

2

SUMMARIZING, ORGANIZING DATA, AND MEASURES OF CENTRAL TENDENCY

Learning Objectives

After reading and studying this chapter, you should be able to do the following:

- Construct a frequency distribution table with individual values to summarize data
- Create a frequency distribution table with equal intervals to summarize data
- Identify the differences between bar graphs and histograms
- Identify the common distribution shapes
- Define the positively skewed and negatively skewed distributions
- Differentiate among the three common central tendency measures: (1) mode, (2) median, and (3) mean
- Calculate mean from the formula
- Adjust the mean formula when using frequency tables

WHAT YOU KNOW AND WHAT IS NEW

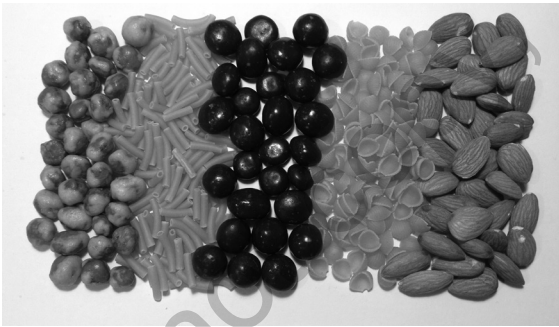
You learned most of the basic vocabulary for statistics in Chapter 1. You learned how to distinguish between samples and populations and how to identify the different scales

FIGURE 2.1 ● Unsorted Data

of measurement. You learned the basics of summation notation and refreshed your mathematical skills. You also learned that different scales of measurement come with distinct mathematical attributes. All of these will be extremely useful and will continue to be relevant throughout this book.

One of the main purposes in statistics is to make sense out of data. It can be very confusing to encounter data without any particular pattern. There are some simple and commonly used methods to make data manageable. In this chapter, we will explore different ways to make sense out of nominal, ordinal, interval, and ratio variables by summarizing and organizing them and finding typical values to represent a group of observations.

With the help of three images, we illustrate why summarizing and organizing help us understand and extract useful information out of data. Data are facts presented in numerical form. Unfiltered or unprocessed data are disorienting and confusing. Data can simply look like an unruly mess such as the big pile of miscellaneous pieces shown in Figure 2.1.

FIGURE 2.2 ● Organized or Sorted Data

When data are organized or sorted according to a particular pattern, such as by type, we can quickly get a rough idea about this pile of miscellaneous pieces. It seems that there are two kinds of pasta: (1) macaroni and (2) shells, and three kinds of snacks: (1) almonds, (2) Wasabi peas, and (3) chocolate-covered espresso beans, as shown in Figure 2.2. Sorting is very useful and commonly used in tracking inventories in retail business.

When data are summarized in a way that provides easy-to-process information about the numeric attributes, we can make sense out of them immediately. The following graph provides the number of pieces in each type. A summary graph might not include all the numerical characteristics of the data such as weight and size of each piece, but it provides meaningful information on the number of pieces in every type as shown in Figure 2.3. The height of the bar represents the quantity of each item. The transformation from

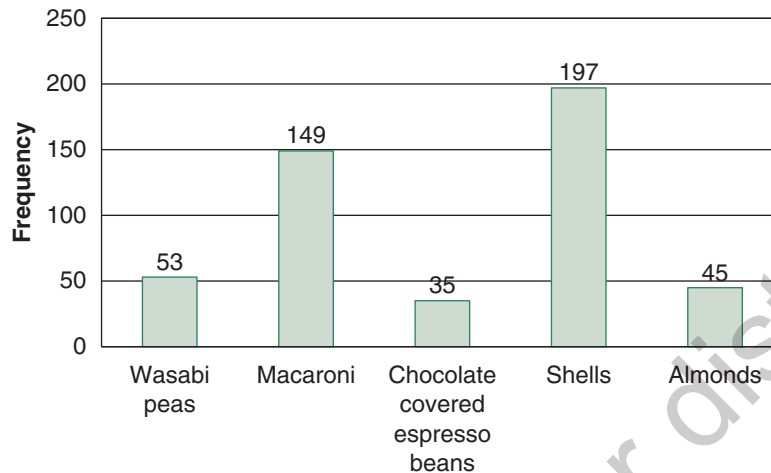
FIGURE 2.3 ■ Number of Pieces in Each Category

Figure 2.1 to Figures 2.2 and 2.3 illustrates the reason why it is helpful to organize and/or summarize data.

FREQUENCY DISTRIBUTION TABLE

Organizing and summarizing of data in a table form presents an easily accessible way to display the data at a glance. A **frequency distribution table** lists all values or categories along with a corresponding tally count for them. **Distribution** is defined as arrangement of values of a variable as they occur in a sample or a population. Both categorical variables (i.e., nominal or ordinal variables) and numerical variables (i.e., interval or ratio variables) can be organized in frequency distribution tables (often simply referred to as “frequency tables”).

Organizing and Summarizing Categorical Variables

When dealing with categorical data (i.e., nominal or ordinal scales of measurement), the process of organizing and summarizing is fairly simple. The observations are organized or sorted into categories, and then the tally count of each category is reported as **frequency**, which is simply referred to as f in this book. When you add up the f s of each category, the sum of all f s equals the sample size, n . Example 2.1 is used to demonstrate such a relationship between f and n .

EXAMPLE 2.1

Suppose that I asked 20 people “what kind of job do you do?” The answers turned out to be “truck driver,” “retail salesperson,” “secretary,” “personal assistant,” “day care worker,” “construction worker,” “car wash worker,” “server,” “bartender,” “librarian,” “customer service representative,” “call center worker,” and so on. To organize these answers, I decided to group similar jobs into occupational categories such as professional, sales, service, clerical, and laborer. Then, I counted the number of answers in each occupation category. The results are shown in Table 2.1.

TABLE 2.1 Frequency Distribution Table of Occupations

X (Occupation)	f (Frequency)	Relative Frequency
Professional	3	.15
Sales	6	.30
Service	5	.25
Clerical	4	.20
Laborer	2	.10
Total	20	1.00

Based on the data in Table 2.1, when I add up all the frequencies, $3 + 6 + 5 + 4 + 2$, the total frequency equals the sample size, $n = 20$. The last column of the frequency distribution table is called the relative frequency. **Relative frequency** is the same as proportion or percentage. It is defined and calculated as the frequency of a category (f) divided by the sample size (n). It provides the percentage of each category. Based on Table 2.1, the majority (i.e., 55%) of the 20 people worked in either sales or service jobs.

The sum of all relative frequencies should add to 1.00. It is a good way to check your math. If the sum does not add to 1.00, there are two possible explanations. First, the math is wrong. You need to double-check your calculations. Second, if the answer is .99, 1.01, or something very close to 1.00, it might be due to rounding.

$$\text{Relative frequency (proportion) for a category} = \frac{\text{Frequency of a category}}{\text{Total frequency}} = \frac{f}{n}$$

$$\text{Percentage for a category} = \frac{\text{Frequency of a category}}{\text{Total frequency}} \times 100\% = \frac{f}{n} \times 100\%$$

Organizing and Summarizing Numerical Variables

When dealing with numerical variables, frequency tables are constructed in a logical way to present the data in an orderly fashion with a reasonable number of categories. An example is used to explain the concepts of “orderly fashion” and “reasonable number of categories.”

EXAMPLE 2.2

Technology has become an integral part of our daily lives. This is especially true in higher education. Computer networks keep us connected electronically. Just like any technology, sometimes problems happen. Our university kept a record of the number of computer network interruptions per day on campus. The following numbers showed the number of computer network interruptions in the month of September.

1 3 1 1 0 1 0 1 1 0

2 2 0 0 0 1 2 1 2 0

0 1 6 4 3 3 1 2 4 0

Please organize the daily number of computer network interruptions into a frequency table and calculate the total number of interruptions for the month.

Let's rearrange the numbers and put them in a frequency distribution table. The table needs to have a reasonable number of categories, so the table is not too long or too short. When the table is too long, the information is too detailed to be absorbed quickly. When the table is too short, there is not enough distinction among the categories. What is a reasonable number of categories? The answer to that question is based on the capacity of human short-term memory: 7 ± 2 . A frequency table with categories between 5 and 9 is reasonable. The short-term memory capacity is defined as the capacity of holding a small amount of information in the brain in an active, immediately available fashion so that we are able to receive, process, and remember the information. According to Miller (1956), the capacity for short-term memory across many different types of stimuli such as words, notes, sounds, or pitches in various laboratory experiments all came to the same magical number 7, plus or minus 2.

Next, let's talk about the concept of "orderly fashion." There are many different kinds of "orders." One is ascending order, in which numbers are arranged from the lowest to the highest. Another is descending order in which numbers are arranged from the highest to the lowest. I prefer constructing a frequency table in an ascending order, from the lowest to the highest. To do this in the example of network interruptions, first, you need to identify the minimal value and maximal value. The minimal value is 0, and the maximal value is 6. Using the number of network interruptions as individual categories, thus, you have seven total categories. Tally the count for each category and report it as the frequency for that category. Therefore, the values of the network interruptions are arranged in an ascending order, lowest to highest, and the frequency of each value is counted, thus, you have constructed a frequency distribution table as shown in Table 2.2.

TABLE 2.2 ● Frequency Distribution Table for Daily Network Interruptions on a University Campus in September

X (Number of Interruptions)	f (Frequency)	Relative Frequency
0	9	.300
1	10	.333
2	5	.167

(Continued)

(Continued)

<i>X</i> (Number of Interruptions)	<i>f</i> (Frequency)	Relative Frequency
3	3	.100
4	2	.067
5	0	.000
6	1	.033
Total	30	1.000

When adding up all frequencies, the total is 30. There were originally 30 numbers reported in the month of September, one for each day of the month. Relative frequency (or proportion) is calculated and reported in the third column of Table 2.2. The sum of all relative frequencies is 1.000. From this table, we learn that about 63% of the time, the network is functioning fairly well with one or no interruptions in a day. Some of you might have noticed that the value 5 is not in the reported daily network interruption data. It is also acceptable to construct the frequency distribution table without the value 5 as shown in Table 2.2a. Because the frequency for the value 5 is 0, it does not affect the frequency count or the calculation of relative frequency. Either Table 2.2 or Table 2.2a is an acceptable way to construct a frequency table for these data.

