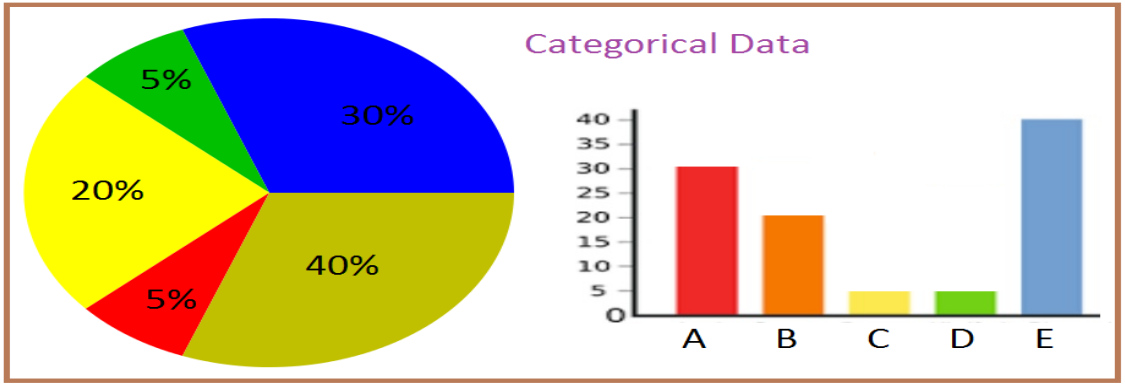
Introductory Statistics: A Problem-Solving Approach

by Stephen Kokoska

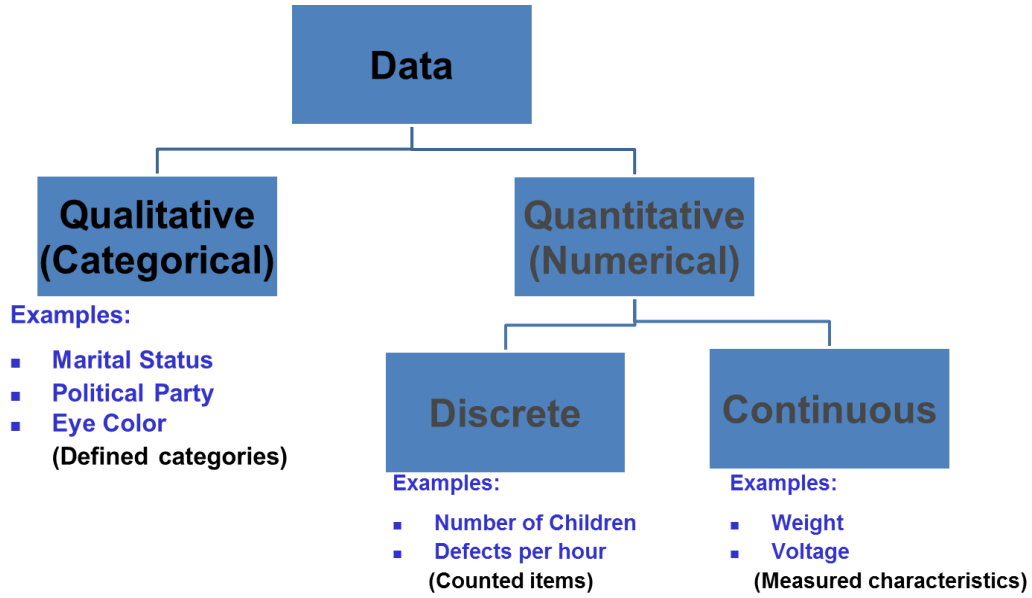
Chapter 2

Tables and Graphs for Summarizing Data



Types of Data

- A data set consisting of observations of only a single characteristic, or attribute, is a **univariate** data set.
- If we measure, or record, two observations of each individual or object, the data set is **bivariate**.
- If there are more than two observations of the same individual or object, the data set is **multivariate**.
- A **categorical**, or **qualitative**, univariate data set consists of non-numerical observations that may be placed in categories.
- A **numerical**, or **quantitative**, univariate data set consists of observations that are numbers.
  - A numerical data set is **discrete** if the set of all possible values is finite, or countably infinite. Discrete data sets are usually associated with counting.
  - A numerical data set is **continuous** if the set of all possible values is an interval of numbers. Continuous data sets are usually associated with measuring.



Displaying Data Distributions with Graphs

To examine a single variable, we graphically display its **distribution.**

- The distribution of a variable tells us what values it takes and how often it takes these values.
- Distributions can be displayed using a variety of graphical tools. The proper choice of graph depends on the nature of the variable.

Categorical variable
Pie chart
Bar graph

Quantitative variable
Histogram
Stem-and-leaf plot

Bar Charts and Pie Charts

A **frequency distribution** for categorical data is a summary table that presents categories, counts, and proportions.

