), and

TABLE 2.14
Wine Purchased by Per Person Rates

Source: "Wine Purchased by Per Person Rates," adapted from Statistics Canada CANSIM Database, http://www5.statcan.gc.ca/cansim/home-accueil?lang=eng, CANSIM Table 183-0006, extracted May 18, 2011.

Region	Volume per Person
Newfoundland and Labrador	6.1
Saskatchewan	7.3
Northwest Territories and Nunavut	8.8
Prince Edward Island	9.0
Manitoba	9.1
New Brunswick	9.6
Nova Scotia	10.1
Ontario	13.2
Canada	15.0
Alberta	15.9
British Columbia	17.9
Quebec	20.1
Yukon	21.0

British Columbia (62,805,000 litres). This would make sense because these are the three largest provinces by population. Table 2.14 contains the volume of wine (in litres) purchased per person (15 years of age or older) across the provinces and territories of Canada for the year 2007. Here we can see that when considering per person rate, Yukon, Quebec, and British Columbia are the top three purchasers.

Recall in an earlier section we found that there was a slight increase in the number of marriages between 2007 and 2008 of 0.36 percent, but during the same time the population also increased by 1.19 percent. We left that example wondering if the marriage rate really increased. Now that we understand percentage change and rate, we can combine the two to answer that question. Table 2.15 provides the information regarding numbers of marriages and population per year.

First, we need to estimate the rate of marriages for both 2007 and 2008 so that we can take the difference in population size in to account.

Rate of Marriages per 1,000 people in 2007:

$$\begin{aligned} rate &= \frac{148{,}296}{32{,}932{,}000} \times 1{,}000 \\ &= 0.00450 \times 1{,}000 \\ &= 4.50 \end{aligned} \tag{2.9}$$

TABLE 2.15
Number of Marriages and Population Per Year

Source: ****Important: Data for 2005 to 2008 is available from Health Statistics Division (hd-ds@statcan.gc.ca) upon request only, as this program is discontinued and data will not be available on CANSIM anymore and neither as Summary Data Table, Marriages by province and territory is discontinued, http://www40.statcan.ca/l01/cst01/famil04-eng.htm.

Year	Number of Marriages	Population
2007	148,296	32,932,000
2008	148,831	33,327,000

Rate of Marriages per 1,000 people in 2008:

$$\begin{aligned} rate &= \frac{148{,}831}{32{,}327{,}000} \times 1{,}000 \\ &= 0.00447 \times 1{,}000 \\ &= 4.47 \end{aligned}$$

We can see that there were 4.50 marriages per 1,000 people in 2007 and 4.47 marriages per 1,000 people in 2008. We can now use equation 2.10 to calculate the percentage change in the rates.

$$\begin{aligned} percentage\ change &= \frac{rate\ in\ 2008 - rate\ in\ 2007}{rate\ in\ 2007} \times 100 \qquad \textbf{(2.10)} \\ &= \frac{4.47 - 4.50}{4.50} \times 100 \\ &= -0.67 \end{aligned}$$

We can now say that although marriages increased by 0.36 percent from 2007 to 2008, the population also increased during that time by 1.19 percent. As a result, when you take the population change into account, the rate of marriages decreased by 0.67 percent (decreased because the number is negative).

Graphing Data

A pictorial representation of data efficiently and effectively transmits information. Pie charts, bar charts, frequency polygons, cumulative percentage frequency polygons, and histograms all help present the information contained in the data.

LO8

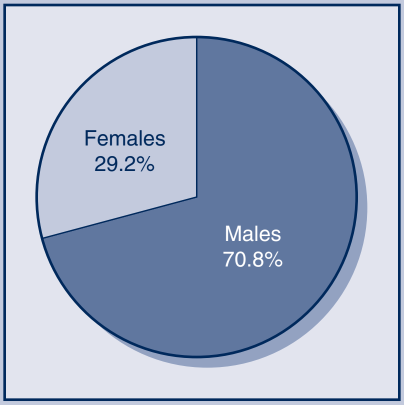
Pie Charts and Bar Charts, For Nominal and Ordinal Data

A **pie chart** displays the distribution of a variable out of 100 percent, where 100 percent represents the entire pie.

Since you have likely already been exposed to pie charts and bar charts during your education, we won't spend a lot of time on them. Pie charts and bar charts are useful for graphically displaying the frequency distribution of nominal and ordinal level data as they can be easily constructed to show the differences in categories within a variable. A **pie chart** displays the distribution of a variable out of 100 percent, where 100 percent represents the entire pie. Either frequency or percentage frequency may be used in constructing the chart. As an example, the pie chart and table in Figure 2.10 displays the frequency of motor vehicle deaths by gender for 2004.

Pie charts are particular useful for nominal and ordinal variables, as the categories are used to separate the pie into the appropriate pieces. While they could be used for interval and ratio level variables using class intervals, frequency polygons and histograms (discussed in the next section) tend to better represent the data.