TABLE 2.2a Alternative Frequency Distribution Table for Daily Network Interruptions on a University Campus in September

<i>X</i> (Number of Interruptions)	<i>f</i> (Frequency)	Relative Frequency	<i>fX</i>
0	9	.300	0
1	10	.333	10
2	5	.167	10
3	3	.100	9
4	2	.067	8
6	1	.033	6
Total	30	1.000	43

The frequency table summarizes the number of daily network interruptions in ascending order with corresponding frequency. Such an organization makes it easy to calculate the total number of network interruptions in September. Instead of adding 30 numbers from the first number to the last number, you can use the modified formula $\sum X = \sum fX$. The frequency table shows every distinct value with its corresponding frequency. In this

example, 0 occurs 9 times, 1 occurs 10 times, 2 occurs 5 times, 3 occurs 3 times, 4 occurs twice, and 6 occurs once. Therefore, you have to multiply the frequency by every distinct value and add them all up. $\sum X = \sum fX = 0 \times 9 + 1 \times 10 + 2 \times 5 + 3 \times 3 + 4 \times 2 + 6 \times 1 = 43$. There are multiplications and additions involved in $\sum fX$; the order of operations dictates that the multiplications have to be done before the additions. The total number of interruptions for the month was 43.

When dealing with data spanning over a large range, creating a frequency table with **equal intervals** to cover the entire range is a good solution. Equal intervals are created by including the same number of values in each interval in a frequency table. We will use a couple of examples to illustrate this process.

EXAMPLE 2.3

A first-grade teacher asked parents to estimate their children's average daily screen time outside of academic learning. The reported screen times for 25 first graders are shown below. The unit of measurement is a minute.

79, 45, 66, 89, 97, 55, 61, 86, 93, 81, 80, 73, 76,
84, 81, 67, 92, 75, 76, 69, 57, 88, 84, 59, 72

Please create a frequency table with a reasonable number of categories in an orderly fashion to organize the reported screen time.

It is difficult to get a sense of the distribution of the times by looking at 25 numbers in no particular order. The minimal screen time was 45 and the maximal value was 97. The range is maximal value minus the minimal value, $97 - 45 = 52$. It is not advisable to create a frequency table using individual values with such a large range. Equal intervals are usually created by the intervals of 5 or 10 for convenience, if feasible. To create equal intervals to cover this range of 52, an interval of 10 is appropriate. The lowest interval is between 40 and 49, so it covers the minimal value, and the highest interval is between 90 and 99, so it covers the maximal value. There should not be any overlap between two adjacent intervals. A frequency table with equal intervals is constructed in Table 2.3 for these 25 first graders.

TABLE 2.3 Frequency Distribution Table for 25 First Graders' Screen Times

Screen Time Interval	f (Frequency)	Relative Frequency
40–49	1	.04
50–59	3	.12
60–69	4	.16
70–79	6	.24

(Continued)

(Continued)

Screen Time Interval	f (Frequency)	Relative Frequency
80–89	8	.32
90–99	3	.12
Total	25	1.00

Table 2.3 summarizes and organizes the 25 first graders' screen times with useful insights. It showed that 16% of the first graders' screen times were below 60 minutes, 56% of the first graders' screen times were between 70 and 89 minutes, and 12% of the first graders' screen times were higher than 90 minutes.

The same principles that were used in constructing a frequency table for discrete variables can also be applied to continuous variables. Continuous variables mean that the observations may take any value between two integers in the forms of fraction or decimal point. It is difficult to obtain precise measures for continuous variables. Therefore, the concept of **real limits** is designed to cover a range of possible values that may be reflected by a continuous measure. The lower limit is the value minus $\frac{1}{2}$ of the unit and the upper limit is the value plus $\frac{1}{2}$ of the unit. Physical measurements, such as height and weight, are continuous variables. For example, when 1 pound is used as a unit to measure a person's weight and John's weight is 180 pounds, the lower limit of John's weight is 179.5 pounds and the upper limit is 180.5 pounds.

Let's apply real limits to online purchase measures in Example 2.4.

EXAMPLE 2.4

Online shopping has become a routine part of our daily activities. We can shop without leaving the comfort of our home, and the merchandise is delivered to our door. There were 59 students in one of my statistics classes. Their answers to the amount of money they spent on online shopping last week were organized in Table 2.4. What was the estimated total online spending for these students last week?

TABLE 2.4 Frequency Distribution Table for 59 Students' Online Purchases Last Week

X (Online Purchases)	f (Frequency)	Relative Frequency
\$0–\$19	5	.08
\$20–\$39	7	.12

X (Online Purchases)	f (Frequency)	Relative Frequency
\$40–\$59	25	.42
\$60–\$79	12	.20
\$80–\$99	7	.12
\$100–\$119	2	.03
\$120–\$139	1	.02
Total	59	.99

When we made online purchases, the transactions were calculated to two places after the decimal point, such as \$34.67 or \$117.92. The dollar amount spent on the online purchase is a continuous variable so real limits apply here. For example, the real limits for the highest interval \$120–\$139 are \$119.5 as the lower limit and \$139.5 as the upper limit. Just to make it clear that each interval includes the lower limit but not the upper limit (i.e., lower limit $\leq X <$ upper limit), there is no overlap between two adjacent intervals. You might have already noticed that the total relative frequency adds to .99. This is simply due to rounding.

To obtain an estimated total dollar amount spent by 59 students' online purchases last week, the midpoint for each interval needs to be calculated. The formula for calculation of the midpoint for an interval is $X_{\text{midpoint}} = \frac{\text{lowend} + \text{highend}}{2}$.

The frequency table with midpoints is shown in Table 2.4a. In a frequency table consisting of equal intervals, the formula for $\sum X$ needs to be modified to $\sum fX_{\text{midpoint}}$. The estimated total amount is the sum of the midpoint of each interval times its corresponding frequency. Again, multiplications and additions are involved in $\sum fX_{\text{midpoint}}$. Multiplications have to be done before additions. The estimated total dollar amount for 59 students' online purchases was \$3,300.50.

TABLE 2.4a Frequency Distribution for 59 Students' Online Purchases Last Week With Midpoint of Each Interval

X (Online Purchases)	f (Frequency)	X_{midpoint}	fX_{midpoint}
\$0–\$19	5	\$9.5	\$47.50
\$20–\$39	7	\$29.5	\$206.50
\$40–\$59	25	\$49.5	\$1,237.50
\$60–\$79	12	\$69.5	\$834.00
\$80–\$99	7	\$89.5	\$626.50
\$100–\$119	2	\$109.5	\$219.00
\$120–\$139	1	\$129.5	\$129.50
Total	59		\$3300.50

Let's review the process of constructing a frequency table with equal intervals. The key points are summarized below:

1. Identify the minimal and maximal values of the variable. Calculate the range.
2. Choose an appropriate and convenient interval, usually in 5s or 10s if feasible, to construct a frequency table with 7 ± 2 equal intervals.
3. Create the equal intervals in an ascending order.
4. Make sure that there is no overlap between two adjacent intervals. For continuous variables, each interval includes the lower limit but not the upper limit.
5. The midpoint for each interval is calculated as

$$X_{\text{midpoint}} = \frac{\text{low end} + \text{high end}}{2}.$$



POP QUIZ

1. In a statistics class, students' quiz scores on a pop quiz with only four questions were reported in the frequency table below. Use the frequency table to calculate the total quiz score for the class.

X , Quiz	f (Frequency)
0	2
1	1
2	5
3	8
4	9

What is the total quiz score for the class?

- a. 10
 - b. 25
 - c. 71
 - d. 100
2. When constructing a frequency table with equal intervals of a continuous variable X , each interval consists of same width. To make sure that there is no overlap between two adjacent intervals, each interval actually includes
 - a. lower limit $< X <$ upper limit.
 - b. lower limit $\leq X <$ upper limit.
 - c. lower limit $< X \leq$ upper limit.
 - d. lower limit $\leq X \leq$ upper limit.

GRAPHS

Creating a graph is another common way to organize and summarize data. There is a specialized field in statistics called data visualization. **Data visualization** is the graphic representation of data. This is especially important in the era of big data because graphs can better communicate information than millions of unsorted data points. A well-organized graph can deliver useful numerical information in a quick and easy-to-understand fashion. Data visualization tools are available in many statistics software packages. In this book, we cover basic chart functions in Excel. Two commonly used graphs are (1) bar graphs and (2) pie charts.

Bar Graphs and Histograms

When **bar graphs** are used to organize categorical data, the horizontal scale (x -axis) represents the values (or categories) of the variable and the vertical scale (y -axis) represents the frequency or relative frequency of each value. When using the data from Table 2.1 to construct a bar graph, the result is shown in Figure 2.4a with the x -axis representing occupations and the y -axis representing the corresponding frequency of each occupation.

It is also possible to create a bar graph with multiple sets of bars to represent multiple categorical variables. For example, male and female students' jobs can be reported separately as shown in Figure 2.4b where males' occupations are represented in lighter blue bars and females' occupations are represented in darker blue bars. There is no darker blue bar in the last occupational category, laborer, because there are no females in this category.

