Chapter 2

Organizing Data

Section 2.1

Variables and Data

Definition 2.1

Variables

Variable: A characteristic that varies from one person or thing to another.

Qualitative variable: A nonnumerically valued variable.

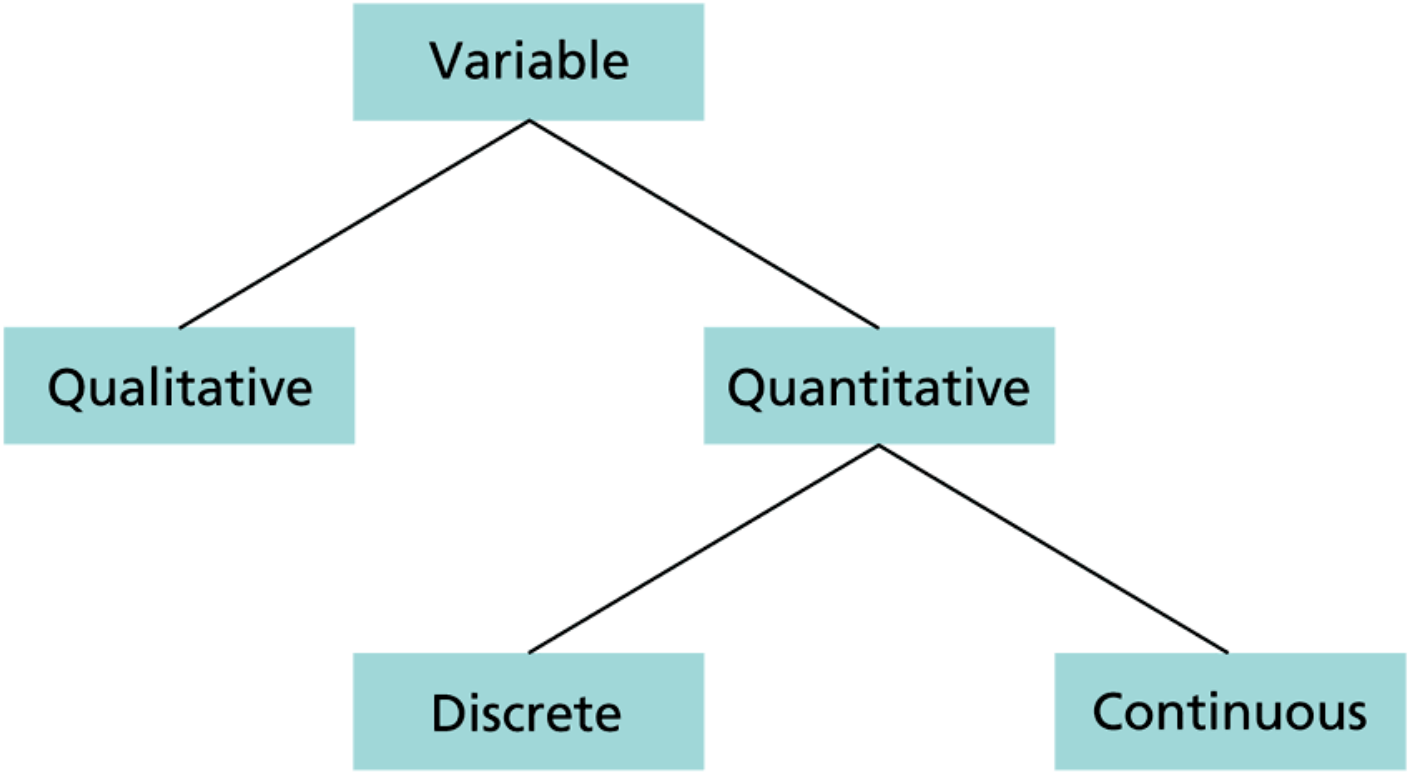
Quantitative variable: A numerically valued variable.

Discrete variable: A quantitative variable whose possible values can be listed.

Continuous variable: A quantitative variable whose possible values form some interval of numbers.

Figure 2.1

Types of variables



Let's produce some data from the class

Watch the following commercials from 10 Famous Funny Commercials (http://www.youtube.com/watch?v=HE9nLWFZ6ac) and answer the following two questions:

1. After watching the first two commercials, do you think a) is better or b) is better (in terms of getting the point through to you)? Write down your choice.
2. Based a score ranged from 1 to 10, please rated both commercials, again. The evaluations are based on whether if the advertisement has get the point through to you, a lowest score of 1 means the advertisement is poorly done, and a highest score of 10 means the advertisement is excellently done.

Definition 2.2

Data

Data: Values of a variable.

Qualitative data: Values of a qualitative variable.

Quantitative data: Values of a quantitative variable.

Discrete data: Values of a discrete variable.

Continuous data: Values of a continuous variable.