1. Each unique value in a categorical data set is a label, or class.
2. The frequency is the count for each class.
3. The relative frequency, or sample proportion, for each class is the frequency of the class divided by the total number of observations.

Example: Online Shopping

A random sample of e-commerce companies in the US was obtained. The category for each company is given here.

Pet products	Health	Sporting goods	Artisanal	Sporting goods
Sporting goods	Artisanal	Sporting goods	Sporting goods	Grocery
Health	Sporting goods	Pet products	Sporting goods	Artisanal
Artisanal	Grocery	Health	Pet Products	Sporting goods
Sporting goods	Sporting goods	Artisanal	Health	Artisanal

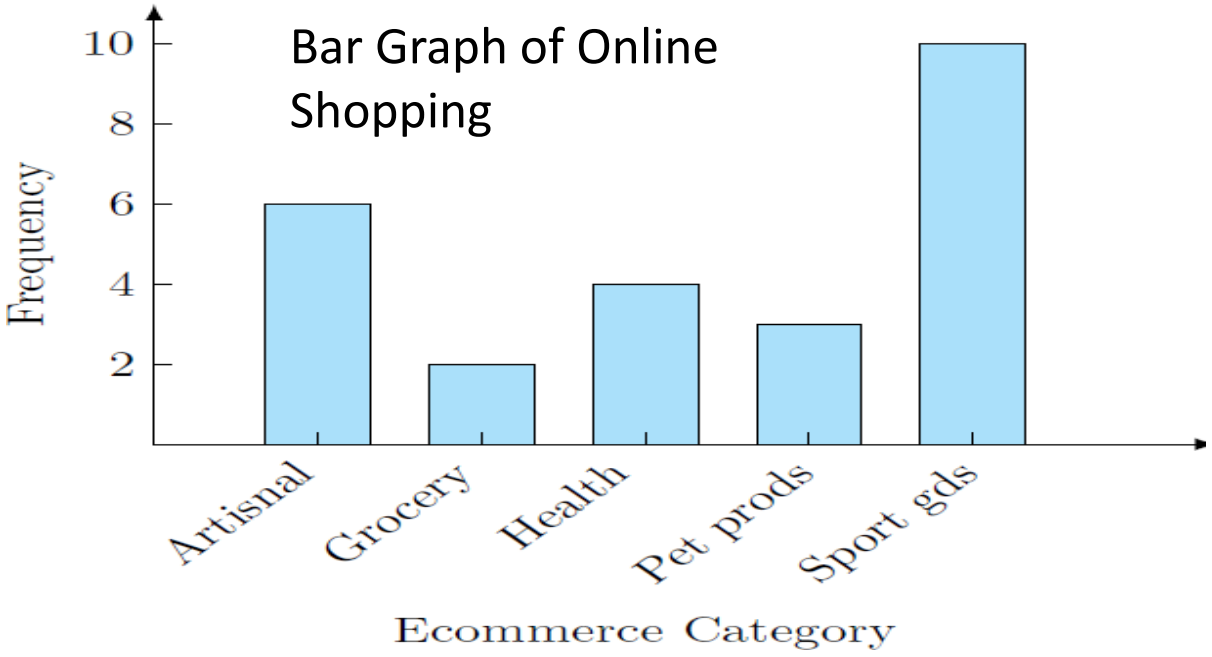
Construct a frequency distribution to describe these data.

What proportion of e-commerce companies are not classified as Health?

Frequency Distribution for Online Shopping

Class	Frequency	Relative Frequency	
Artisanal	6	0.24	(= 6/25)
Grocery	2	0.08	(= 2/25)
Health	4	0.16	(= 4/25)
Pet products	3	0.12	(= 3/25)
Sporting goods	10	0.40	(= 10/25)
Total	25	1.00	

The proportion of e-commerce categories that are Health related is 4/25 = 0.16. The total proportion is always 1.00. Therefore, the proportion of e-commerce categories that are not Health related is 1.00 – 0.16 = 0.84