FIGURE 2.4a Bar Graph of Occupational Categories

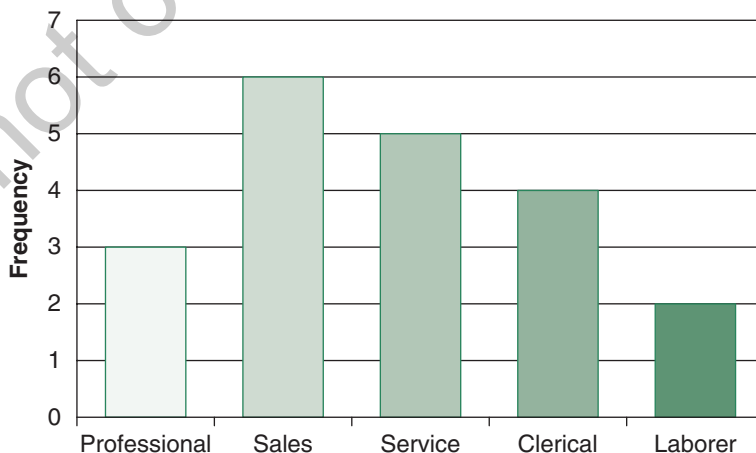
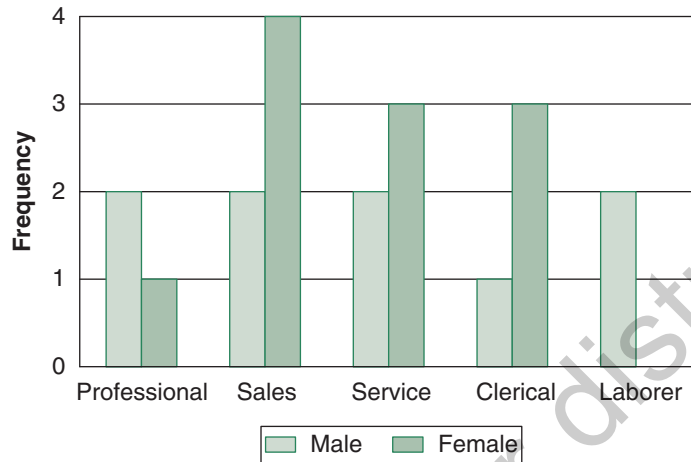
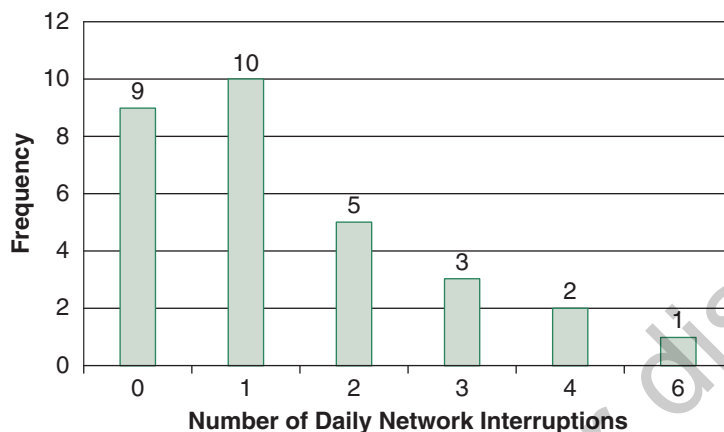


FIGURE 2.4b Bar Graph of Males' and Females' Occupations



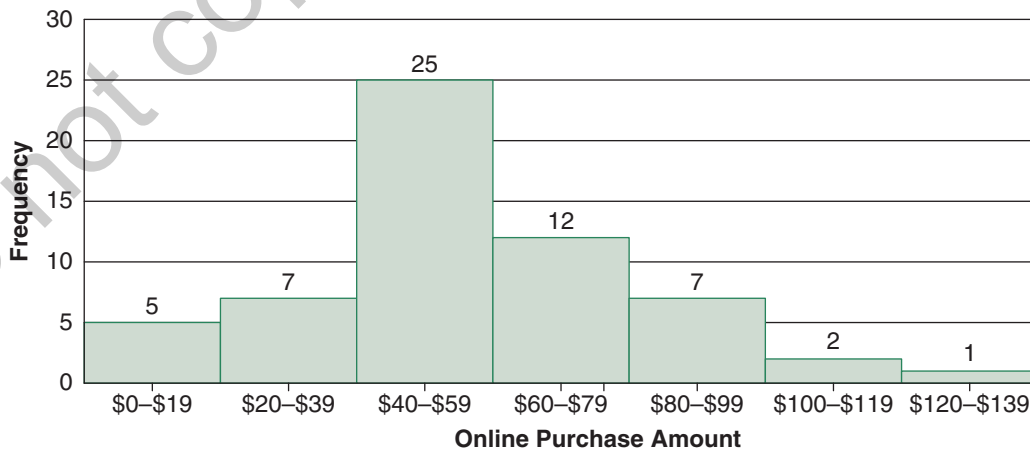
Bar charts can be applied to discrete data. When using data from Table 2.2 to construct a bar graph, the x -axis represents the number of reported daily network interruptions, and they are lined up from the smallest to the largest value from left to right. The y -axis represents the frequency, and the bar height represents frequency of each value. Frequency can only be integers. There are no valid answers between two integers: (1) k and (2) $k + 1$, because the number of daily network interruptions has to be a whole number. It is not possible to have 2.15 network interruptions in any particular day. Therefore, the bars are separate from one another, showing a gap in between two adjacent values. In bar graphs, bars do not touch each other as shown in Figure 2.5. Looking at Figure 2.5, you got an idea that most days (i.e., 19 out of 30 days), the university computer networks operated smoothly with no interruptions or only one interruption. According to Table 2.2 data, none of the days had five interruptions so there was not a bar that represented 5 interruptions. The last bar represented six daily interruptions happened once in September.

When using a graph to summarize a continuous variable, we need to consider the real limits of the values. Due to the fact that there are numerous possible answers between two adjacent values, there are no gaps between two values. We need to create a new type of graph called the **histogram**. A histogram is defined as a graphical presentation of a continuous variable. The bars of a histogram are of equal width to represent equal intervals of values, and they are touching each other to illustrate that the values are continuous. The difference between a bar graph and a histogram is that the two adjacent bars are touching each other in a histogram, but there are gaps between the bars in a bar

FIGURE 2.5 Bar Graph for the Number of Daily Network Interruptions in September

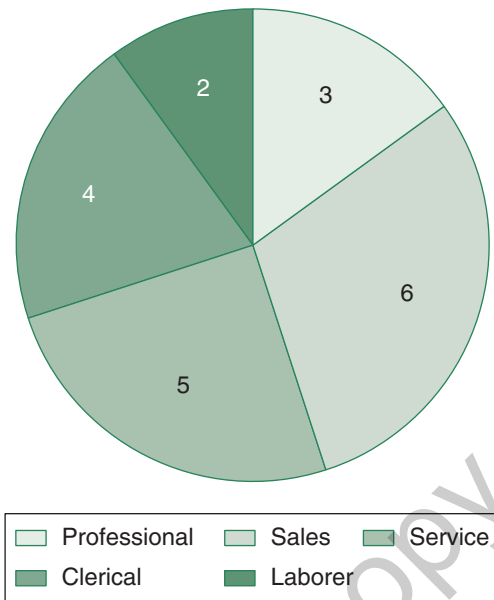
graph. Bar graphs are used for discrete variables, and histograms are used for continuous variables. The height of the bar represents the corresponding frequency for each interval. Using online purchase data in Table 2.4, we can construct a histogram with \$20 intervals as shown in Figure 2.6.

The highest amount interval \$120–\$139 is actually bounded by a lower limit of \$119.5 and an upper limit of \$139.5. The lower limit is included but the upper limit is not

FIGURE 2.6 Histogram of 59 Students' Online Purchase Amount Last Week

included in each interval; therefore, the real limits of the interval \$120–\$139 refer to $\$119.5 \leq X < 139.5$, and the real limits of the second highest interval \$100–\$119 refer to $\$99.5 \leq X < \119.5 . Such a rule is to make sure that there are no overlaps between intervals for continuous variables. The tallest bar represents \$40–\$59, which means 25 students spent such an amount for online purchases last week. More students belong in this category than any other category. A step-by-step instruction on how to create this histogram in Excel is provided at the end of this chapter.

FIGURE 2.7 Pie Chart of the Occupational Categories



Pie Charts

A **pie chart** usually depicts categorical data in a circle where the size of each slice of the pie corresponds to the frequency or relative frequency of each category. Pie charts are visually intuitive because the entire circle is 360 degrees. All categories add up to 100%, which constitute the entire 360 degrees. Using the occupational data in Table 2.1, a pie chart is constructed in Figure 2.7. It presents a clear image that sales and service job categories cover more than half of the pie, which means more than 50% of respondents hold jobs in these two categories.

Pie charts are extremely popular in presenting reports on budgets, government spending, or TV polling. When used appropriately, pie charts can be a powerful visual aid to present information. The right conditions to use pie charts include (a) the total adds to a meaningful sum, (b) there are no overlaps between categories, and (c) there is a reasonable number of categories. As you can imagine, if the number of categories is more

than 10, each slice becomes very small. There will be too much information crowded into the pie chart. The disadvantage of a pie chart is that it doesn't present any statistical information other than frequency or proportion.



POP QUIZ

3. Which one of the following variables is most likely to be used to create histograms?
 - a. Nominal variables
 - b. Ordinal variables
 - c. Discrete variables
 - d. Continuous variables

COMMON DISTRIBUTION SHAPES

A distribution refers to all possible values and the number of times every value occurs in a sample or a population. In either a bar graph or a histogram where values are lined up from the smallest (the left side of the x -axis) to the largest (the right side of the x -axis), the height of the bar represents frequency or relative frequency (i.e., proportion) of each value; therefore, the shape of the distribution is in full display. Common shapes of data distributions include uniform distribution, normal distribution, and skewed distribution.

Uniform Distribution

In a **uniform distribution**, every value appears with the same frequency, proportion, or probability. Many things have uniform distribution such as dice, coins, and cards. For example, a die has six sides with 1, 2, 3, 4, 5, or 6 dots on each side. The relative frequency or probability of throwing a die and obtaining any of 1, 2, 3, 4, 5, or 6 dots is evenly distributed as $1/6$, which is shown in Table 2.5 and Figure 2.8. The probability distribution of the number of dots on a die is called a uniform distribution. A fair die produces the same probability of landing on every side.

Normal Distribution

When you throw two dice at the same time and add the number of dots on the dice, the answers are 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, or 12 dots. The probabilities of obtaining those dots and various combinations for creating the number of dots are listed in Table 2.6. The total probability adds to 1.001 due to rounding. The bar graph of the sum of dots from two dice is shown in Figure 2.9.