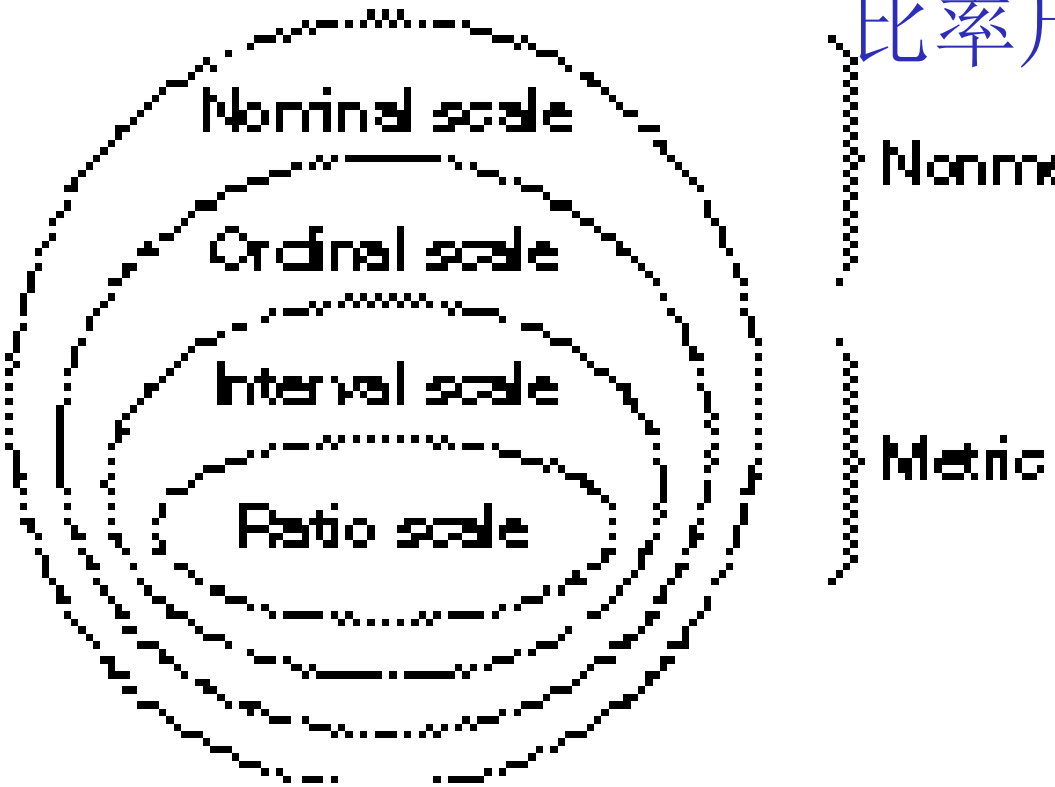
變數資料之類型

名目尺度(nominal scale)

序列尺度(ordinal scale)

區間尺度(interval scale)

比率尺度(ratio scale)



Nominal Level Data

Nominal level: Data that is classified into categories and cannot be arranged in any particular order, because Nominal level data are mutually exclusive and exhaustive. Furthermore, data categories have no logical order.

Mutually exclusive: An individual, object, or measurement is included in only one category.

Exhaustive: Each individual, object, or measurement must appear in one of the categories.

Nominal Level Data

Numbers are used to classify or categorize

- Example: Employment Classification
  - 1 for Educator
  - 2 for Construction Worker
  - 3 for Manufacturing Worker
- Example: Ethnicity
  - 1 for African-American
  - 2 for Anglo-American
  - 3 for Hispanic-American

Ordinal Level Data

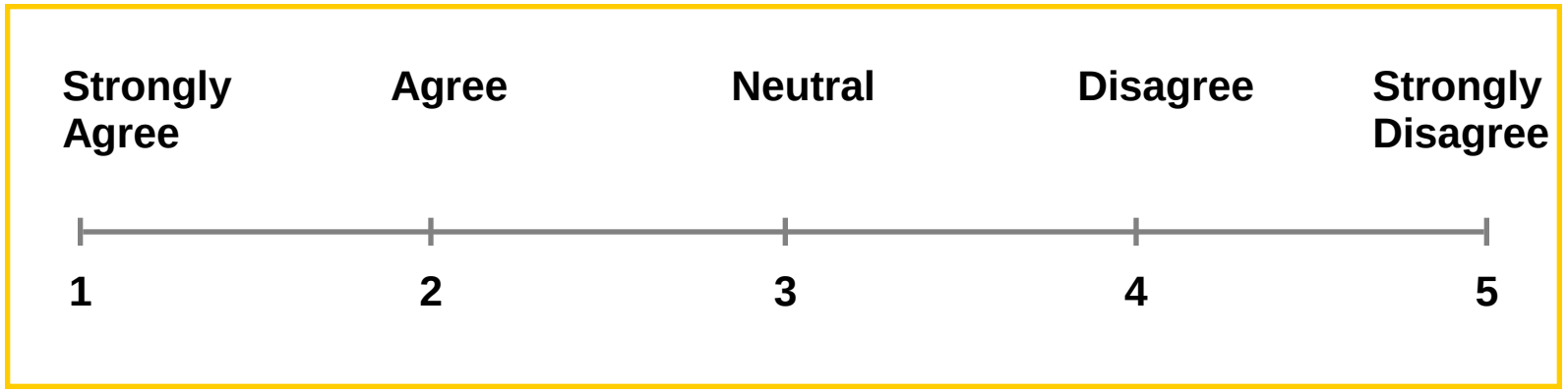
Numbers are used to indicate rank or order

- Relative magnitude of numbers is meaningful
- Differences between numbers are not comparable

- Example: Ranking productivity of employees
- Example: Taste test ranking of three brands of soft drink
- Example: Position within an organization
  - 1 for President
  - 2 for Vice President
  - 3 for Plant Manager
  - 4 for Department Supervisor
  - 5 for Employee

Ordinal Data

Faculty and staff should receive preferential treatment for parking space.



Interval Level Data

- Distances between consecutive integers are equal
  - Relative magnitude of numbers is meaningful
  - Differences between numbers are comparable
  - Location of origin, zero, is arbitrary
  - Vertical intercept of unit of measure transform function is not zero

- Example: Fahrenheit Temperature
- Example: Calendar Time
- Example: Monetary Utility

Interval Level Data

Interval level: similar to the ordinal level, with the additional property that meaningful amounts of differences between data values can be determined. There is no natural zero point.