FIGURE 2.10 Example of a Pie Chart



	Frequency	Percentage Frequency
	(*f*)	(%*f*)
Males	2,035	70.8
Females	840	29.2
Total	2,875	100.0

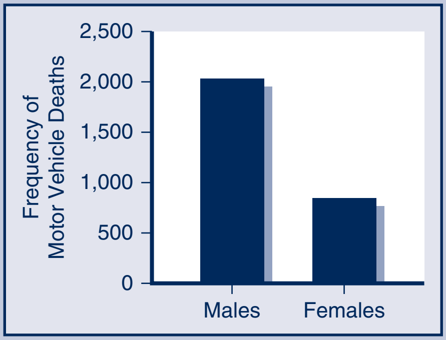
Source: "Example of a Pie Chart—'Motor vehicle accident deaths, 1979 to 2004,'" adapted from Statistics Canada publication *Health Reports,* Catalogue 82-003-XIE, Vol. 19, No. 3, http://www.statcan.gc.ca/pub/82-003-x/2008003/article/10648/5202440-eng.htm.

A **bar chart** displays the frequency of a variable with the variable categories along the *x*-axis and the variable frequencies on the *y*-axis.

Similar to a pie chart, a **bar chart** displays the frequency of a variable with the categories of the variable along the *x*-axis and the frequency of the variable on the *y*-axis. Figure 2.11 displays the same motor vehicle accident data in bar chart format.

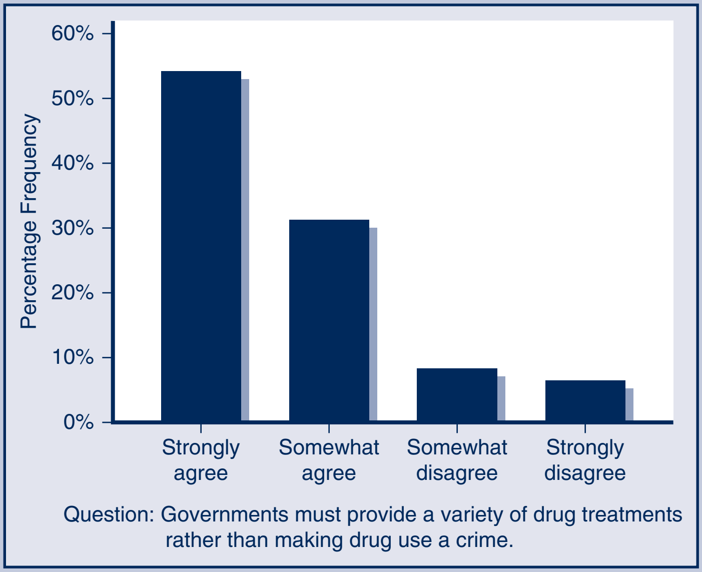
As is the case with pie charts, bar charts are best suited to nominal and ordinal data as the categories of the variables can be easily transferred to the *x*-axis of the chart. However, they can sometimes be useful for displaying interval and ratio level variables using class intervals. Figure 2.12 is an example of interval data from a survey item, measured using a 4-point Likert-type scale (four response options), in the 2004 Canadian Addiction Survey. Figure 2.13 provides a ratio example where class intervals are used to categorize age in response to a question about the ease of access to marijuana.

FIGURE 2.11 Example of a Bar Chart



	Frequency	Percentage Frequency
	(*f*)	(%*f*)
Males	2,035	70.8
Females	840	29.2
Total	2,875	100.0

FIGURE 2.12
Example of a Bar Chart For a 4-Point Likert-type Scale



	Frequency	Percentage Frequency	Cumulative Percentage Frequency
Strongly agree	2,433	54.1%	51.1%
Somewhat agree	1,403	31.2%	85.3%
Somewhat disagree	373	8.3%	93.6%
Strongly disagree	290	6.4%	100.0%
Total	4,499	100.0%	

LO9

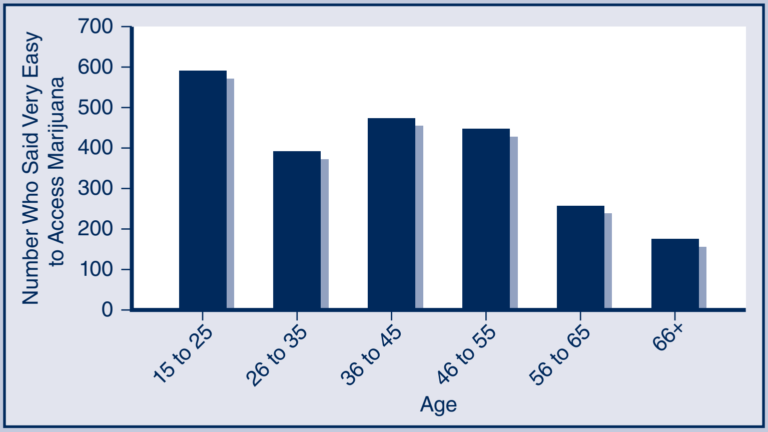
Frequency Polygon and Cumulative Percentage Frequency Polygon, For Interval and Ratio Level Data

A **frequency polygon** is a line graph of the frequency of interval or ratio data.

A **frequency polygon** is a line graph of the frequency distribution of interval or ratio data and is constructed by placing the class intervals on the *x*-axis and the frequencies (or percentage frequencies) on the *y*-axis. They are useful for graphically displaying the shape of the frequency distribution. Figure 2.14 is an example of a frequency polygon with its associated frequency table. This data represents a sample of 209 respondents, from Statistics Canada's 2004 Canadian Addiction Survey, from the ages of 11 to 21 who had previously used or continue to use marijuana, cannabis, or hashish. The frequency polygon and table represent the age at which the respondent began using marijuana, cannabis, or hashish. Looking at the frequency polygon, you can see that the largest frequency is age 16 (the highest point in the line).