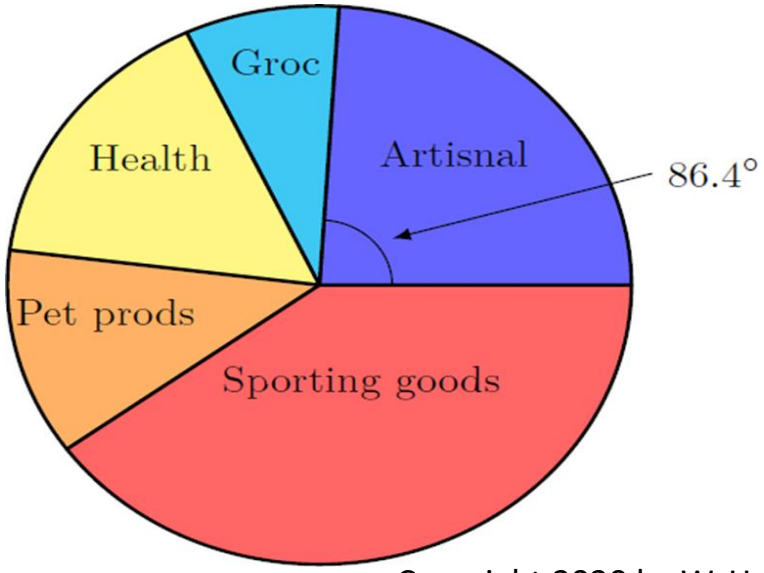


Pie Charts

How to Construct a Pie Chart

1. Divide a circle (or pie) into slices or wedges so that each slice corresponds to a class.
2. The size of each slice is measured by the angle of the slice. To compute the angle of each slice, multiply the relative frequency by 360° (the number of degrees in a whole or complete circle).
3. The first slice of a pie chart is usually drawn with an edge horizontal and to the right (0°). The angle is measured clockwise. Each successive slice is added counterclockwise with the appropriate angle.

Example: Pie Chart for Online Shopping



Class	Relative Frequency	Angle
Artisanal	0.24	86.4° (= 0.24 × 360°)
Grocery	0.08	28.8° (= 0.08 × 360°)
Health	0.16	57.6° (= 0.16 × 360°)
Pet products	0.12	43.2° (= 0.12 × 360°)
Sporting goods	0.40	144.0° (= 0.40 × 360°)
Total	1.00	

Example Stem-and Leaf Plot

Stem-and-leaf plot: To create this plot, each observation in the data set must have at least two digits. Think of each observation as consisting of two pieces: a stem and a leaf.

Kerepakupai Meru, or Angel Falls, is the highest waterfall in the world. Suppose the table lists the total height, in meters, of several waterfalls in the world.

```
The decimal point is 1
digit(s) to the right
of the |

60 | 000
61 | 00000000002
62 | 0059
63 | 158
64 | 0056
65 | 01
66 | 005
67 | 114
68 | 0
69 | 3
70 | 05677
71 | 559
72 | 057
73 | 289
74 | 5
```

693	745	631	635	625	629	739	738	732	725
720	719	715	715	707	707	706	705	700	680
674	671	671	665	660	660	650	646	645	640
640	638	620	620	612	610	610	610	610	610
610	610	610	610	610	600	600	600	651	727

Figure 2.13
R stem-and-leaf plot.

Constructing a Frequency Distribution

A **frequency distribution** for numerical data is a summary table that displays classes, frequencies, relative frequencies, and cumulative relative frequencies.

How to Construct a Frequency Distribution

1. Choose a range of values that captures all of the data. Divide it into non-overlapping (usually equal length) intervals. Each interval is called a class, or class interval. The endpoints of each class are the class boundaries.
2. We use the left-endpoints. An observation equal to an endpoint is allocated to the class with that value as its lower endpoint. Hence, the lower-class boundary is always included in the interval, and the upper-class boundary is never included. This ensures that each observation falls into exactly one interval.
3. Use 5-20 intervals.
4. Count the number of observations in each class interval. This count is the class frequency or frequency.
5. Compute the proportion of observations in each class. This ratio is the relative frequency.

Find the cumulative relative frequency (CRF) for each class: the sum of all the relative frequencies of classes up to and including that class. This column is a running total, or accumulation, of relative frequency by row.

Example: Frequency Distribution

Torque is a measure of the force needed to cause an object to rotate. It is usually measured in foot-pounds (ft-lb). As part of a quality-control program, Whirlpool inspectors measure the initial torque needed to loosen the balancing bolts on each leg of a clothes washer. Construct a frequency distribution for this random sample of measurements.

20.4	41.3	13.0	24.1	11.0	44.4	28.4	37.5	16.9	53.4
36.4	14.9	62.1	25.6	63.7	31.7	43.5	57.2	23.1	45.7
24.2	38.1	35.5	51.1	26.4					

e.g., we have 7 measurements in the 20-30 class and they represent 28%.

From the CRF: About 64% of the measurements are less than 40.