TABLE 2.5 ■ Probability of Throwing a Die and Obtaining a Specific Number of Dots







X (Number of Dots)	f (Frequency)	Relative Frequency (Probability)
1 	1	$1/6 = .167$
2 	1	$1/6 = .167$
3 	1	$1/6 = .167$
4 	1	$1/6 = .167$
5 	1	$1/6 = .167$
6 	1	$1/6 = .167$
Total	6	1.002

FIGURE 2.8 ◆ Uniform Distribution of Number of Dots on a Die

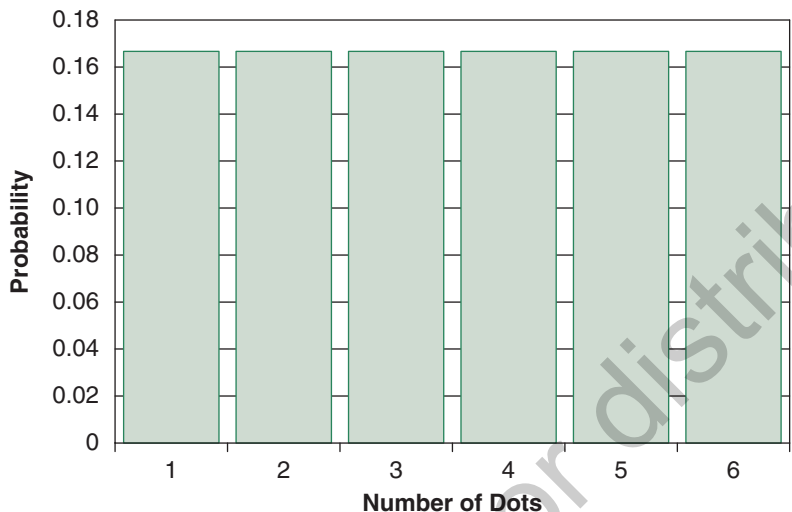
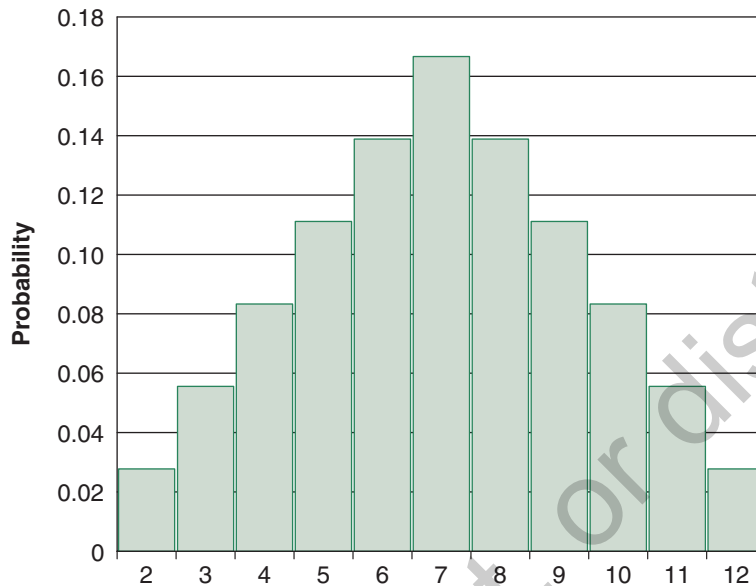


TABLE 2.6 ◆ Total Number of Dots From Two Dice and the Probabilities

X (Number of Dots)	f (Frequency)	Probability
2 (1, 1)	1	$1/36 = .028$
3 (1, 2); (2, 1)	2	$2/36 = .056$
4 (1, 3); (2, 2); (3, 1)	3	$3/36 = .083$
5 (1, 4); (2, 3); (3, 2); (4, 1)	4	$4/36 = .111$
6 (1, 5); (2, 4); (3, 3); (4, 2); (5, 1)	5	$5/36 = .139$
7 (1, 6); (2, 5); (3, 4); (4, 3); (5, 2); (6, 1)	6	$6/36 = .167$
8 (2, 6); (3, 5); (4, 4); (5, 3); (6, 2)	5	$5/36 = .139$
9 (3, 6); (4, 5); (5, 4); (6, 3)	4	$4/36 = .111$
10 (4, 6); (5, 5); (6, 4)	3	$3/36 = .083$
11 (5, 6); (6, 5)	2	$2/36 = .056$
12 (6, 6)	1	$1/36 = .028$
Total	36	1.001

FIGURE 2.9 ■ Probability of Number of Dots From Two Dice

The bar graph shown in Figure 2.9 has the typical characteristics of a **normal distribution**: (a) the distribution peaks at the center, (b) the distribution symmetrically tapers off on both sides and the left side is a mirror image of the right side, and (c) 50% of the distribution is above the mean and the other 50% is below the mean.

Many statistical procedures require sample data with a distribution that is approximately normally distributed. A normal distribution is a very important data distribution pattern in statistics. When one examines a distribution from a large sample size, the edges of the histograms are likely to be smoothed out and almost become a curve as seen in Figure 2.10, showing the bar graph of the probability of getting a certain number of heads from 100 coin tosses.

When you trace the outline of the bar graph in Figure 2.10, you create a **line graph**. A line graph is defined as a graph that displays quantitative information with a line or curve that connects a series of adjacent data points. This particular line graph shows a normal distribution curve. A normal distribution curve peaks at the center of the data (i.e., the mean) and then symmetrically tapers off on both sides with 50% of the distribution above the mean and 50% of the distribution below the mean. Draw a line straight through the mean, and the left side and the right side are mirror images of each other as shown in the line graph in Figure 2.11.

FIGURE 2.10 The Probability of Getting a Certain Number of Heads From 100 Coin Tosses

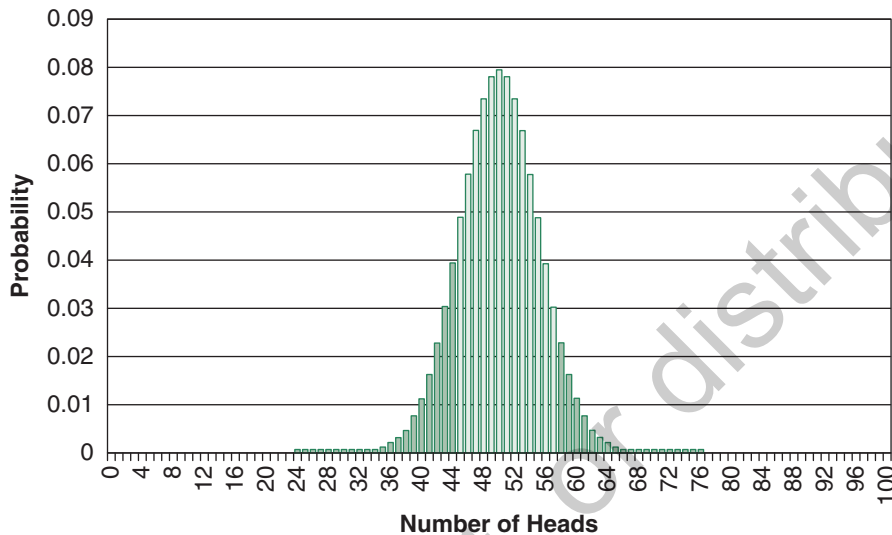
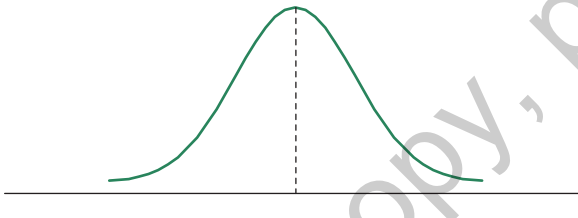


FIGURE 2.11 Normal Distribution Curve

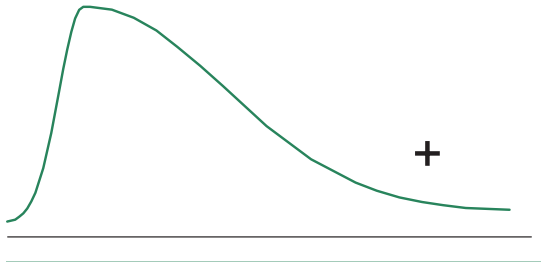
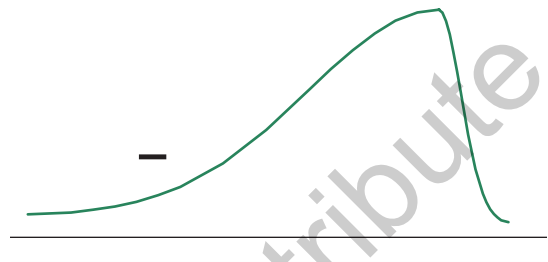


Skewed Distribution

Not all data are normally distributed. A **skewed distribution** happens when values are not symmetrical, and they concentrate more on one side than on the other. When the distribution mostly concentrates on the left side with a long tail to the right side, it is labeled as **positively skewed distribution** or right

skewed as shown in Figure 2.12. For example, income distribution in the United States is positively skewed with a very long tail to the right side. It means that most people earn a modest income while the top 1% is making hundreds times more than regular working-class salary.

When the distribution mostly concentrates on the right side with a long tail to the left side, it is labeled as **negatively skewed distribution** or left skewed as shown in Figure 2.13. For example, when students are asked what grades they expect to get on a statistics exam, the majority of the students answer 80s or 90s with very few low scores. The labeling of the skewness depends on the direction of the tail. When the tail points to the high end of the distribution, the data distribution is positively skewed (or right skewed) and when the tail points to the low end of the distribution, the data distribution is negatively skewed (or left skewed).

FIGURE 2.12 ■ Positively Skewed or Right-Skewed Distribution**FIGURE 2.13** ■ Negatively Skewed or Left-Skewed Distribution**POP QUIZ**

4. A _____ skewed distribution is a distribution with a high concentration of data points at the low end of the values and a long tail toward the high end of the values.
- positively
 - negatively
 - doubly
 - exponentially

MEASURES OF CENTRAL TENDENCY

Besides tables and charts, we also use numeric numbers to describe attributes of a sample. **Central tendency** is defined as a single number to describe the center of a distribution. The center of a distribution is usually the most typical or representative value of all the values. There are three measures that are commonly used to represent central tendency: *Mean*, *median*, and *mode*. These three common measures of central tendency will be presented in order of mathematical complexity from the simplest to the most sophisticated. Therefore, we will discuss mode first, median second, and mean last.

Mode

The **mode** is defined as the value or category with the highest frequency in a distribution. It means the value or category that occurs most often. To identify the mode in a distribution, you may do a simple tally count of individual values or look into a frequency table and identify the *value* with the highest frequency without performing any other mathematical operation. Frequency counts can be applied to all four scales of measurement: (1) nominal, (2) ordinal, (3) interval, and (4) ratio. Let's look at an example.

EXAMPLE 2.5

According to T. Rowe Price's most recent annual Parents, Kids, Money Survey, one of the questions for the parents was "Which of the following best describe how you feel about savings for your kids' college education?" The frequency and percentage of the answer to this question is shown in Table 2.7. Based on Table 2.7, what is the mode of how parents feel about saving for their kids' college education?