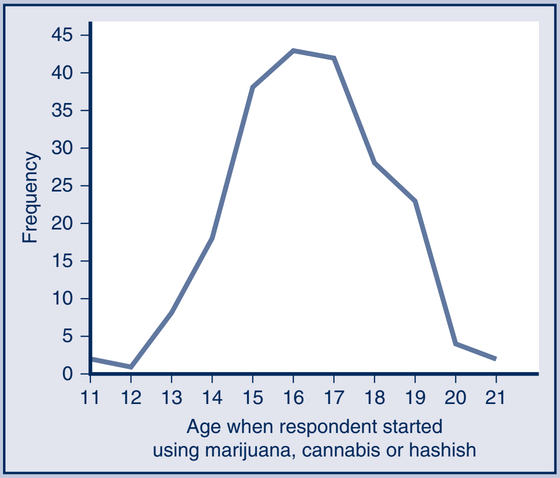
FIGURE 2.13

Example of a Bar Chart With Class Intervals



Age	Frequency of Very Easy to Access Marijuana	Percentage Frequency	Cumulative Percentage Frequency
15 to 25	592	25%	25%
26 to 35	392	17%	42%
36 to 44	476	20%	62%
46 to 55	448	19%	81%
56 to 65	259	11%	92%
66+	174	8%	100%
Total	2,341	100%	

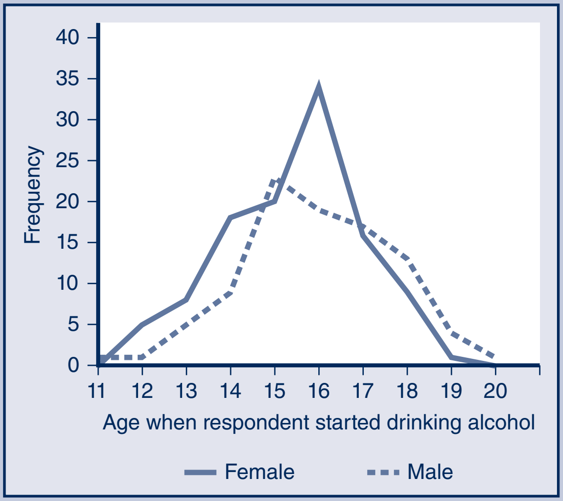
FIGURE 2.14 Frequency Polygon



Age	Frequency (*f*)	Percentage Frequency (%*f*)	Cumulative Percentage Frequency (*c*.%*f*)
11	2	0.96	0.96
12	1	0.48	1.44
13	8	3.83	5.26
14	18	8.61	13.88
15	38	18.18	32.06
16	43	20.57	52.63
17	42	20.10	72.73
18	28	13.40	86.12
19	23	11.00	97.13
20	4	1.91	99.04
21	2	0.96	100.00
Total	209	100.00	

Source: Permission granted by CAS via the Data Liberation Initiative.

FIGURE 2.15 **Frequency Polygon by Gender**



Age	MALES Frequency (*f*)	FEMALES Frequency (*f*)
11	1	0
12	1	5
13	5	8
14	9	18
15	23	20
16	19	34
17	17	16
18	13	9
19	4	1
20	1	0
Total	93	111

Source: Permission granted by CAS via the Data Liberation Initiative.

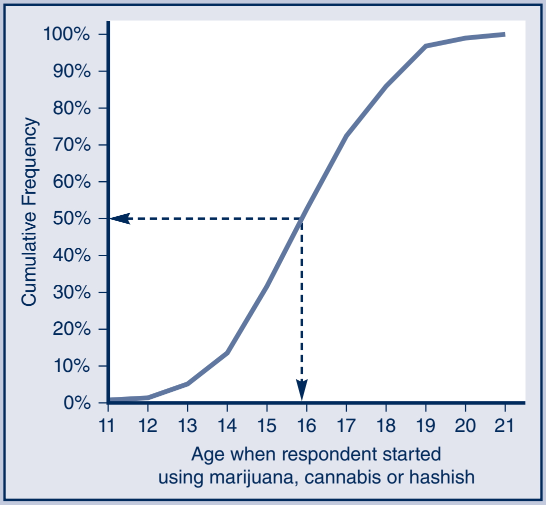
Frequency polygons can also be used to compare the distribution of a variable across groups of respondents. For example, the 2004 Canadian Addiction Survey asked respondents to indicate at what age they began to drink alcohol (if ever). Using a sample of males and females from the ages of 11 to 20, Figure 2.15 compares the frequency distributions (males versus females) of age in which the respondents stated they began drinking alcohol. You can see from the frequency polygons, and respective frequency tables, that for males the highest age frequency is 15 years of age and for females is 16 years of age.

A **cumulative frequency polygon** is a frequency polygon that graphs the cumulative percentage frequency column in a frequency table.

A **cumulative percentage frequency polygon** is a frequency polygon that graphs the cumulative percentage frequency column in a frequency table. Similar to frequency polygons, they can be used for comparing frequencies of a variable across groups. Figure 2.16 provides the cumulative percentage frequency of the example used in Figure 2.14. You can see that approximately 50 percent of the respondents had started using marijuana, cannabis, or hashish by approximately 16 years of age.