Ratio Level Data

Highest level of measurement

- Relative magnitude of numbers is meaningful
- Differences between numbers are comparable
- Location of origin, zero, is absolute (natural)
- Vertical intercept of unit of measure transform function is zero

Examples: Height, Weight, and Volume

Example: Monetary Variables, such as Profit and Loss, Revenues, and Expenses

Example: Financial ratios, such as P/E Ratio, Inventory Turnover, and Quick Ratio.

Ratio Level Data

Ratio level: the interval level with an inherent zero starting point. Differences and ratios are meaningful for this level of measurement.

Data Level, Operation, & Statistical Method

Data Level	Meaningful Operations	Statistical Methods
Nominal	Classifying and Counting	Nonparametric
Ordinal	All of the above plus Ranking	Nonparametric
Interval	All of the above plus Addition, Subtraction, Multiplication, and Division	Parametric
Ratio	All of the above	Parametric

Section 2.2

Organizing Qualitative Data

Definition 2.3

Frequency Distribution of Qualitative Data

A **frequency distribution** of qualitative data is a listing of the distinct values and their frequencies.

Table 2.1

Students of a Statistical class were asked to state their political party affiliations as Democratic (D), Republican (R), or Other (O). The responses of the 40 students in the class are given in Table 2.1. Determine a frequency distribution of these data.

D	R	O	R	R	R	R	R
D	O	R	D	O	O	R	D
D	R	O	D	R	R	O	R
D	O	D	D	D	R	O	D
O	R	D	R	R	R	R	D

Political party affiliations of the students in the class

Table 2.2

Table for constructing a frequency distribution for the political party affiliation data in Table 2.1

Party	Tally	Frequency
Democratic	𝍸 𝍸 |||	13
Republican	𝍸 𝍸 𝍸 |||	18
Other	𝍸 ||||	9
		40

Definition 2.4

Relative-Frequency Distribution of Qualitative Data

A **relative-frequency distribution** of qualitative data is a listing of the distinct values and their relative frequencies.

In addition to the frequency that a particular distinct value occurs, we are often interested in the **relative frequency,** which is the ratio of the frequency to the total number of observations:

$$\text{Relative frequency} = \frac{\text{Frequency}}{\text{Number of observations}}.$$

For instance, as we see from Table 2.2, the relative frequency of Democrats in Professor Weiss's introductory statistics class is

$$\text{Relative frequency of Democrats} = \frac{\text{Frequency of Democrats}}{\text{Number of observations}} = \frac{13}{40} = 0.325.$$

Table 2.3

Relative-frequency distribution for the political party affiliation data in Table 2.1

Party	Relative frequency	
Democratic	0.325	← *13/40*
Republican	0.450	← *18/40*
Other	0.225	← *9/40*
	1.000	

Figure 2.2

Pie chart of the political party affiliation data in Table 2.1

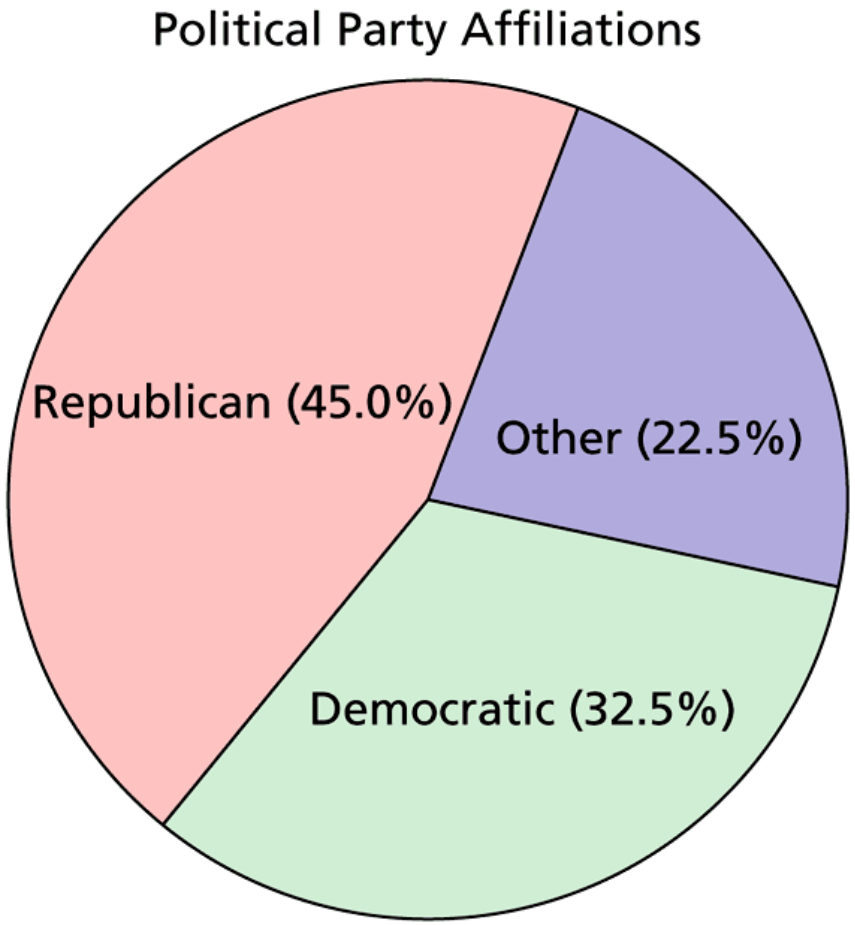
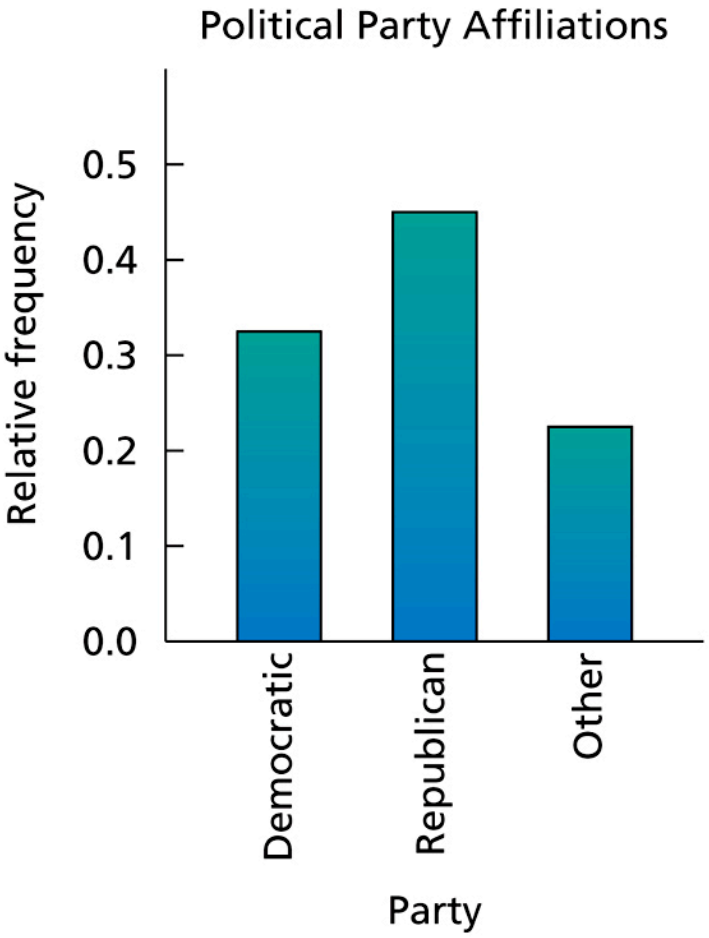