TABLE 2.7 Frequency and Percentage of How Parents Feel About Saving for Kids' College Education

Which of the Following Best Describe How You Feel About Saving for Your Kids' College Education?	Frequency	Percentage (Relative Frequency)
Was not able to pay ANY of the cost	164	18.5%
Was able to pay SOME of the cost	397	44.8%
Was able to pay MOST of the cost	219	24.7%
Was able to pay ALL of the cost	107	12.1%
Total	887	100.1%

Note: Data from T. Rowe Price, "11th Annual Parents, Kids & Money Survey—College Savings Results" (2019) via Slideshare.com.

The savings for college education question was answered on a nominal scale of measurement. Mode can be applied to a nominal variable. The category with the highest frequency was "able to pay SOME of the cost." Notice that 44.8% of parents belonged in this category. Only 12.1% of the parents felt that they were able to pay all of the cost for their kids' college education, 24.7% of the parents were able to pay most of the cost, and still 18.5% of the parents were not able to pay any of the cost. Therefore, the vast majority of the students would still have to be responsible for at least part of the cost for their college education. The total percentage added up to 100.1% due to rounding.

Mode can be applied to all four different scales of measurement. In the next example, we will demonstrate how mode can be applied to an ordinal scale of measurement.

EXAMPLE 2.6

A middle school conducted an anonymous survey on the occurrences, severity, and types of bullying incidents at school. The answers to the question "How often have you been bullied at school in the past couple of months?" were tallied and reported in Table 2.8. What is the mode of this distribution?

TABLE 2.8 Occurrences of Bullying Incidents at a Middle School

Occurrences of Bullying Incidents	Frequency	Relative Frequency
I have not been bullied at school	320	65%
It has only happened once or twice	121	24%
2 or 3 times a month	25	5%
About once a week	17	3%
Several times a week	13	3%
Total	496	100%

The frequency table reported the occurrences of bullying incidents in five categories. The answers to “How often have you been bullied at school in the past couple of months?” are classified as an ordinal scale of measurement. The answers indicate an increasing level of occurrences of bullying incidents from “not being bullied” to “several times a week.” The mode is the category with the highest frequency. The highest frequency is 320. The mode is the value or category with the highest frequency, so the answer is “I have not been bullied at school in the past couple of months.” As is indicated by the relative frequency column, 65% of children reported that they were not bullied at school. It was a good news that a majority of the students did not have the personal experience of being bullied. The results also indicated that there were 30 students who had encountered regular weekly bullying. The school needs to take action to protect these students so they don’t continue to suffer such mistreatment or abuse at school.

What you need to know about mode is that

1. it can be applied to nominal, ordinal, interval, or ratio data, and
2. a data set may have one mode, multiple modes, or no mode. For example, if two values or categories both have the same highest frequency, then the distribution has two modes, which is called a bimodal distribution. When none of the values is repeated in a distribution, or when each value is repeated the same number of times, there is no mode. Usually, it is common to have one mode in a distribution.

Median

The **median** is defined as the value right in the middle of the distribution when all values are sorted from the lowest to the highest. Based on the definition, the data need to be sorted. Therefore, they need to have at least ordinal attributes. It is important to note that median can be meaningfully applied to only ordinal, interval, or ratio variables.

The median is the middle position of sorted observed values. Let's clarify the term *middle position*. When the sample size n is an odd number, there is only one middle position in the distribution. When the sample size n is an even number, there are two middle positions in the distribution. Finding the median involves finding the middle location(s) of the distribution. This process is illustrated by Examples 2.7 and 2.8.

EXAMPLE 2.7

The following numbers are the fuel efficiency, in miles per gallon, reported on five randomly selected cars of the same model by the car manufacturer: 31, 28, 46, 39, and 41. What is the mode and median fuel efficiency of these five cars?

These five values all occur once in the sample. There is no mode in this case.

The sample size, $n = 5$, is an odd number. There is one middle position when sample size is an odd number. However, the values were reported in an arbitrary order so they need to be sorted first. The sorted values from the lowest to the highest are 28, 31, 39, 41, and 46. When n is an odd number, the middle location is determined by $L = \frac{n}{2}$ then round it up to the next integer. In this example, $L = \frac{5}{2} = 2.5$ then round up to 3. The third position of the sorted data is the median. The value of the third position is 39. Therefore, the median fuel efficiency for these five cars is 39 miles per gallon.

EXAMPLE 2.8

Let's use the frequency table of internet network interruption example in Table 2.2 to illustrate how to figure out the mode and median by introducing a new concept, cumulative relative frequency. **Cumulative relative frequency** is defined as the accumulation of the relative frequency for a particular value together with all the relative frequencies for the values lower than it. The cumulative relative frequency only applies to variables with orders (i.e., ordinal, interval, or ratio variables). Median can be identified in a frequency table as the first value with the cumulative relative frequency over .50. Let's use the network interruptions example with cumulative relative frequency calculated in Table 2.9.

TABLE 2.9 Frequency Table for Daily Network Interruptions With Cumulative Relative Frequency

X (Number of Interruptions)	f (Frequency)	Relative Frequency	Cumulative Relative Frequency
0	9	.300	.300
1	10	.333	.633
2	5	.167	.800

<i>X</i> (Number of Interruptions)	<i>f</i> (Frequency)	Relative Frequency	Cumulative Relative Frequency
3	3	.100	.900
4	2	.067	.967
5	0	.000	.967
6	1	.033	1.000
Total	30	1.000	

Median can be identified as the first value with cumulative relative frequency over .50. In this case, the first cumulative relative frequency over .50 is .633, and its corresponding value is 1. If you use the same method as Example 2.7 by identifying two middle locations in a sample size $n = 30$, the values of 15th and 16th positions are both 1; therefore, they are averaged to obtain the median = 1. You get the same results using different methods to obtain the median.

Here is a summary of the steps to identify the median in a distribution. Median can only be obtained in ordinal, interval, or ratio variables.

1. SORT the data from the smallest value to the largest value.
2. If n is an odd number, the location for the middle position is $L = n/2$, then round up to the next integer. Identify median as the value of the L th location in the distribution.
3. If n is an even number, there are two middle positions in the data. They are located in L th and $(L + 1)$ st positions. Median is the average of the values at those two middle locations.
4. In a frequency table, the median is the first value with cumulative relative frequency over .50.

What you need to know about the median is that

1. its calculation does not involve every value in the variable, and
2. it is not affected much by extreme values of the variable. It remains relatively stable even with extreme values in the variable.

Mean

The **mean** (also referred to as the average, or arithmetic mean) is a measure of central tendency that is determined by adding all the values and dividing by the number of

observations. The formula for the sample mean is $\bar{X} = \frac{\Sigma X}{n}$ and the formula for the population mean is $\mu = \frac{\Sigma X}{N}$.

The common element of the formulas for \bar{X} and μ is ΣX . As mentioned in Chapter 1 in the section on Statistical Notations,

$$\Sigma X = X_1 + X_2 + X_3 + \cdots + X_n$$

The summation of X , ΣX , denotes adding all values of X , from the first value to the last value. The mean is calculated by the sum of all values divided by the number of observations, $\bar{X} = \frac{\Sigma X}{n}$ or $\mu = \frac{\Sigma X}{N}$.

When applying the mean formula using frequency tables where individual values occur different numbers of times, the formula needs to be modified to $\bar{X} = \frac{\Sigma X}{n} = \frac{\Sigma fX}{n}$. Such a modification allows the impact of frequencies to be fully incorporated. The same modification also applies to calculation of the population mean. We will use an example to illustrate this process.

EXAMPLE 2.9

The numbers of times a group of college students used Grubhub to have food delivered to them during last week are reported in Table 2.10. What are the mode, median, and mean of these students' number of Grubhub usage last week?

TABLE 2.10 Students' Grubhub Usage

X (Grubhub Usage)	f (Frequency)	Relative Frequency	Cumulative Relative Frequency
0	1	.02	.02
1	2	.04	.06
2	2	.04	.10
3	14	.28	.38
4	15	.30	.68
5	16	.32	1.00
Total	50		

The highest frequency is 16 in Table 2.10 and its corresponding value is 5. The mode is 5.

The median is easy to find with the cumulative relative frequency column. It can be readily seen that 4 is the first value with cumulative relative frequency over .50. The median is 4.

The formula for mean is adjusted to $\bar{X} = \frac{\sum fX}{n}$ because data are presented in a frequency table. According to the formula, the numerator and the denominator need to be figured out before the division. For the numerator, the order of operations dictates the multiplication of fX be conducted first. Therefore, the column of fX is added to Table 2.10a. Then at the end, the summation $\sum fX$ is calculated and the answer

is 188. For the denominator, the sample size is 50. The mean is $\bar{X} = \frac{\sum X}{n} = \frac{\sum fX}{n} = \frac{188}{50} = 3.76$. The rule

of rounding in reporting median or mean is to report the answer with one more decimal place than the original data. Since the original quiz scores are reported as integers, the answer for the mean should be rounded one place after the decimal point. Mean is 3.8. That is, the students' mean number of usage of Grubhub is 3.8.

TABLE 2.10a Students' Grubhub Usage With fX Column

X (Grubhub Usage)	f (Frequency)	fX
0	1	0
1	2	2
2	2	4
3	14	42
4	15	60
5	16	80
Total	50	188

When dealing with data spanning over a large range, creating a frequency table with equal intervals is a good solution. Identifying measures of central tendency involved in frequency tables with equal intervals needs to be explicitly stated here. Let's use Example 2.10 to illustrate this process.

EXAMPLE 2.10

Depression can affect anyone. Table 2.11 shows the ages for a random sample of $n = 35$ adult patients who are diagnosed with depression. What are the measures of central tendency for the patients' ages?