The examples in figures 2.14 and 2.16 have class intervals with a width of one year, which makes them a little easier to understand. However, you could have class intervals with greater widths. Say you created the four class intervals 11–13, 14–16, 17–19, and 20+. In this case your *x*-axis would include the numbers from the upper class limit of each interval, 13, 16, 19, and 20+. You would then plot the frequency up to the upper limits of each interval. Therefore, 13 would be 11, 16 would be 99, 19 would be 93, and 20+ would

FIGURE 2.16 **Cumulative Frequency Polygon**



Age	Frequency (*f*)	Percentage Frequency (%*f*)	Cumulative Percentage Frequency (*c.*%*f*)
11	2	0.96	0.96
12	1	0.48	1.44
13	8	3.83	5.26
14	18	8.61	13.88
15	38	18.18	32.06
16	43	20.57	52.63
17	42	20.10	72.73
18	28	13.40	86.12
19	23	11.00	97.13
20	4	1.91	99.04
21	2	0.96	100.00
Total	209	100.00	

Source: Permission granted by CAS via the Data Liberation Initiative.

be 6. The same applies to cumulative frequency polygons, except that you would use the cumulative frequencies.

LO10

A **histogram** is a plot of the frequency of interval or ratio data.

Histograms

A **histogram** is a plot of the frequency of interval or ratio data. Histograms are useful ways of graphically representing interval and ratio variables because they are able to show the continuous nature of the data without necessarily creating class intervals.

Statistics Canada's 2004 Canadian Addiction Survey provides the age that participants stated they had their first alcoholic drink. Figure 2.17 provides two histograms of the variable "Age of First Alcohol Drink." In both histograms, the *x*-axis represents the ages from 10 to 21. In the histogram on the left, the *y*-axis provides the frequency with which each age is observed in the data. In this histogram you can see that the most frequently occurring ages are 16 and 18. The histogram on the right provides the same information, only this time *y*-axis is percentage frequency as opposed to frequency. Again, we can see that the most frequently occurring ages are 16 and 18, but this time we also see that these ages account for approximately 36 percent of the cases (approximately 18 percent each).

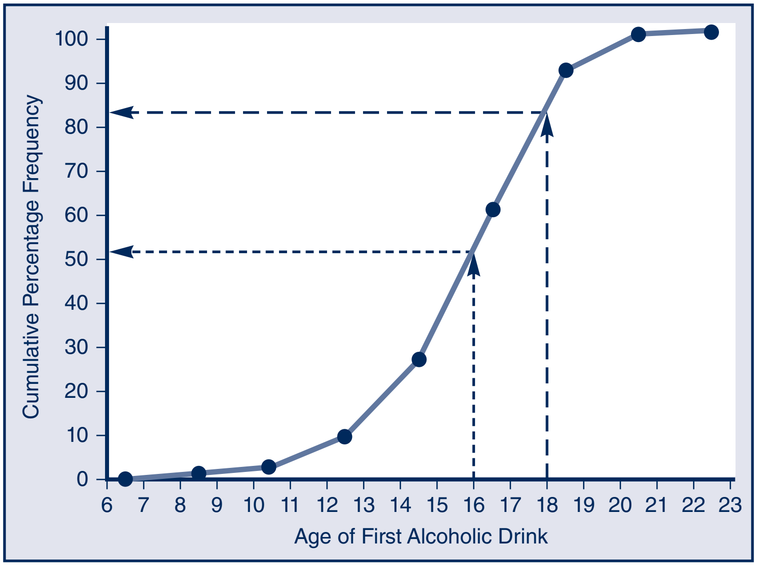
Although it is possible to use histograms with class intervals, doing so loses some of the interpretive value of the histogram. Consider the two histograms in Figure 2.18. On the left is a histogram representing the age of 10,060 individuals from the ages of 20 to 60 who participated in Statistics Canada's 2004 Canadian

For Your Information

A cumulative percentage frequency polygon (line chart) is useful for answering questions, such as "What percentage of participants had their first alcoholic drink at 16 years of age and at 18 years of age?"

To answer this, start at age 16 on the *x*-axis and move upwards (following the red arrow) until you reach the frequency line. Then move across to the *y*-axis to find the percentage. In interpreting this you might say that of those participants who have previously drank alcohol, approximately 52 percent had their first alcoholic drink by the age of 16. Similarly, if you look at the age 18 (green arrow) you might say that of those participants who have previous drank alcohol, approximately 84 percent had their first alcoholic drink by the age of 18. By subtracting the two (84 – 52) we can say that approximately 32 percent had their first alcoholic drink between the ages of 16 and 18.

Source: Permission granted by CAS via the Data Liberation Initiative.