Figure 2.3

Bar chart of the political party affiliation data in Table 2.1



Section 2.3

Organizing Quantitative Data

Table 2.4

Number of TV sets in each of 50 randomly selected households.

1	1	1	2	6	3	3	4	2	4
3	2	1	5	2	1	3	6	2	2
3	1	1	4	3	2	2	2	2	3
0	3	1	2	1	2	3	1	1	3
3	2	1	2	1	1	3	1	5	1

Table 2.5

Frequency and relative-frequency distributions, using single-value grouping, for the number-of-TVs data in Table 2.4

Number of TVs	Frequency	Relative frequency
0	1	0.02
1	16	0.32
2	14	0.28
3	12	0.24
4	3	0.06
5	2	0.04
6	2	0.04
	50	1.00

Example: *Days to Maturity for Short-Term Investments*

Table 2.6 displays the number of days to maturity for 40 short-term investments. The data are from *BARRON'S* magazine. Use limit grouping, with grouping by 10s, to organize these data into frequency and relative-frequency distributions.

70	64	99	55	64	89	87	65
62	38	67	70	60	69	78	39
75	56	71	51	99	68	95	86
57	53	47	50	55	81	80	98
51	36	63	66	85	79	83	70

Table 2.7

Frequency and relative-frequency distributions, using limit grouping, for the days-to-maturity data in Table 2.6

Days to maturity	Tally	Frequency	Relative frequency
30–39	|||	3	0.075
40–49	|	1	0.025
50–59	𝍸 |||	8	0.200
60–69	𝍸 𝍸	10	0.250
70–79	𝍸 ||	7	0.175
80–89	𝍸 ||	7	0.175
90–99	||||	4	0.100
		40	1.000

Definition 2.7

Terms Used in Limit Grouping

Lower class limit: The smallest value that could go in a class.

Upper class limit: The largest value that could go in a class.

Class width: The difference between the lower limit of a class and the lower limit of the next-higher class.

Class mark: The average of the two class limits of a class.

Definition 2.8

Terms Used in Cutpoint Grouping

Lower class cutpoint: The smallest value that could go in a class.

Upper class cutpoint: The largest value that could go in the next-higher class (equivalent to the lower cutpoint of the next-higher class).

Class width: The difference between the cutpoints of a class.

Class midpoint: The average of the two cutpoints of a class.