(Continued)

(Continued)

TABLE 2.11 Frequency Table of 35 Adult Depression Patients' Ages

Patient's Age	<i>f</i> (Frequency)
20–29	7
30–39	9
40–49	8
50–59	7
60–69	4
Total	35

The mode of patients' ages is 30–39 and the median 40–49. When you are asked to estimate the mean of the patients' ages from the frequency table, there is no way to find where the ages are located within the interval. The best single value to estimate each interval is the midpoint of the interval, which is defined by the average of the endpoints in the interval, $X_{\text{midpoint}} = \frac{\text{low end} + \text{high end}}{2}$. When

the frequency table consists of equal intervals, the formula for the mean $\bar{X} = \frac{\sum X}{n} = \frac{\sum fX}{n}$ needs to be modified to $\bar{X} = \frac{\sum fX_{\text{midpoint}}}{n}$. The midpoint for the first interval 20–29 is $(20 + 29)/2 = 24.5$. The midpoints are shown in Table 2.11a. The estimated total score is the sum of the midpoint of each interval times its corresponding frequency. Again, multiplication and addition are involved in $\sum fX_{\text{midpoint}}$. Multiplication operations have to be done before addition operations. The estimated total age for 35 patients is 1477.5. The mean is $1477.5/35 = 41.2$. Clearly, this estimated mean is slightly different from the actual mean if you have all the individual patient's ages and can add them all up. However, when you are presented with a frequency table without all individual values, this estimated mean provides a quick answer that is very close to the actual mean due to the fact that some numbers in the interval are higher than the midpoint and others are lower than the midpoint. The positive differences and the negative differences might cancel each other out to some extent.

TABLE 2.11a Frequency Table of 35 Adult Depression Patients' Ages With $\sum fX_{\text{midpoint}}$

Patient's Age	<i>f</i> (Frequency)	X_{midpoint}	fX_{midpoint}
20–29	7	24.5	171.5
30–39	9	34.5	310.5
40–49	8	44.5	356
50–59	7	54.5	381.5
60–69	4	64.5	258
Total	35		1477.5

TABLE 2.12 Summary Table on What Central Tendency Measures Apply to Which Scales of Measurement

Central Tendency	Scales of Measurement			
Mode	Nominal	Ordinal	Interval	Ratio
Median		Ordinal	Interval	Ratio
Mean			Interval	Ratio

It is important to know which statistical procedures can be appropriately applied to what scales of measurement. In Table 2.12, I summarize what measures of central tendency can be appropriately applied to which scales of measurement.

In the next sections, we will discuss calculating mean in different situations that involve slight modification of the mean formula.

Weighted Mean

The **weighted mean** is defined as calculating the mean when data values are assigned different weights, w . The formula for weighted mean is $\bar{X} = \frac{\sum wX}{\sum w}$. Weighted mean is appropriate when every value in the data is not treated equally. One of the most relevant examples for calculating a weighted mean for students is to show how GPA (grade point average) is calculated. Here is a statement I've often heard from students, "I got an A in one course and a C from another course, my average for this semester was a B." Was this statement correct? Not necessarily, as it depends on whether these two courses had the same number of credit hours. If John got an A in a 1-credit-hour American Sign Language course and a C in a 4-credit-hour Behavioral Science Statistics course, then his semester average would be closer to C than to B. When calculating GPA, courses with a higher number of credit hours are considered more heavily than courses with a lower number of credit hours. The credit hours of the courses are the weights that need to be considered. In a four-point system, the grades are worth A = 4, B = 3, C = 2, D = 1, and F = 0 points. Full-time undergraduate students usually take four to six different courses in a semester. Let's demonstrate how to calculate a student's GPA by using an example.

EXAMPLE 2.11

Chris received a B in Abnormal Psychology (3 credit hours), an A in Behavioral Science Statistics (4 credit hours), an A in Rock and Blues (1 credit hour), a C in Psychology of Women (3 credit hours), and a B in Biology (5 credit hours) last semester. What is Chris's GPA for last semester?

(Continued)

(Continued)

Let's start by organizing all the courses in a table including the courses, grades, and credit hours as shown in Table 2.13.

TABLE 2.13 Chris's Courses, Grades, and Credit Hours

Course	X (Grade)	w (Credit Hours)
Abnormal Psychology	B = 3	3
Behavioral Science Statistics	A = 4	4
Rock and Blues	A = 4	1
Psychology of Women	C = 2	3
Biology	B = 3	5

According to the formula for weighted mean, $\bar{X} = \frac{\sum wX}{\sum w}$, the multiplication wX needs to be conducted first. Therefore, the column of wX is created in Table 2.13a.

TABLE 2.13a Chris's Courses, Grades, and Credit Hours With wX Column

Course	X (Grade)	w (Credit Hours)	wX
Abnormal Psychology	B = 3	3	9
Behavioral Science Statistics	A = 4	4	16
Rock and Blues	A = 4	1	4
Psychology of Women	C = 2	3	6
Biology	B = 3	5	15
Total		16	50

Once the column of wX is created, the $\sum wX = 50$ is calculated at the end of the column, which provides the total weighted points earned. Then the value of the total weighted points is divided by the total number of credit hours (the total weight). This division gives the weighted mean = $50/16 = 3.125$. Again, the rounding off rule states the answer of a calculation is one more decimal place than the original data. The original grades are reported as integers. The weighted mean should be reported at one place after the decimal point, Chris's GPA for last semester is 3.1.

What you need to know about mean is that

1. it applies to interval or ratio measures which come with equal units,
2. its calculation involves every value in the variable, therefore,
3. it is sensitive to the impact of *outliers*.

Outliers refer to extreme values in a distribution. Outliers usually stand far away from the rest of the data points. They are not representative of the sample or population. In some statistical procedures, outliers can have heavy influences on the results. In central tendency measures, mean is the only one that is heavily influenced by outliers. Mode and median are not likely to be influenced by outliers.

Locations of Mean, Median, and Mode in Different Shapes of Distribution

In a normal distribution, the highest frequency is in the middle, then it symmetrically tapers off on both sides. The right side is a mirror image of the left side; thus, mean, median, and mode are located in the same spot, as shown in Figure 2.14.

The mode is the value with the highest frequency, which is at the peak exactly in the middle. The median is the mid-point of the distribution, so it is also precisely in the middle. Because of the symmetrical tapering off on both sides, the values above the mean on the right side compensate for the values below the mean on the left side perfectly. Therefore, the mean stays exactly in the middle as well. In a normal distribution, the mean, median, and mode are located at the same spot, exactly in the middle.

The locations of the mean, median, and mode in skewed distributions are quite different. In a positively skewed (right skewed) distribution, the mode is the value with the highest frequency. This occurs at the peak, which is located to the left side as marked in Figure 2.15. The mean is the arithmetic average of all values and is sensitive to the impacts of outliers. Therefore, the long tail to the right side is pulling the mean toward the tail. The median is in between the mode and the mean.

In a negatively skewed distribution, the peak is located to the right side as marked in Figure 2.16. The peak is the mode which is the value with the highest frequency. The mean is the arithmetic average of all values and is sensitive to the impacts of outliers.

FIGURE 2.14 Normal Distribution Curve With Mean, Median, and Mode

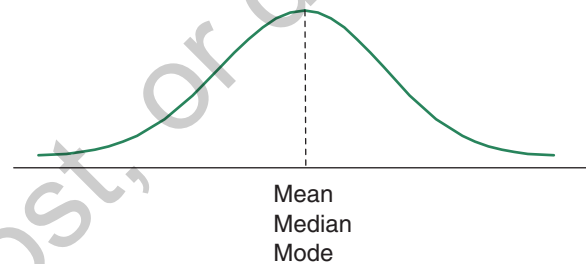


FIGURE 2.15 Positively Skewed (Right Skewed) Distribution With Mean, Median, and Mode

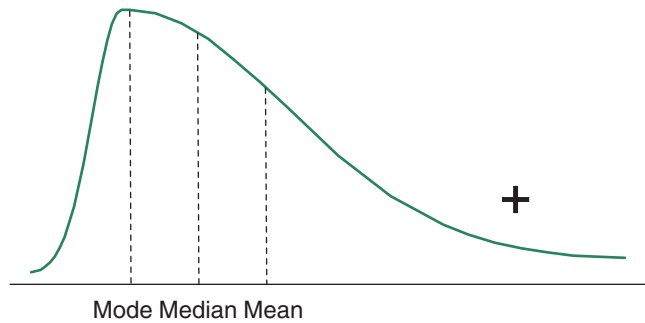

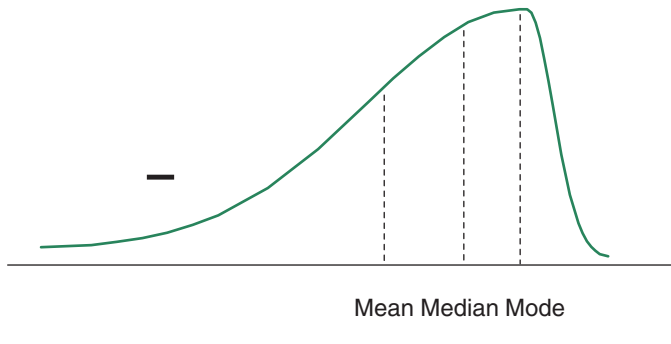


FIGURE 2.16  **Negatively Skewed (Left Skewed) Distribution With Mean, Median, and Mode**



Therefore, the long tail to the left side is pulling the mean toward the tail. The median is in between the mode and the mean.

There are three measures for central tendency. Which measure of central tendency is the best? The answer to that question is not simple. It depends on the shape of the distribution and the purpose for the central tendency. For example, let's use annual income in the United States as an example. Income is known to have a positive

skewed distribution. Few outliers with extremely high annual salary would have inflated the mean salary; therefore, the mean is less accurate in describing ordinary people's annual salary. The mean is higher than the median in a positively skewed distribution, because every number is used to calculate the mean, the outliers pull the mean toward the tail (high end). To avoid outliers unduly influencing the central tendency measures, the median is a more appropriate measure to represent ordinary people's annual salary than is the mean.



POP QUIZ

5. In a negatively skewed distribution, the mean would be
 - a. larger than the median.
 - b. smaller than the median.
 - c. equal to the median.
 - d. none of the above.
6. The normal distribution curve is symmetric around which point of the distribution?
 - a. mean
 - b. median
 - c. mode
 - d. all of the above

EXCEL STEP-BY-STEP INSTRUCTION FOR CALCULATING $\sum X$ FROM A FREQUENCY TABLE AND CONSTRUCTING A BAR GRAPH AND A HISTOGRAM

To calculate $\sum X$ using a frequency table in Excel, we use the network interruption example. First, enter the data showing daily network interruptions in September in Excel as shown in columns **A** and **B** in Figure 2.17.