Class	Frequency	Relative frequency	Relative frequency	Cumulative relative frequency	Cumulative relative frequency
10-20	4	0.16	(=4/25)	0.16	(=0.16)
20-30	7	0.28	(=7/25)	0.44	(=0.16 + 0.28)
30-40	5	0.20	(=5/25)	0.64	(=0.44 + 0.20)
40-50	4	0.16	(=4/25)	0.80	(=0.64 + 0.16)
50-60	3	0.12	(=3/25)	0.92	(=0.80 + 0.12)
60-70	2	0.08	(=2/25)	1.00	(=0.92 + 0.08)
Total	25	1.00			

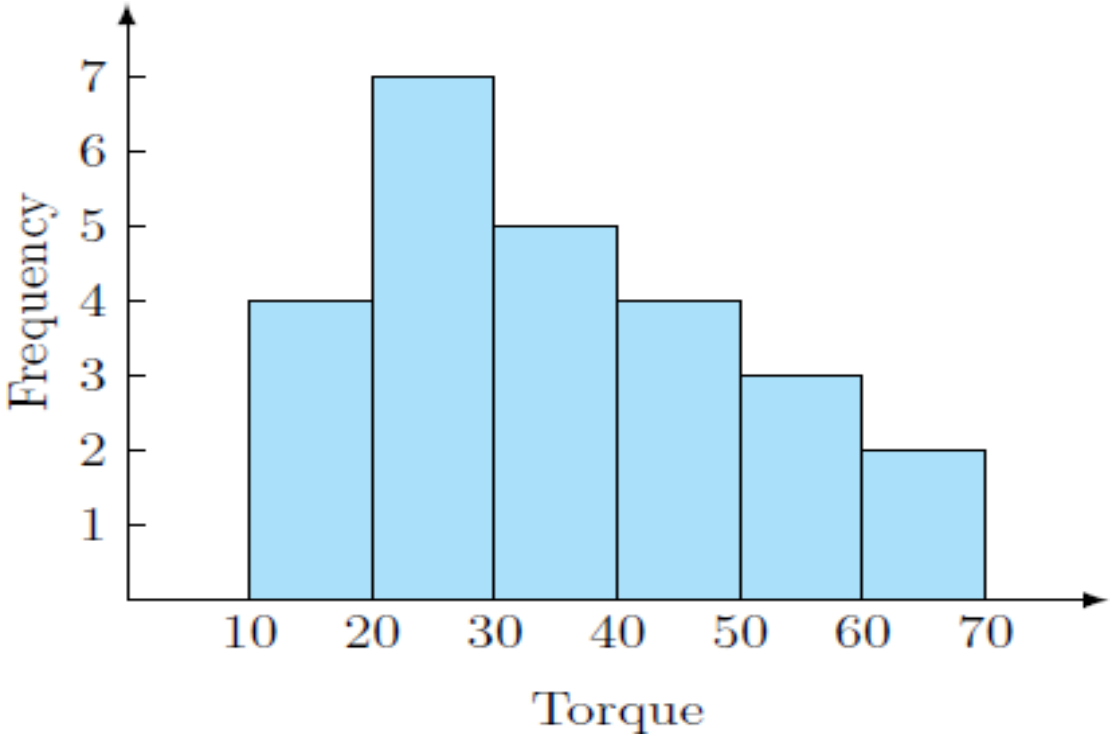
Histograms

A histogram is a graphical representation of a frequency distribution, a plot of frequency versus class interval.

How to Construct a Histogram

1. Draw a horizontal axis and place tick marks corresponding to the class boundaries
2. Draw a vertical axis and place tick marks corresponding to frequency. Label each axis
3. Draw a rectangle above each class with height equal to frequency

Class	Frequency
10-20	4
20-30	7
30-40	5
40-50	4
50-60	3
60-70	2
Total	25



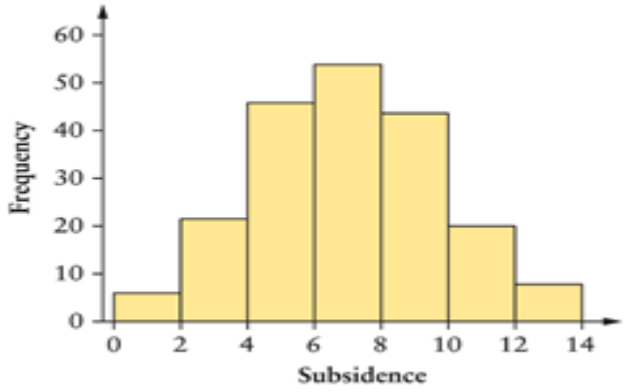
Frequency Histogram of Torque

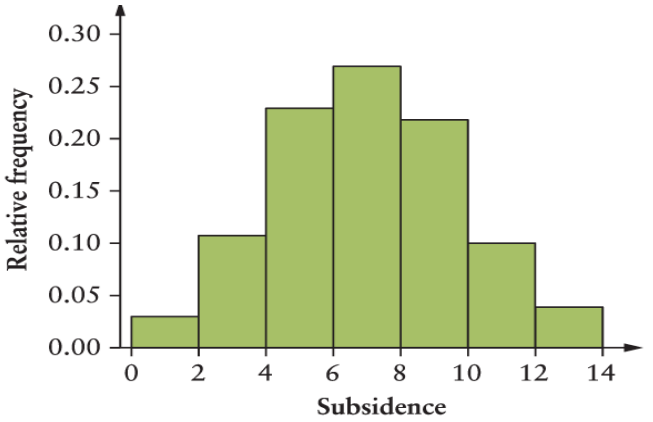
Frequency Versus Relative Frequency Histograms

Example: A random sample of locations in the San Francisco Bay area was obtained, and the yearly subsidence rate (in mm) was measured.