Definition 2.9

Histogram

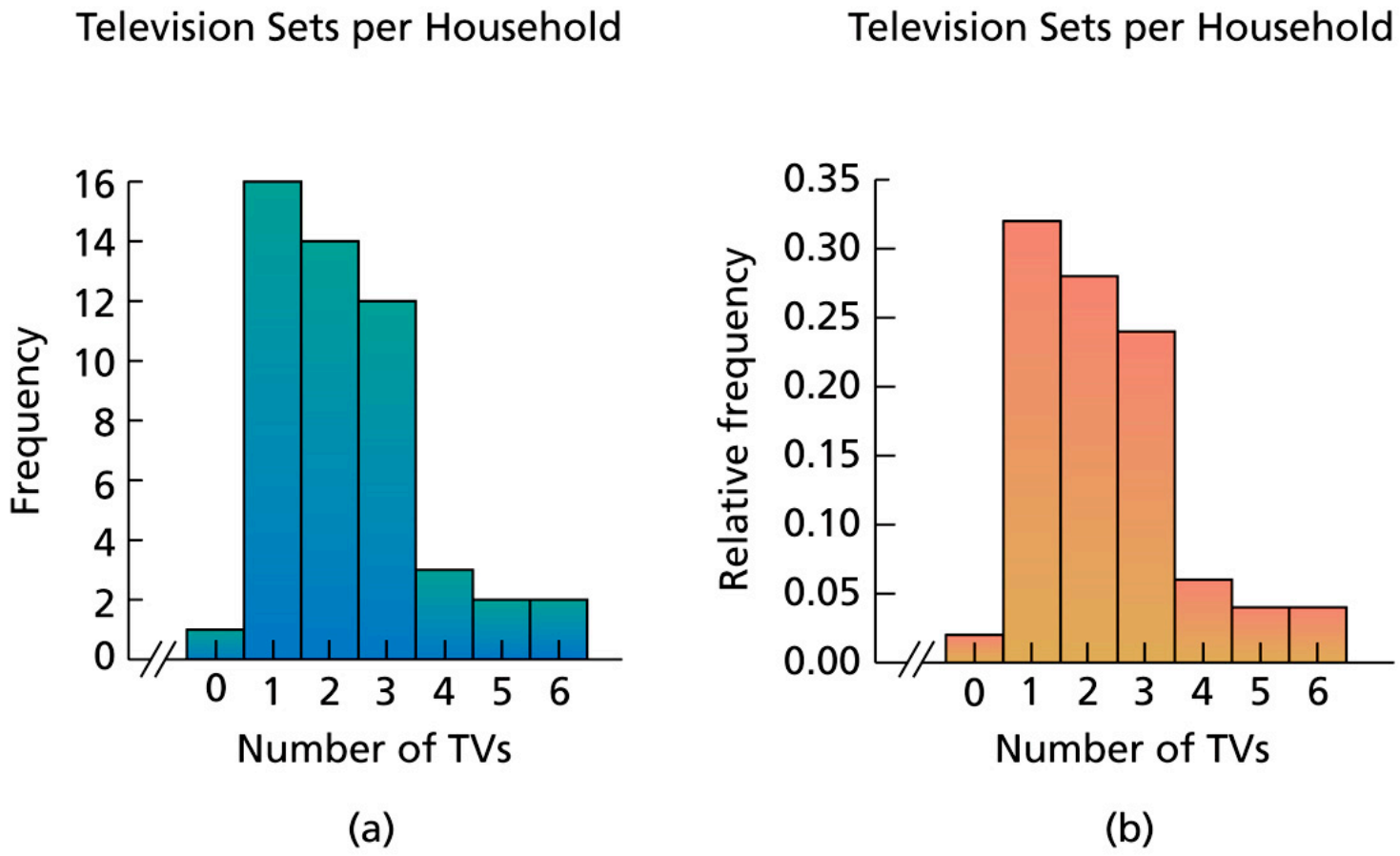
A **histogram** displays the classes of the quantitative data on a horizontal axis and the frequencies (relative frequencies, percents) of those classes on a vertical axis. The frequency (relative frequency, percent) of each class is represented by a vertical bar whose height is equal to the frequency (relative frequency, percent) of that class. The bars should be positioned so that they touch each other.

- For single-value grouping, we use the distinct values of the observations to label the bars, with each such value centered under its bar.
- For limit grouping or cutpoint grouping, we use the lower class limits (or, equivalently, lower class cutpoints) to label the bars.

Note: Some statisticians and technologies use class marks or class midpoints centered under the bars.

Figure 2.4

Single-value grouping. Number of TVs per household:
(a) frequency histogram; (b) relative-frequency histogram



Example: *Days to Maturity for Short-Term Investments*

Table 2.6 displays the number of days to maturity for 40 short-term investments. The data are from *BARRON'S* magazine. Use limit grouping, with grouping by 10s, to organize these data into frequency and relative-frequency distributions.

Max= 99
Min= 36

70	64	99	55	64	89	87	65
62	38	67	70	60	69	78	39
75	56	71	51	99	68	95	86
57	53	47	50	55	81	80	98
51	36	63	66	85	79	83	70

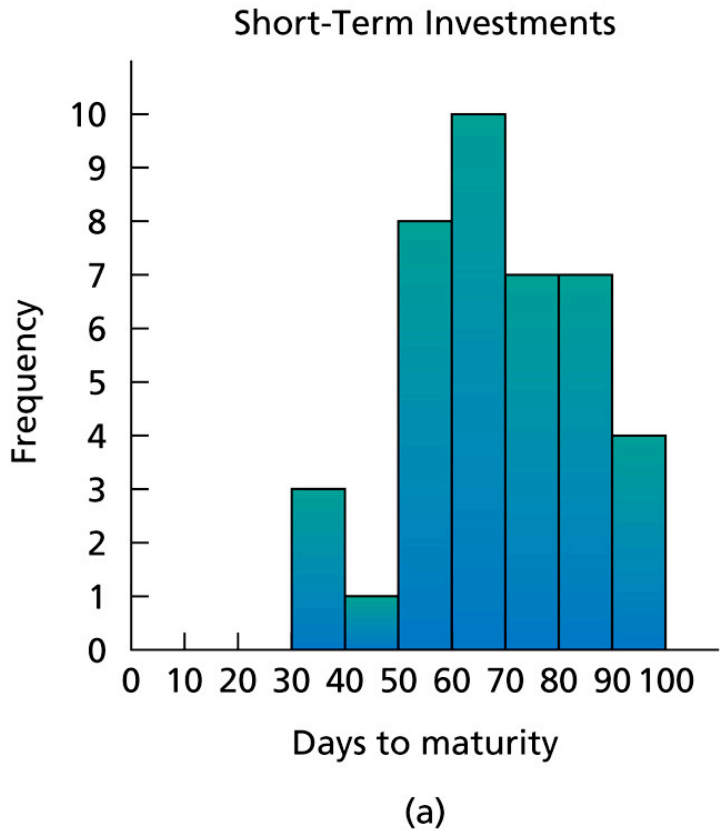
Example: *Days to Maturity for Short-Term Investments*