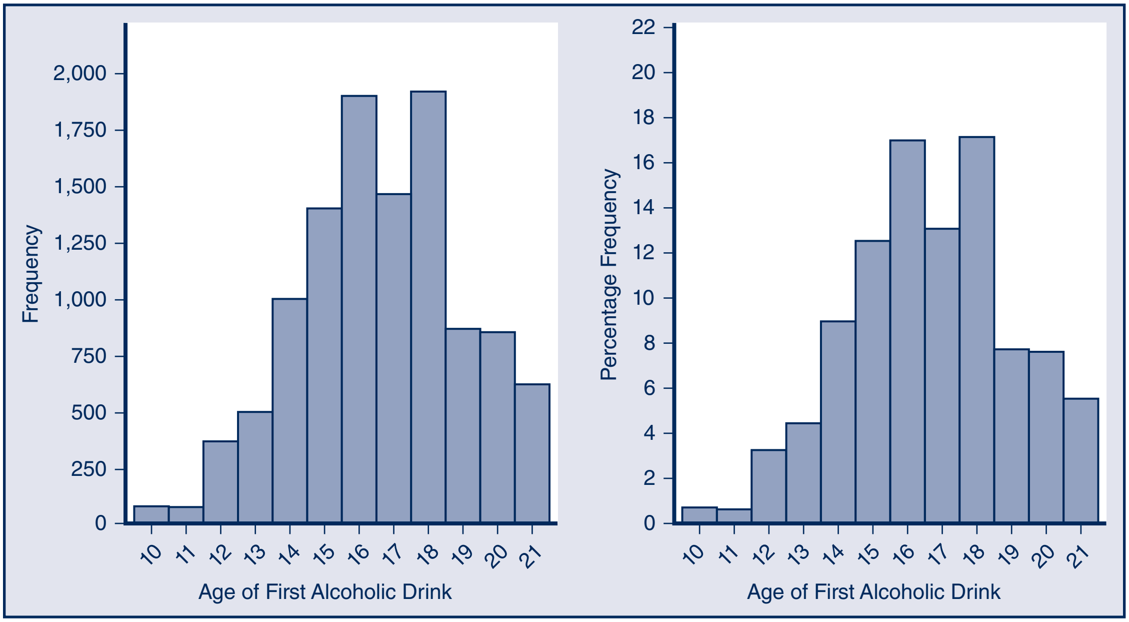
Addiction Survey. In this example, class intervals were constructed to reduce the variable from a ratio level to four classes. On the right is a histogram of age of the same 10,060 individuals, only this time class intervals were not used. You can see immediately that the histogram on the right provides you with a lot more detailed information than the one on the left, even though the same data and same number of cases are included.

A **stem-and-leaf plot** is a plot of the frequency of interval or ratio data similar to the histogram but more informative since it provides actual data values.

Stem-and-Leaf Plots

A **stem-and-leaf plot** is similar to a histogram in that it provides a graphical representation of the frequency of interval or ratio data. Where it differs from

FIGURE 2.17 **Age of First Alcoholic Drink**



the histogram is that the data within the frequency table is included in the plot. Suppose you randomly select 20 high school students and record the number of text message each individual sends in one day. Table 2.16 represents the number of text messages for one day.

FIGURE 2.18
Histogram With and Without Class Intervals

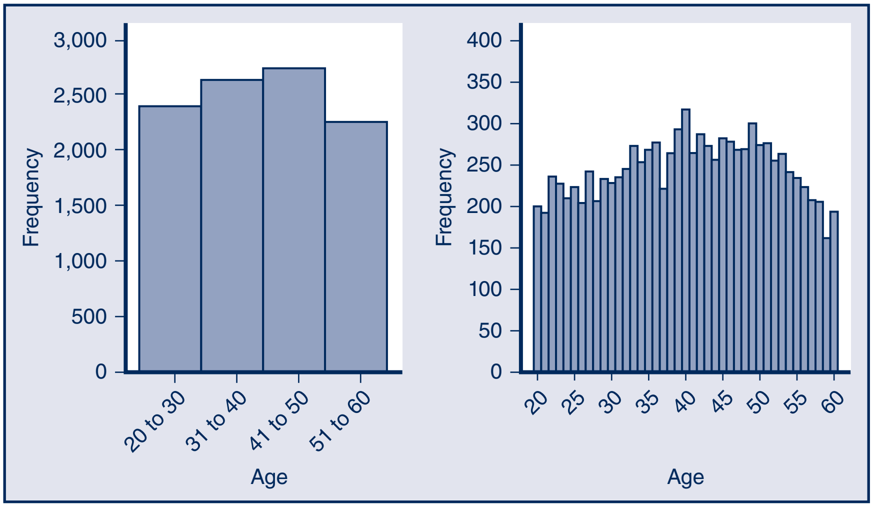


TABLE 2.16 Sample Text Message Data

13	21	22	31	35	39	39	42	43	46
46	47	48	50	51	58	60	67	70	79

Based on what we know so far, we could create a frequency table with seven class intervals and from there create a histogram (see Figure 2.19). While both the frequency table and histogram are useful ways to display the data, the drawback is that the actual values of the data are not shown. For example, from both the frequency table and the histogram we can see that there are four values in the interval 30 to 39, but without seeing the actual data, we don't know what those values are.

Figure 2.20 provides the stem-and-leaf plot for the same data. The stem represents the value(s) on the left side of each individual number. In this case, the stem represents the tens column of the number. So for the values in the 10 to 19 interval, 1 is the stem, for the values in the 20 to 29 interval, 2 is the stem, and so on. If our values were more than two digits, we could adjust the stem to represent hundreds (as in 1 for 100) or thousands (3 for 3,000). The value you set for the stem is based on your judgment of the best way to present the data.

The leaf represents the value(s) on the right side of each number. In this case, the leaf represents the ones column of the number. For the value 13, 1 is the stem and 3 is the leaf. The place value of the numbers representing the leaf (i.e., ones, tens, hundreds, etc.) depends on the place value you use for the stems.

Looking at Figure 2.20, we can see that the value 35 is shown with 3 in the stem and 5 in the leaf, while 51 is shown with 5 in the stem and 1 in the leaf. All values with the same stem are included on the same line, which is why you see 3 as the stem and 1 5 9 9 in the leaf, representing the values 31, 35, 39, and 39.