Class	Frequency	Relative frequency	Cumulative relative frequency
0–2	6	0.03	0.03
2–4	22	0.11	0.14
4–6	46	0.23	0.37
6–8	54	0.27	0.64
8–10	44	0.22	0.86
10–12	20	0.10	0.96
12–14	8	0.04	1.00
Total	200	1.00	

- A histogram tells us about the shape, center, and variability of the distribution, and allows us to quickly identify any outliers.
- To construct a histogram by hand, you must construct the frequency distribution first. Calculators and computers can construct the histogram directly from the data.
- To construct a *relative* frequency histogram, plot relative frequency versus class interval. The only difference between a frequency histogram and a relative frequency histogram is the scale on the vertical axis. The two graphs are identical in appearance.
- Histograms should not be used for inference. They provide a quick look at the distribution, and only suggest certain characteristics.





Shape of a Distribution

- The shape of a distribution, represented in a histogram, is an important characteristic.
- To help describe the various shapes, we draw a smooth curve along the tops of the rectangles that captures the general nature of the distribution (as shown in Figure 2.25).
- To help identify and describe distributions quickly, a smoothed histogram is often drawn on a graph without a vertical axis, without any tick marks on the measurement axis, and without any rectangles (as shown in Figure 2.26).

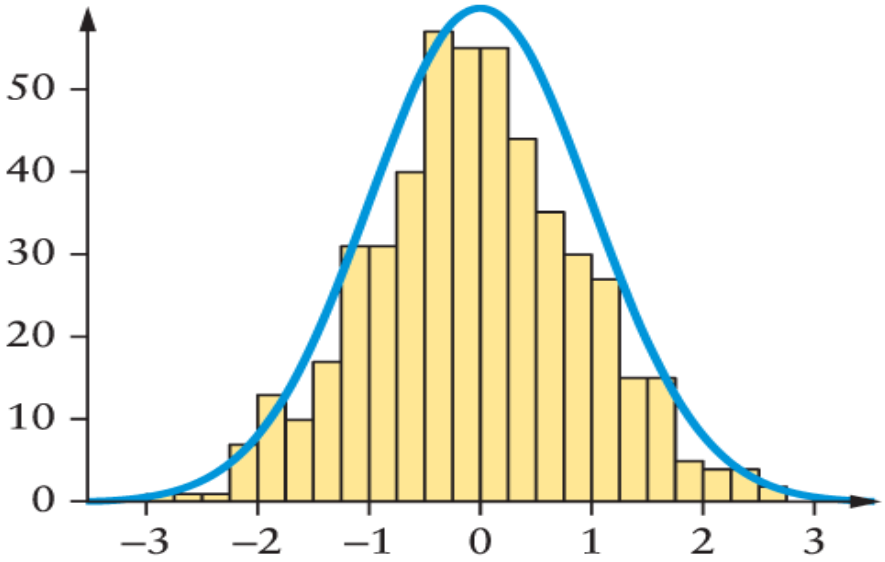


Figure 2.25

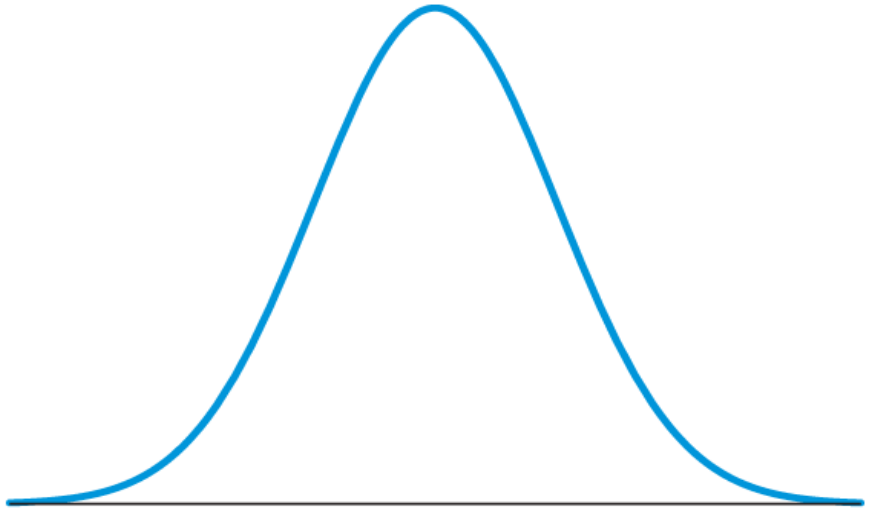
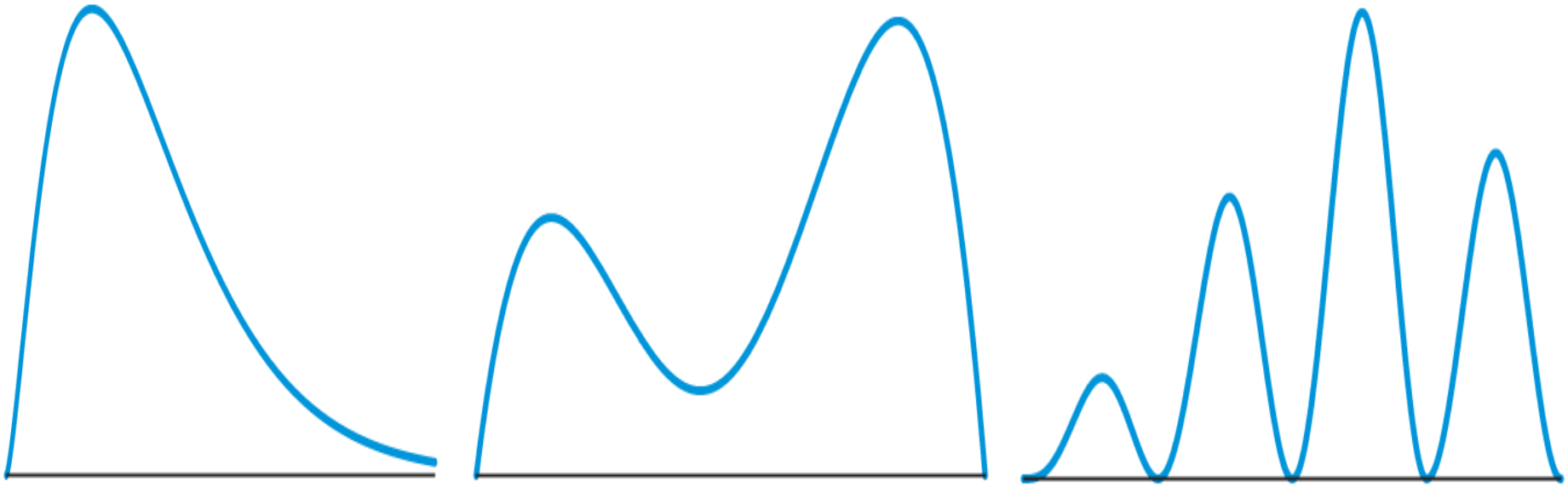


Figure 2.26

Shape of a Distribution

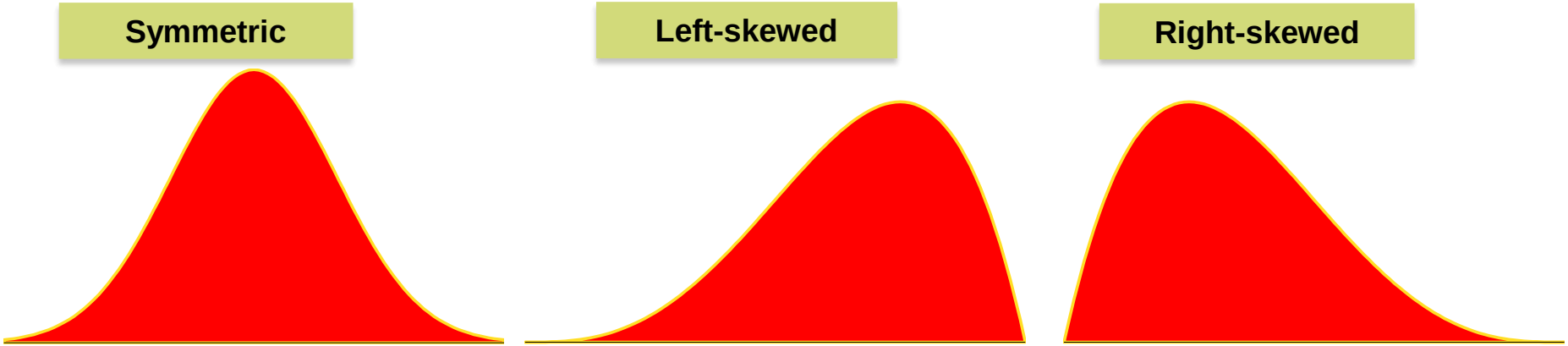
Definition

1. A **unimodal** distribution had one peak. It is very common and almost all distributions have a single peak.
2. A **bimodal** distribution has two peaks. It is very common and may occur if data from two different populations are accidentally mixed.
3. A **multimodal** distribution has more than one peak and it is very rare.