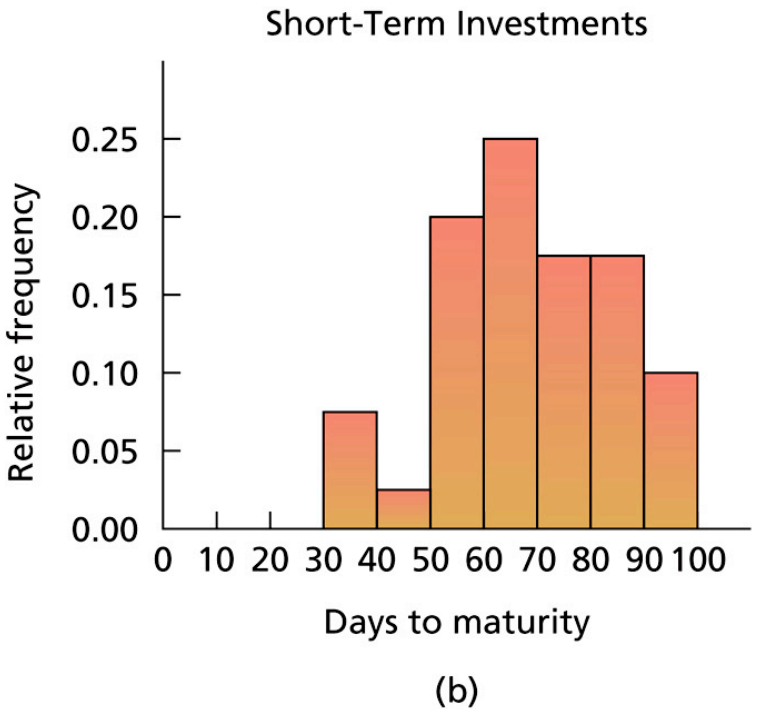
Frequency and relative-frequency distributions, using limit grouping, for the days-to-maturity data in Table 2.6

Days to maturity	Tally	Frequency	Relative frequency
30–39	|||	3	0.075
40–49	|	1	0.025
50–59	𝍸 |||	8	0.200
60–69	𝍸 𝍸	10	0.250
70–79	𝍸 ||	7	0.175
80–89	𝍸 ||	7	0.175
90–99	||||	4	0.100
		40	1.000

Figure 2.5

Limit grouping. Days to maturity: (a) frequency histogram; (b) relative-frequency histogram





Important guidelines for grouping

1. *The number of classes should be small enough to provide an effective summary but large enough to display the relevant characteristics of the data. p.s. A rule of thumb is that the number of classes should be between 5 and 20.*

2. *Each observation must belong to one, and only one, class.*

3. *Whenever feasible, all classes should have the same width.*

Example: Prices of DVD Players

One of Professor Weiss's sons wanted to add a new DVD player to his home theater system. He used the Internet to shop and went to pricewatch.com. There he found 16 quotes on different brands and styles of DVD players. Table 2.11 lists the prices, in dollars.

210	219	214	197
224	219	199	199
208	209	215	199
212	212	219	210

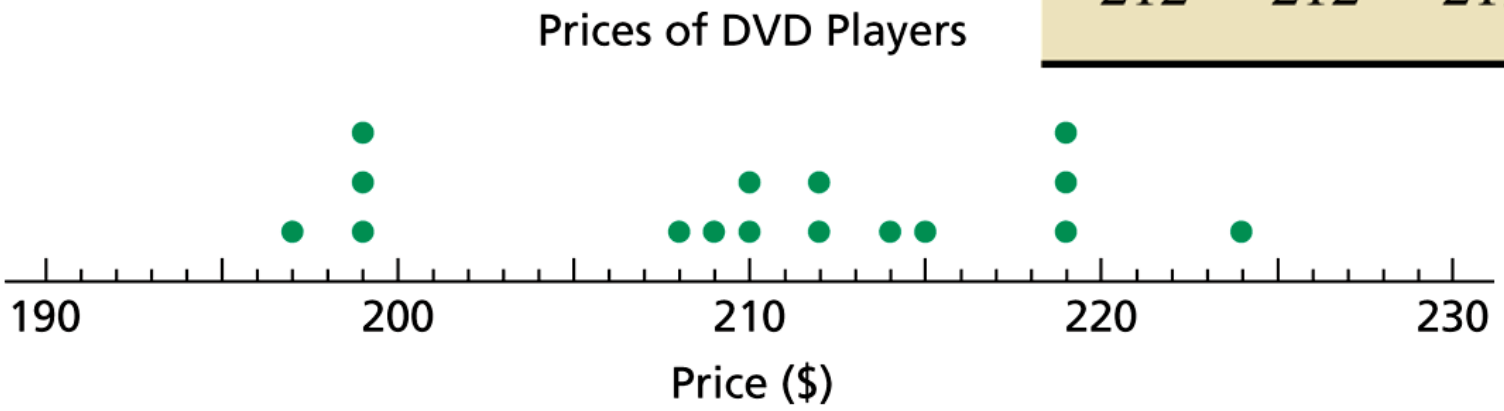


Table 2.11 & Figure 2.7

Prices, in dollars, of 16 DVD players

210	219	214	197
224	219	199	199
208	209	215	199
212	212	219	210

PROCEDURE 2.7
Construction of a Stem-and-Leaf Diagram

To Construct a Stem-and-Leaf Diagram

Step 1 Think of each observation as a stem—consisting of all but the rightmost digit—and a leaf, the rightmost digit.

Step 2 Write the stems from smallest to largest in a vertical column to the left of a vertical rule.

Step 3 Write each leaf to the right of the vertical rule in the row that contains the appropriate stem.