The $\sum X$ formula needs to be adjusted to $\sum fX$ using the frequency table. You need to calculate the multiplicative product of the frequency (f) times the value X by moving the cursor to **C1** and label it as fX , then move to **C2** and type “ $=A2*B2$ ” for the multiplication. All formulas start with “=,” **A2** is the first value of X , and **B2** is the corresponding frequency. Then hit **enter**. Notice **C2** shows 0, which is the result of $A2*B2 = 0 \times 9$ as shown in Figure 2.17.

Then move the cursor to the lower right corner of **C2** until + shows up; then left click and hold the mouse and drag it to **C7**, and then release the mouse. You will see that the rest of the multiplicative products of fX automatically show up in **C3** to **C7**. Now you need to sum up all the fX values from **C2** to **C7**. Move the cursor to **C8** and type “ $=SUM(C2:C7)$ ”. Then hit **enter**. The answer for the total number of network interruptions is 43 as shown in Figure 2.18. If you have trouble following the description of steps here, I have good news for you. I created a 3-minute YouTube video for you. You may find it by using this URL <https://www.youtube.com/watch?v=14G91DzRMvs>.

The following graphing instructions apply to Microsoft Excel 2019. Graphing functions are likely to vary across different versions of Excel. However, Excel 2013, 2016, and 2019 all have the same functionality for creating a bar graph. To create a bar graph of daily network interruptions, first highlight the data range including the X values, their

FIGURE 2.17 Programming fX in Excel

	A	B	C	D	E
1	X	f	fX		
2	0	9	0		
3	1	10			
4	2	5			
5	3	3			
6	4	2			
7	6	1			
8					

FIGURE 2.18 Programming $\sum fX$ in Excel

	A	B	C	D	E
1	X	f	fX		
2	0	9	0		
3	1	10	10		
4	2	5	10		
5	3	3	9		
6	4	2	8		
7	6	1	6		
8			43		
9					

corresponding frequencies and also include the variable names; **A1:B7**. Once these two columns are highlighted, click the **Insert** tab on the top of the tool bar; then click on **Recommended Charts** as shown in Figure 2.19. A new pop-up window appears to display the recommended charts, choose the second chart, **Clustered Column**, on the left panel as shown in Figure 2.20, then click **OK**. Congratulations, you just created a basic bar chart inside Excel. If you want to refine the chart, you can create a chart title and axis title and place frequency on top of each bar in the chart, so your refined chart looks like the one shown in Figure 2.21. Learning by doing is the best way to get familiar with software functions. Click on different parts of the graphs to see options available. Try different options to see if you like the results. If you make mistakes (we all do), there is always the undo button on the upper left corner to restore the graph. Have fun creating your own graphs.

FIGURE 2.19 Finding the Bar Graph Function in Excel

	A	B	C	D	E	F	G	H	I	J
1	X	f	fX							
2	0	9	0							
3	1	10	10							
4	2	5	10							
5	3	3	9							
6	4	2	8							
7	6	1	6							
8			43							
9										

FIGURE 2.20 A Pop-Up Window Displaying Recommended Charts

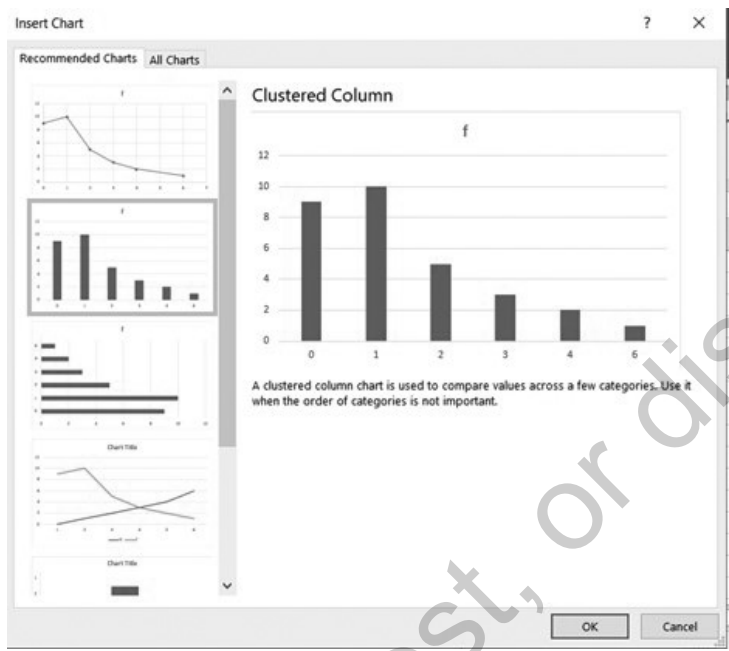
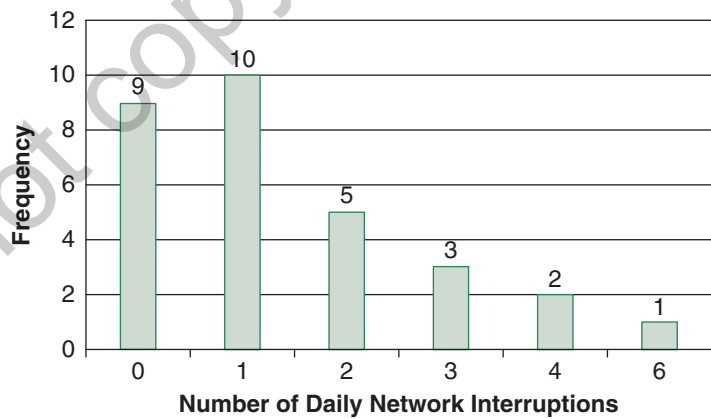


FIGURE 2.21 Refined Bar Chart With Axis Titles and Frequency on Top of Each Bar



EXERCISE PROBLEMS

- Workouts are necessary to keep you healthy. Although we all know the importance of keeping an active lifestyle, school and work demands usually get in the way. Here is a list of 20 college students' number of workouts during last week.
5, 3, 4, 4, 3, 3, 5, 2, 3, 1,
1, 2, 4, 3, 0, 5, 4, 3, 5, 5
 - Organize them into a frequency table of the number of workouts for the 20 college students.
 - Construct a bar graph based on the frequency table.
 - Identify the measures of central tendency for the number of workouts for these 20 college students.
- Student loans are at the historical record high. Many college students are incurring large amount of debts while pursuing their degrees. Here is a frequency table of student loans from a random sample of $n = 35$ students in Table 2.14.
 - Create a histogram based on the frequency table and comment on the shape of the distribution.
 - What is the estimated mean for the student loans based on the frequency table?

TABLE 2.14 Frequency Table of Student Loans

Student Loan (in Dollars)	f (Frequency)
0–9,999	1
10,000–19,999	1
20,000–29,999	2
30,000–39,999	3
40,000–49,999	4
50,000–59,999	6
60,000–69,999	8
70,000–79,999	10

Solutions

- The minimal value of the number of workouts is 0 and the maximal value is 5. The frequency table can be created using the individual values.

- a. The frequency table of 20 students' number of workouts is shown in Table 2.15.

TABLE 2.15 Frequency Table of Number of Workouts Among 20 Students Last Week

Number of Workouts	<i>f</i> (Frequency)
0	1
1	2
2	2
3	6
4	4
5	5
Total	20

- b. The bar graph of the number of workouts among 20 college students is shown in Figure 2.22.

FIGURE 2.22 Bar Graph of 20 Students' Number of Workouts Last Week



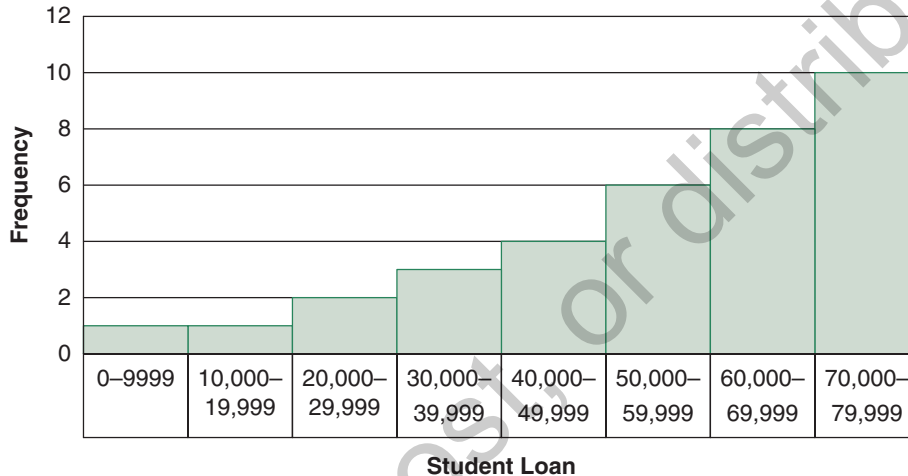
- c. The measures of central tendency are mode, median, and mean. The mode is 3, median is 3, and the mean is $\bar{X} = \frac{\sum X}{n} = \frac{\sum fX}{n} = \frac{65}{20} = 3.25$.

2.

- a. The histogram was created based on the frequency table of a random sample of 32 students' loans in Table 2.14. The distribution of student loans has a long tail to the left and peaks at the high end of the distribution. The distribution of student loans has the characteristics of a negatively skewed distribution. The histogram is shown in Figure 2.23. The difference between creating a bar graph versus a histogram in **Excel** is in the **Chart Tools**. The

default setting automatically creates bar graphs where bars are not touching one another. Click on **Format** tab, and click **Format Selection** on the upper left side. Then a pop-up window showing **Format Data Series** appears. The second sliding ruler shows **Gap Width**. Move the ruler all the way to the left, 0% showing no gap between bars. Your bar graph turns into a histogram.

FIGURE 2.23 Histogram of Student Loans



- b. To calculate the estimated mean of the student loan based on the frequency table, you need to calculate the midpoint in each interval as shown in the X_{midpoint} column in Table 2.15a. Consider the impact of the frequency, you need to do the multiplication of $f \times X_{\text{midpoint}}$ as shown in the last column of Table 2.15a then sum up the numbers at the bottom of the table. The formula for the estimated mean needs to be modified to

$$\bar{X} = \frac{\sum fX_{\text{midpoint}}}{n} = \frac{1954983}{35} = 55856.6$$

The estimated mean for the

35 students' loan is \$55,856.6. Many students have expressed that the student loan feels like a weight tied around their neck forcing them to assume jobs with the highest paychecks instead of pursuing their passion.