FIGURE 2.19 Frequency Table and Histogram of Text Message Data

Frequency of Text Messages

Class Interval	Frequency (*f*)
10 to 19	1
20 to 29	2
30 to 39	4
40 to 49	6
50 to 59	3
60 to 69	2
70 to 79	2

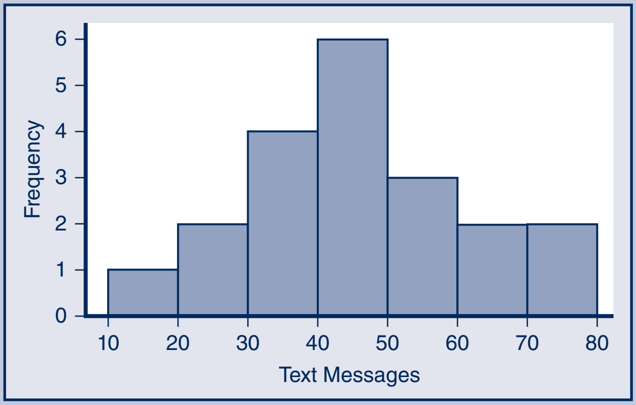
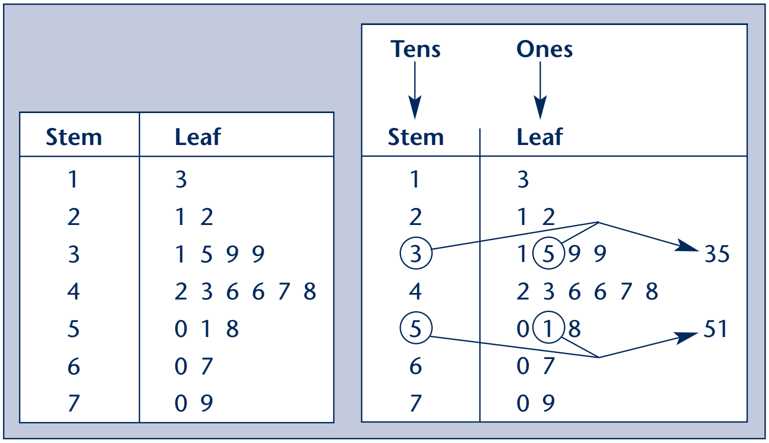


FIGURE 2.20
Stem-and-Leaf Plot of Text Message Data



Stem	Leaf
1	3
2	1 2
3	1 5 9 9
4	2 3 6 6 7 8
5	0 1 8
6	0 7
7	0 9

If you look at the histogram and the stem-and-leaf plot you will see that the shape of the leaf corresponds to the shape of the histogram. This is where you can see the value of the stem-and-leaf plot. It is a good way of showing data distribution (like a histogram), but without losing any of the information regarding the values of each number.

EXERCISE 2.2

Using the following numbers, create a stem-and-leaf plot.

41	45	50	52	54	54	61	67	67	68
69	71	75	84	86	87	90	91	92	92

Boxplots

A **boxplot** (also known as a box and whisker plot) is a graphical summary of the data based on percentiles.

A **boxplot** (also known as a box and whisker plot) is a graphical summary of the data based on percentiles. Figure 2.21 is a boxplot of the text messaging data from Table 2.16. The box represents the distribution of the data between the 25th and 75th percentile. The light line in the middle represents the median (50th percentile). The lines coming out of the box (also known as whiskers) extend to the lowest and highest value in the data, which provides you with the range. As we can see in Figure 2.21, the lowest value is 13 and the highest is 79. The 75th percentile sits at 56.25 and the 25th percentile is 36. Finally the median value, which represents the 50th percentile, is 46.

One advantage of boxplots is that you can compare the data distribution of multiple groups. For example, suppose we gathered the number of text messages sent per day from 20 students in School A and 20 students in School B. We could create two boxplots to compare them (Figure 2.22).

FIGURE 2.21
Example of a Boxplot

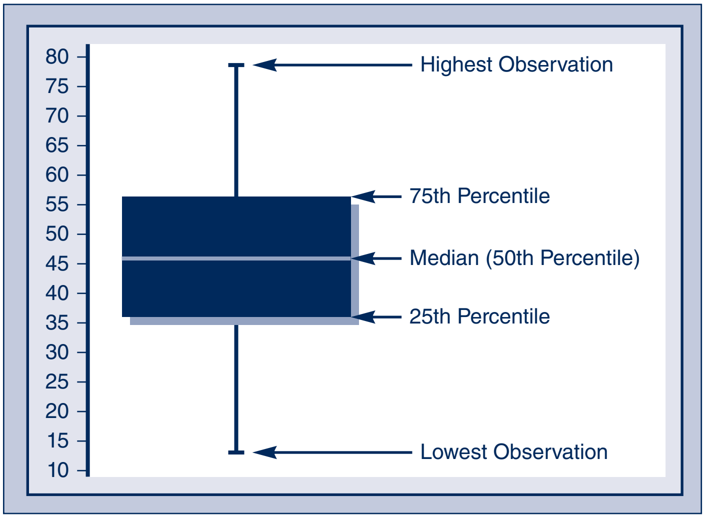
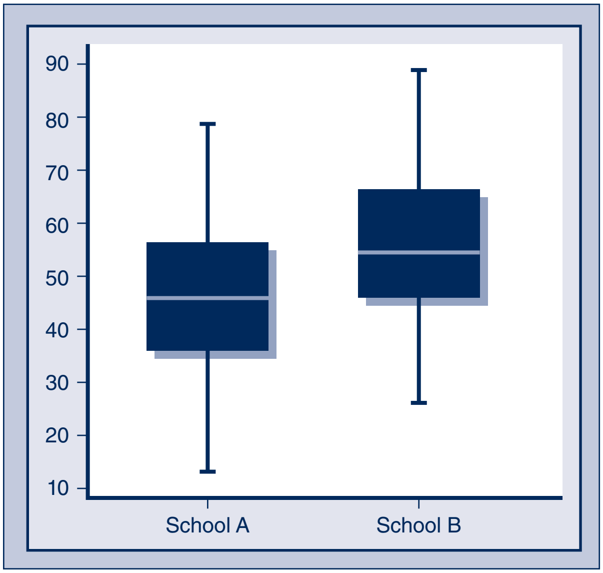