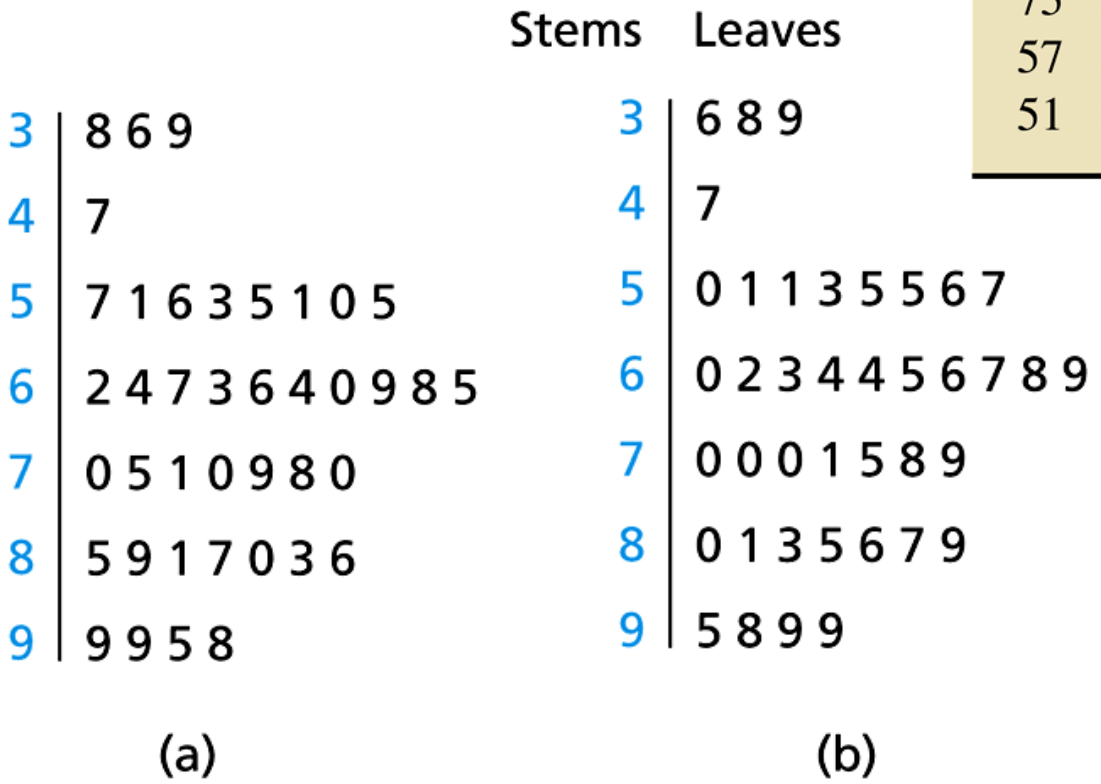
Step 4 Arrange the leaves in each row in ascending order.

Table 2.12 & Figure 2.8

Days to maturity for
40 short-term investments

70	64	99	55	64	89	87	65
62	38	67	70	60	69	78	39
75	56	71	51	99	68	95	86
57	53	47	50	55	81	80	98
51	36	63	66	85	79	83	70

Constructing a stem-and-leaf diagram
for the days-to-maturity data



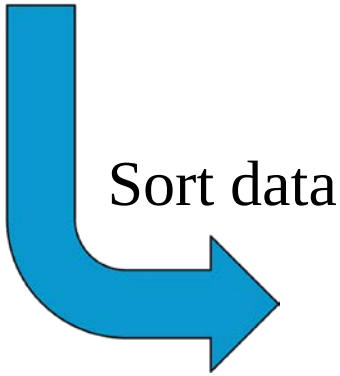
Example: Cholesterol Levels

According to the *National Health and Nutrition Examination Survey*, published by the Centers for Disease Control, the average cholesterol level for children between 4 and 19 years of age is 165 mg/dL. A pediatrician tested the cholesterol levels of several young patients and was alarmed to find that many had levels higher than 200 mg/dL. Table 2.13 presents the readings of 20 patients with high levels. Construct a stem-and-leaf diagram for these data by using:

a. one line per stem. **b.** two lines per stem.

Table 2.13

210	209	212	208
217	207	210	203
208	210	210	199
215	221	213	218
202	218	200	214



199
200
202
203
207
208
208
209
210
210
210
210
212
213
214
215
217
218
218
221

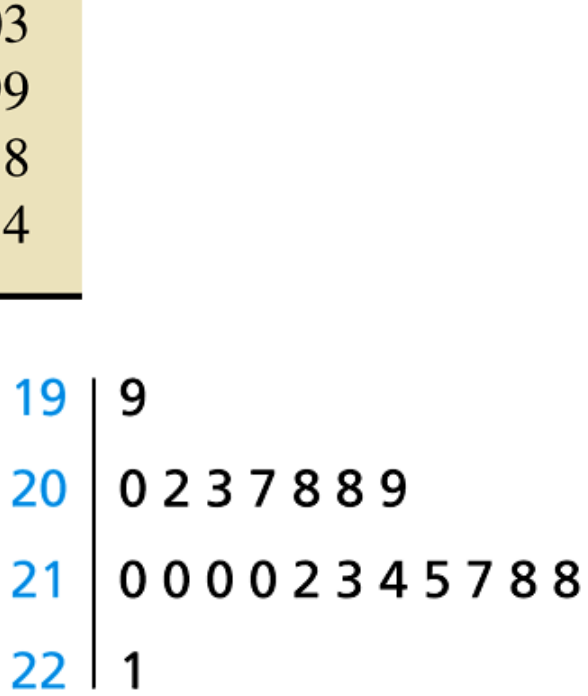
19	9
20	0 2 3 7 8 8 9
21	0 0 0 0 2 3 4 5 7 8 8
22	1