Examining Distributions

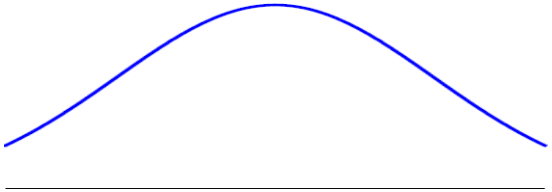
- In any graph of data, look for the overall pattern and for striking deviations from that pattern.
- You can describe the overall pattern by its shape, center, and spread.
- An important kind of deviation is an outlier, an individual that falls outside the overall pattern.
- A distribution is **symmetric** (bell-shaped) if the right and left sides of the graph are approximately mirror images of each other (there is a vertical line of symmetry in the distribution).
- If a unimodal distribution is not symmetric, then it may be skewed.
  - A distribution is **skewed to the right** (right-skewed) if the right side of the graph is much longer than the left side.
  - It is **skewed to the left** (left-skewed) if the left side of the graph is much longer than the right side.



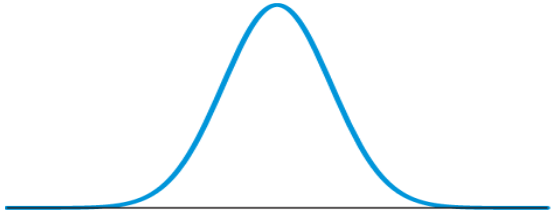
Examining Distributions

Spread

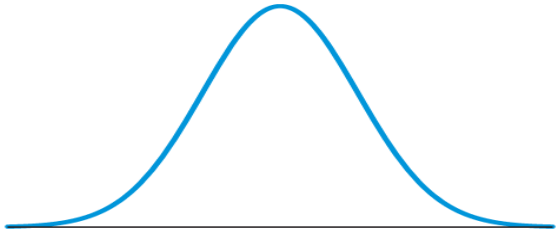
The **lower tail** of a unimodal distribution is the leftmost portion of the distribution. The **upper tail** is the rightmost portion of the distribution.



A distribution with heavy tails



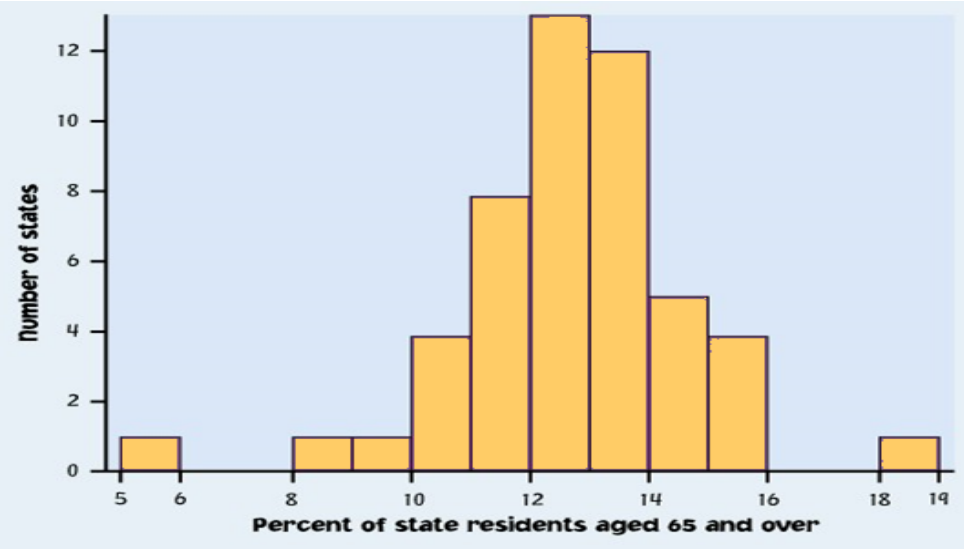
A distribution with light tails



Normal curve

An important kind of deviation is an **outlier.** Outliers are observations that lie outside the overall pattern of a distribution. Always look for outliers and try to explain them.

- The overall pattern is fairly symmetrical except for two states that clearly do not belong to the main pattern. Alaska and Florida have unusually small and large percents, respectively, of elderly residents in their populations.
- A large gap in the distribution is typically a sign of an outlier.