FIGURE 2.22
Example of a Boxplot Comparing Two Groups



Conclusion

In this chapter we discussed ways in which you can describe your data using frequencies, cross-tabulations (or cross-tabs), and graphs. Frequency tables are useful in displaying the distribution of scores on a variable and often include relative, percentage, and cumulative frequency information to aid in the interpretation of the data. Frequency tables can be used with variables assessed at all levels of measurement (nominal, ordinal, interval, and ratio data), although keep in mind that with interval and ratio data, you often need to construct class intervals first. Cross-tabs are used when you want to display summary information about two or more variables. They allow you to observe how the distribution of one variable relates to that of another variable. Similar to frequency tables, cross-tabs work with nominal, ordinal, interval, and ratio data.

When you need to compare summary results, using percentage change, ratios, or rates is helpful. Percentage change allows you to see the difference in the variable across different time periods. Ratios allow you to compare two values of a variable based on their frequency. Rates allow you to compare variables where population size or time needs to be accounted for.

Graphical representation of data is also useful. You now know how to present data using pie and bar charts, frequency and cumulative frequency polygons, histograms, stem-and-leaf plots, and boxplots. Each type of graph has a different purpose, and their use depends on the information you need to convey. For example, you could use a frequency polygon to show the dollar amounts of social assistance provided over the course of 10 years, and a pie chart to show percentage of the total amount provided to males versus females.

In the next chapter, we will look at descriptive statistics as another way of describing data. This will provide you with the first look at how the level of measurement influences the types of statistical analysis that can be used.

Key Chapter Concepts and Terms

Empirical data, 37
Frequency, 39
Frequency distribution, 40
Relative frequency, 40
Percentage frequency, 41
Cumulative percentage frequency, 41
Range, 42
Class interval, 44
Class width, 44
Class limits, 44
Percentage change, 51
Ratio, 52
Rate, 54
Pie chart, 57
Bar chart, 58
Frequency polygon, 59
Cumulative frequency polygon, 61
Histogram, 62
Stem-and-leaf plot, 63
Boxplot, 66

Frequently Asked Questions

1. Does per capita represent a specific number of people? Or does it simply mean proportionately? For example, does per capita automatically mean per 1,000 people or does it vary?

   Per capita is translated as ‘per head.’ So per capita spending is total spending divided by the number of people. However, per capita murders, or even per capita for all crimes, can be a very small number. In those cases the per capita rate is changed to be per 1,000 people or per 10,000 people for ease. So if there are two murders per 1,000 people, we could write 0.002 as the per capita rate, but for many people, 2 per 1,000 is easier to understand.

2. Can you give me an example of the difference between a frequency and relative frequency?

   Frequency is the count of observed values in the variable. For example, in the May 2, 2011 federal election, the frequency of seats won by political party was Conservatives 167, NDP 102, Liberals 34, Bloc Québécois 4, and Green Party 1.

   Relative frequency is the frequency divided by the total number of observations (308). So the relative frequency is Conservatives 0.5422 (167 ÷ 308), NDP 0.3312 (102 ÷ 308), Liberals 0.1104 (34 ÷ 308), Bloc Québécois 0.0129 (4 ÷ 308), and Green Party 0.0033 (1 ÷ 308).

3. What makes the midpoint of a frequency histogram important to know?

   If all we have is a frequency table, then we can use this information to generate approximate values for the sample mean or sample standard deviation. To do so, we assume that the data points are evenly dispersed over an interval and the midpoint represents the average of the values in that interval.

4. When would we use a boxplot instead of a histogram?

   A boxplot is useful when you want to provide information about the range and percentiles of your data. It is also useful for comparing the distribution of a variable across groups. When you are asked to display information about the percentiles, boxplots are the best option. When you want to show information about the frequency of observations per class interval, then histograms are the best choice. Remember: boxplots don't provide information about the class intervals.

5. If we have class intervals in a frequency table, why do we need class boundaries?

   Class intervals tell you what the start and end points are for groups of data in the frequency. Say you have a class interval for age that ranges from 18 to 25 and 26 to 33. Class boundaries tell you where to put numbers when the values fall on the edge of these limits. For example, what if a participant was 25 years and 2 months old? Where would you put this individual since he or she is not 25 and but not yet 26? The class boundary for the 18 to 25 class interval would likely be 17.5 (17 years and 6 months) to less than 25.5 (25 years and 6 months). So anyone older than or equal to 17 years and 6 months old and younger than 25 years and 6 months old would be placed in the class interval 18 to 25.