TABLE 2.15a Frequency Table of Student Loans With Midpoint in Each Interval

Student Loan (in Dollars)	f (Frequency)	X_{midpoint}	fX_{midpoint}
0–9,999	1	4,999.5	4,999.5
10,000–19,999	1	14,999.5	14,999.5

Student Loan (in Dollars)	f (Frequency)	X_{midpoint}	fX_{midpoint}
20,000–29,999	2	24,999.5	49,999
30,000–39,999	3	34,999.5	104,998.5
40,000–49,999	4	44,999.5	1,79,998
50,000–59,999	6	54,999.5	3,29,997
60,000–69,999	8	64,999.5	5,19,996
70,000–79,999	10	74,999.5	7,49,995
Total	35		19,54,983

What You Learned

In this chapter, you have learned to construct frequency tables and graphs to present a quick glance and easy-to-understand data summary. When constructing a frequency table, here are the keys to success:

1. Identify the minimal and maximal values of the variable. Calculate the range.
2. Construct a frequency table with 7 ± 2 categories. When the range is too large to construct a table with individual values, choose an appropriate and convenient interval, usually in 5s or 10s, if feasible.
3. Create the categories or equal intervals in an ascending order.
4. Make sure that there are no overlaps between any two adjacent intervals. For continuous variables, each interval includes the lower limit but not the upper limit.

In bar graphs and histograms, the graphs consist of bars of equal width. The bar width presents the value or interval. The bar height represents the frequency or relative frequency of each value or interval. Bar graphs are used to organize discrete variables, and histograms are used to organize continuous variables. In pie charts, the graphs consist of different slices. The size of the slice is proportional to the frequency or relative frequency of each category.

Central tendency is defined as using one single value to represent the center of the data. There are three common measures: mode, median, and mean.

1. Mode is the value with the highest frequency in the distribution.
2. Median is the midpoint of a sorted distribution. Median is not affected much by extreme outliers.

3. Mean is the arithmetical average of all values. The formula for the population mean is $\mu = \frac{\sum X}{N}$ and the formula for the sample mean is $\bar{X} = \frac{\sum X}{n}$. When using a frequency table to calculate mean, the formula needs to be modified to $\bar{X} = \frac{\sum fX}{n}$ to incorporate the impact of different frequencies on individual values in the table. Same modification also applies to $\mu = \frac{\sum fX}{N}$. Mean is highly sensitive to the impact of outliers.
4. When estimated mean is calculated from a frequency table with equal intervals, the formula needs to be modified to $\bar{X} = \frac{\sum fX_{\text{midpoint}}}{n}$.

Common distribution shapes include uniform distribution, normal distribution, and skewed distribution.

Key Words

Bar graph: A bar graph uses bars of equal width to show frequency or relative frequency of discrete categorical data (i.e., nominal or ordinal data). Adjacent bars are not touching each other. 37

Central tendency: Central tendency is defined as utilizing a single value to represent the center of a distribution. There are three commonly used measures for central tendency: mean, median, and mode. 45

Cumulative relative frequency: Cumulative relative frequency is defined as the accumulation of the relative frequency for a particular value and all the relative frequencies of lower values. 48

Data visualization: Data visualization is the graphic representation of data. 37

Distribution: Distribution is the arrangement of values of a variable as it occurs in a sample or a population. 29

Equal intervals: Equal intervals are created in a frequency table by including the same number of values in each interval. 33

Frequency: Frequency is a simple count of a particular value or category occurring in a sample or a population. 29

Frequency distribution table: A frequency distribution table lists all distinct values or categories arranged in an orderly fashion in a table, along with tally counts for each value or category in a data set. 29

Histogram: A histogram is a graphical presentation of a continuous variable. The bars of a histogram are touching each other to illustrate that the values are continuous. 38

Line graph: A line graph is a graphical display of quantitative information with a line or curve that connects a series of adjacent data points. 43

Mean: the mean is defined as the arithmetic average of all values in a data distribution. 49

Median: the median is defined as the value in the middle position of a sorted variable arranged from the lowest value to the highest value. 47

Mode: The mode is defined as the value or category with the highest frequency in a data distribution. 45

Negatively skewed distribution: When the data points mostly concentrate on the high-end values with a long tail to the low-end values, the distribution is negatively skewed. 44

Normal distribution: A normal distribution curve peaks at the mean of the values and symmetrically tapers off on both sides with 50% of the data points above the mean and 50% of the data points below the mean. Draw a line straight through the mean, and the left side and the right side are mirror images of each other. 43

Outliers: Outliers refer to extreme values in a distribution. Outliers usually stand far away from the rest of the data points. 55

Pie chart: It is a circular graph that uses slices to show different categories. The size of each slice is proportional to the frequency or relative frequency of each category. 40

Positively skewed distribution: When the data points mostly concentrate on the low-end values with a long tail to the high-end values, the distribution is positively skewed. 44

Real limits: Real limits cover a range of possible values that may be reflected by a single continuous measure. The lower limit is the value minus $\frac{1}{2}$ of the unit and the upper limit is the value plus $\frac{1}{2}$ of the unit. 34

Relative frequency: Relative frequency is defined as the frequency of a particular value or category divided by the total frequency (or sample size). Relative frequency is also called proportion. 30

Skewed distribution: A skewed distribution happens when data points are not symmetrical, and they concentrate more on one side of the mean than on the other. 44

Uniform distribution: In a uniform distribution, every value appears with the same frequency, proportion, or probability. 41

Weighted mean: The weighted mean is defined as calculating the mean when data values are assigned different weights, w . The formula for weighted mean is $\bar{X} = \frac{\sum wX}{\sum w}$. 53

Learning Assessment

Multiple Choice Questions: Choose the Best Answer in the Following Questions.

1. A _____ skewness is a distribution with a high concentration of data points at the high end of the distribution and a long tail toward the low end of the distribution.

- a. negative
 - b. positive
 - c. discrete
 - d. continuous
2. A proportion describes the frequency (f) of a category in relation to the total frequency or sample size (n), f/n . It is also called a _____.
 - a. frequency
 - b. relative frequency
 - c. cumulative relative frequency
 - d. correlation
 3. The appropriate number of categories in a frequency table is
 - a. 5 ± 2
 - b. 7 ± 2
 - c. 9 ± 2
 - d. 11 ± 2
 4. What is the sample size according to the frequency table of students' homework grades?

X (Homework Grade)	f (Frequency)
0	2
6	4
7	4
8	12
9	14
10	17

- a. 40
 - b. 41
 - c. 51
 - d. 53
5. Calculate mean quiz score of students based on the following frequency table.

X (Quiz Grade)	f (Frequency)
0	2
1	4

X (Quiz Grade)	f (Frequency)
2	4
3	12
4	14
5	17

- 2.7
 - 3.1
 - 3.6
 - 4.8
6. What is wrong with the following frequency table?

X (Exam Grade)	f (Frequency)
40–50	2
50–60	4
60–70	4
70–80	12
80–90	4
90–100	7

- The frequency table does not consist of equal intervals
 - The number of intervals is too large
 - The number of intervals is too small
 - There are overlaps between two adjacent intervals
7. In a sample where outliers are present, the worst statistical measure to represent the central tendency is _____.
- the mean
 - the median
 - the mode
 - all of the above
8. In a positively skewed distribution, scores with the highest frequencies are _____.
- on the high end (right side) of the distribution
 - on the low end (left side) of the distribution
 - in the middle of the distribution
 - represented at two distinct peaks

9. In a frequency distribution, the _____ is the score or category that has the highest frequency.
- mean
 - median
 - mode
 - range
10. When there are 12 observations in a small sample, the median is calculated by
- the sixth position of an ascending sorted data
 - the average value of the sixth position and the seventh position of an ascending sorted data
 - the sixth position of the data as they were presented without sorting
 - the average value of the sixth position and the seventh position of the data as they were presented without being sorted
11. Mean is an appropriate statistical measure for central tendency for _____ scales of measurement.
- nominal, ordinal, interval, and ratio
 - ordinal, interval, and ratio
 - both interval and ratio
 - only ratio
12. In a positively skewed distribution, the mean would be
- larger than the median
 - smaller than the median
 - equal to the median
 - none of the above

Free Response Questions

13. There is a small city in Ohio with a population of 179 people and 0.27 miles on a busy interstate highway. It is famous for its speed traps. A speed trap is defined as a section of a road where police, radar, or traffic cameras frequently check the speed of motorists and strictly enforce traffic regulations. A municipality can intentionally lower the speed limits to catch motorists cruising by without paying attention to the speed signs to boost its revenues. Table 2.16 shows a summary of citation amount intervals and frequency associated with each interval. You may answer the question by using your calculator and draw the histogram by hand or you may use the Excel step-by-step instructions to answer the following questions.
- What was the total dollar amount collected from the traffic citations in this small city in the past 12 months?

- b. What is distribution shape of the traffic citations? Create a histogram and comment on the shape of the distribution of traffic citations.
- c. What are the mode, median, and mean of the traffic citations?

TABLE 2.16 Frequency Table of Traffic Citation Amounts

X (Citation Amount in Dollars)	f (Frequency)
90–109	12
130–149	25
150–169	46
170–189	105
190–209	852
210–229	907
230–249	981

14. A summary of all employees' annual salary from a local company is shown in Table 2.17. You may answer the question by using your calculator and draw the histogram by hand or you may use the Excel step-by-step instructions to answer the following questions.
- a. What is the distribution shape of the employees' income? Create a histogram of employees' annual salary and comment on the shape of the distribution.
- b. What are the mode, median, and the estimated mean of employees' annual salary?

TABLE 2.17 Frequency Table of Employees' Annual Salary

X (Employee Annual Salary)	f (Frequency)
20,000–39,999	329
40,000–59,999	342
60,000–79,999	198
80,000–99,999	60
100,000–119,999	13
120,000–139,999	3
140,000–159,999	2
160,000–179,999	1

15. A brief survey of college students revealed their reported hand washing routine immediately after returning to their private living space from outside during COVID-19 pandemic. The results are shown in Table 2.18. Use the frequency numbers in the table to create a bar chart to describe the shape of the distribution of college students' hand washing behavior during the pandemic.

TABLE 2.18 Frequency Table of College Students' Hand Washing Behavior

X (Hand Washing Behavior)	f (Frequency)
Never	2
Seldom	4
Sometimes	23
Most of the time	39
Always	53

Answers to Pop Quiz Questions

1. c
2. b
3. d
4. a
5. b
6. d