EXAMPLE 2.14 Solar Power

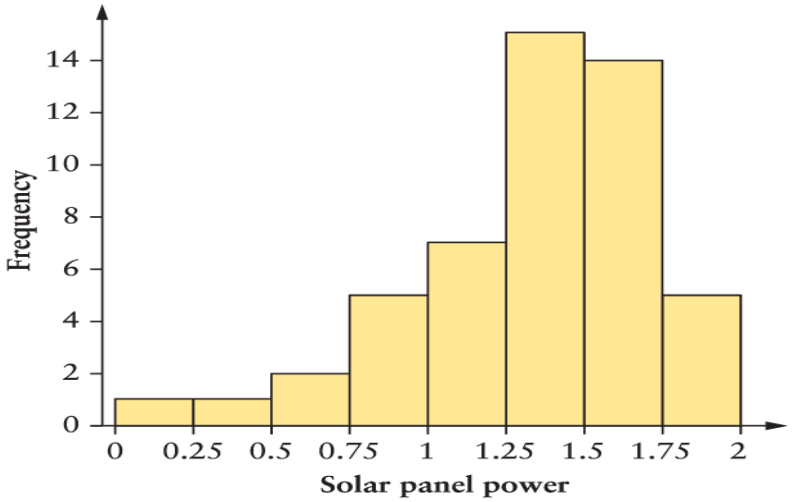
The amount of energy generated by a solar panel depends on the amount of direct sunlight and the panel's theoretical power production. Suppose a sample of 50 solar panels located on the roofs of residential homes was obtained and a daily power output of each was measured (in kWh). The data are given in the textbook.

a. Describe the shape, center, and spread of the distribution.
b. What proportion of observations is less than 1 kWh?
c. What proportion of observations is at least 1.5 kWh?
d. How many solar panels that their daily power output are less than 0.75 kWh?
e. How many solar panels that their daily power output are at least 1.50 kWh?
f. What proportion of observations is between 1.25 and 1.50 kWh?

Class	Frequency	Relative frequency		Cumulative relative frequency	
0.00–0.25	1	0.02	$(= 1/50)$	0.02	$(= 0.02)$
0.25–0.50	1	0.02	$(= 1/50)$	0.04	$(= 0.02 + 0.02)$
0.50–0.75	2	0.04	$(= 2/50)$	0.08	$(= 0.04 + 0.04)$
0.75–1.00	5	0.10	$(= 5/50)$	0.18	$(= 0.08 + 0.10)$
1.00–1.25	7	0.14	$(= 7/50)$	0.32	$(= 0.18 + 0.14)$
1.25–1.50	15	0.30	$(= 15/50)$	0.62	$(= 0.32 + 0.30)$
1.50–1.75	14	0.28	$(= 14/50)$	0.90	$(= 0.62 + 0.28)$
1.75–2.00	5	0.10	$(= 5/50)$	1.00	$(= 0.90 + 0.10)$
Total	50	1.00			

EXAMPLE 2.14 Solar Power

- The distribution is negatively skewed. The observations are clustered in the upper tail, and the lower tail extends farther than the upper tail.
- To estimate the center of the distribution, use the histogram to identify a value such that approximately half of the observations are below that number and half are above that number. A number between 1.25 and 1.50 appears to divide the ordered data in half. Typical values for this data set are in this range, and an estimate of the center is 1.38.
- The variability is typically described as either compact (data that are compressed or squeezed together) or spread out (observations that extend over a wide range). Although this distinction is somewhat subjective for now, this data set seems fairly compact.



EXAMPLE 2.14 Solar Power

b. The proportion of observations less than 1 is 0.18

c. The proportion of observations that is at least 1.5

i. 0.28 + 0.10 = 0.38

ii. Proportion of observations ≥ 1.50 = 1 − (proportion of observations < 1.50)

= 1 − 0.62 = 0.38

d. Solar panels that their daily power output are < 0.75 = 4

e. Solar panels that their daily power output are at least 1.50 = 19

f. The proportion of observations that is between 1.25 and 1.50 = 30%

Class	Frequency	Relative frequency		Cumulative relative frequency	
0.00–0.25	1	0.02	$(= 1/50)$	0.02	$(= 0.02)$
0.25–0.50	1	0.02	$(= 1/50)$	0.04	$(= 0.02 + 0.02)$
0.50–0.75	2	0.04	$(= 2/50)$	0.08	$(= 0.04 + 0.04)$
0.75–1.00	5	0.10	$(= 5/50)$	0.18	$(= 0.08 + 0.10)$
1.00–1.25	7	0.14	$(= 7/50)$	0.32	$(= 0.18 + 0.14)$
1.25–1.50	15	0.30	$(= 15/50)$	0.62	$(= 0.32 + 0.30)$
1.50–1.75	14	0.28	$(= 14/50)$	0.90	$(= 0.62 + 0.28)$
1.75–2.00	5	0.10	$(= 5/50)$	1.00	$(= 0.90 + 0.10)$
Total	50	1.00			