Research Example:

In 2010, Dr. Carole Orchard, of the University of Western Ontario, and her three colleagues published their paper "Integrated nursing access program: An approach to prepare aboriginal students for nursing careers" in the *International Journal of Nursing Education and Scholarship*. One of the purposes of their paper was to describe the progress of a national program to assist Canadian Aboriginal students in meeting the requirements for admissions to university nursing programs.

As part of the program student participants completed a survey of their readiness to engage in self-directed learning. The survey, with the acronym SLDRS-NNES, consisted of 34 questions measured using a 5-point Likert-type scale. Dr. Orchard and colleagues detailed the results of the 35 student participants ranging from 34 (low) to 170 (High) in a frequency table similar to the one below.

Frequency Table of SLDRS-NNES Scores

Score	*f*	*f*/*n*	%*f*	c.%*f*
34–106	3	0.086	8.6	8.6
107–145	29	0.828	82.8	91.4
146–170	3	0.086	8.6	100.0
Total	35	1.000	100.0	

Source: Orchard, Carole A., Paula Didham, Cathy Jong, and June Fry (2010). "Integrated Nursing Access Program: An Approach to Prepare Aboriginal Students for Nursing Careers," *International Journal of Nursing Education Scholarship:* Vol. 7, Iss. 1, Article 10.

Research Example Questions:

Question 1: What are the class intervals in the frequency table?

Question 2: What are the midpoints for each of the class intervals in the frequency table?

Question 3: What can be said of the ratio of participants scoring from 107 to 145 compared to all other participant scores?

Question 4: Suppose the class intervals and frequencies were as shown in the following table. Complete the frequency table and a histogram using these class intervals with frequency on the *y*-axis.

Frequency Table for Question 4

Score	*f*	*f*/*n*	%*f*	c.%*f*
34–68	1			
69–103	2			
104–138	25			
139–173	7			
Total	35			

Research Example Answers:

Question 1: What are the class intervals in the frequency table?
The class intervals are 34 to 106, 107 to 145, and 146 to 170.

Question 2: What are the midpoints for each of the class intervals in the frequency table?

The midpoint for the class interval 34–106 is 70 [(34 + 106) ÷ 2].
The midpoint for the class interval 107–145 is 126 [(107 + 145) ÷ 2].
The midpoint for the class interval 146–170 is 158 [(146 + 170) ÷ 2].

Question 3: What can be said of the ratio of participants scoring between 107–145 compared to all other participant scores?

$$Ratio = \frac{f_{v_1}}{f_{v_2}} = \frac{\#\text{of scores between } 107 - 145}{\#\text{of scores between } 34 - 106 \text{ and } 146 - 170} = \frac{29}{6} = 4.83 \quad \textbf{(2.6)}$$

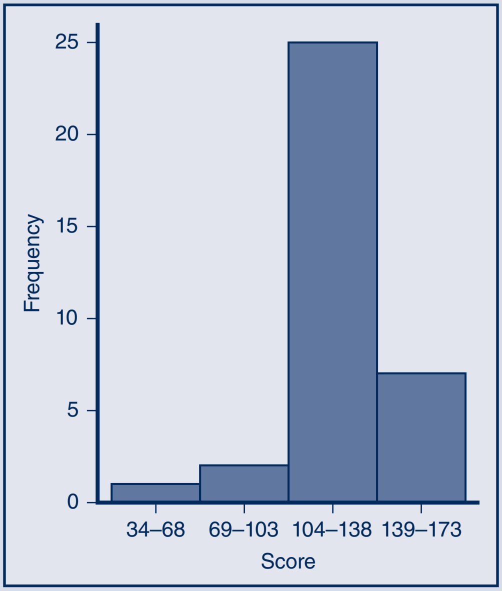
We can say that there were 4.83 scores within the 107 − 145 interval for every one score outside of that class interval (4.83:1).

Question 4: Suppose the class intervals and frequencies were as shown in the following table. Complete the frequency table and a histogram using these class intervals with frequency on the *y*-axis.

Frequency Table for Question 4

Score	*f*	*f*/*n*	%*f*	c.%*f*
34–68	1	0.029	2.9	2.9
69–103	2	0.057	5.7	8.6
104–138	25	0.714	71.4	80.0
139–173	7	0.200	20.0	100
Total	35	1.000	100.0	

Histogram for Question 4



Problems

Below is a complete list of the grades obtained in a statistics class last semester.

G	M	G	M	G	M	G	M	G	M
F	53	M	40	M	58	F	89	M	12
M	6	M	27	F	65	M	79	F	70
M	73	M	83	M	80	M	15	M	10
F	65	F	85	M	25	F	55	F	62
M	56	M	77	F	54	F	43	M	49

Column headings: G = gender; M = mark
Within rows: M = male; F = female

1. Prepare the grade frequency distribution using class intervals equal to 10.
2. Once completed, include the relative frequency and the cumulative frequency.
3. If 60 percent is considered a passing grade, what proportion of the students passed? Prepare a cross-tabulation with the gender variable in the columns and grade interval in the rows.
4. What percentage of the males passed?
5. What percentage of the females failed?
6. For each grade interval, indicate the ratio of males to females.
7. As a percentage, how many more males are there in the class than females?