Table 2.13 & Figure 2.9

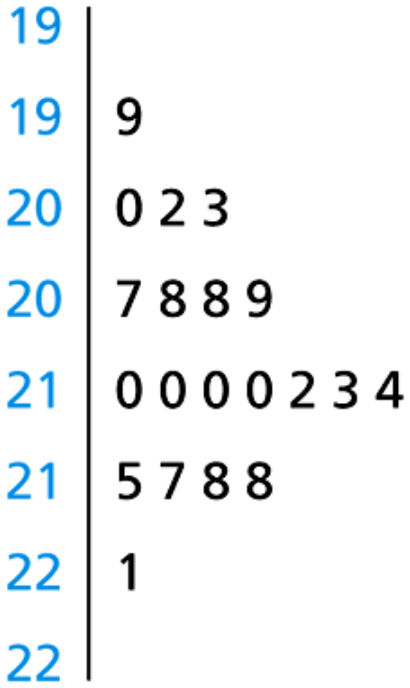
Cholesterol levels
for 20 high-level patients

210	209	212	208
217	207	210	203
208	210	210	199
215	221	213	218
202	218	200	214

Stem-and-leaf diagram for cholesterol levels: (a) one line per stem; (b) two lines per stem



(a)



(b)

Section 2.4

Distribution Shapes

Definition 2.10

Distribution of a Data Set

The **distribution of a data set** is a table, graph, or formula that provides the values of the observations and how often they occur.

Figure 2.10

Relative-frequency histogram and approximating smooth curve for the distribution of heights

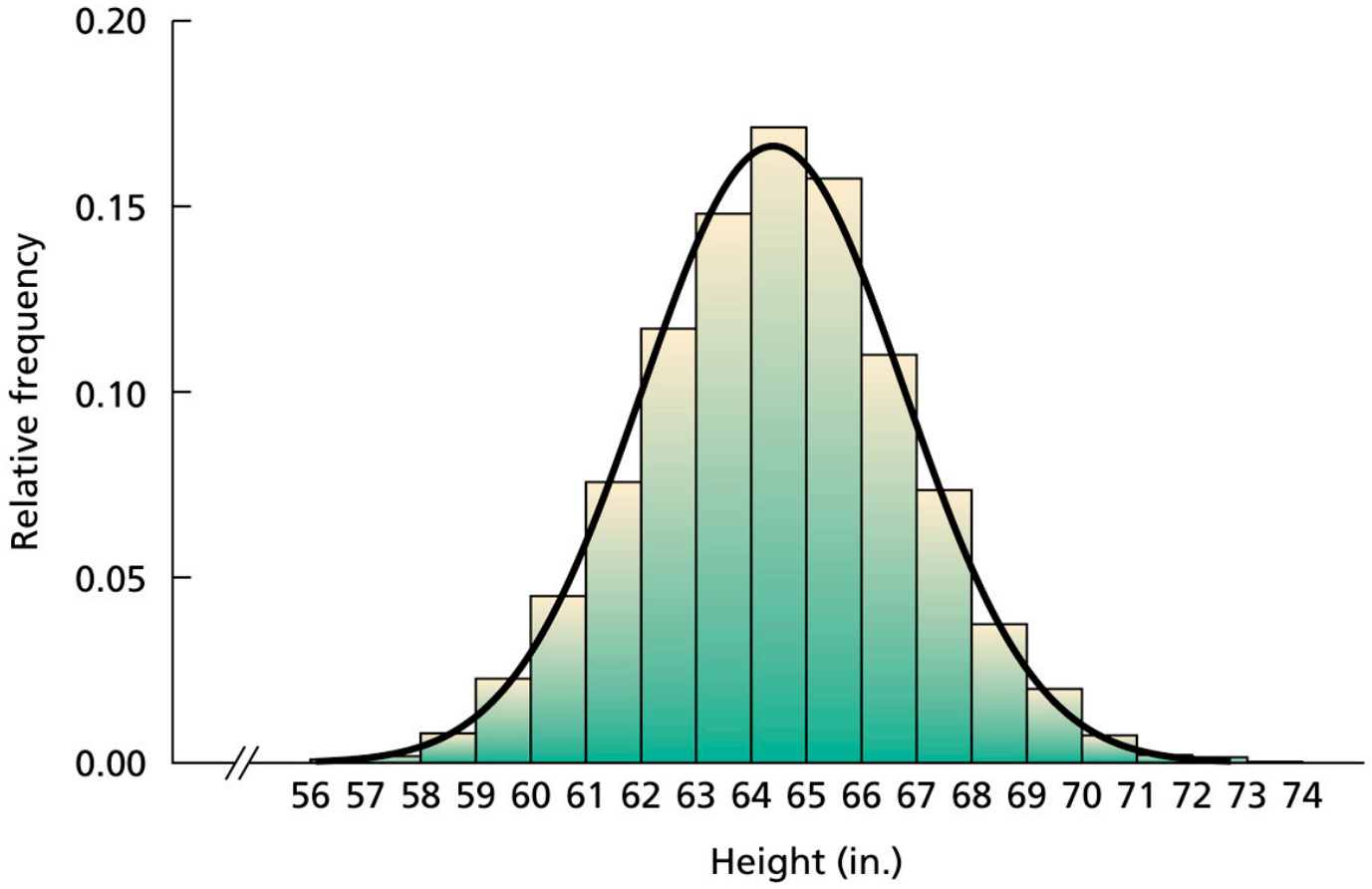


Figure 2.11

Common distribution shapes

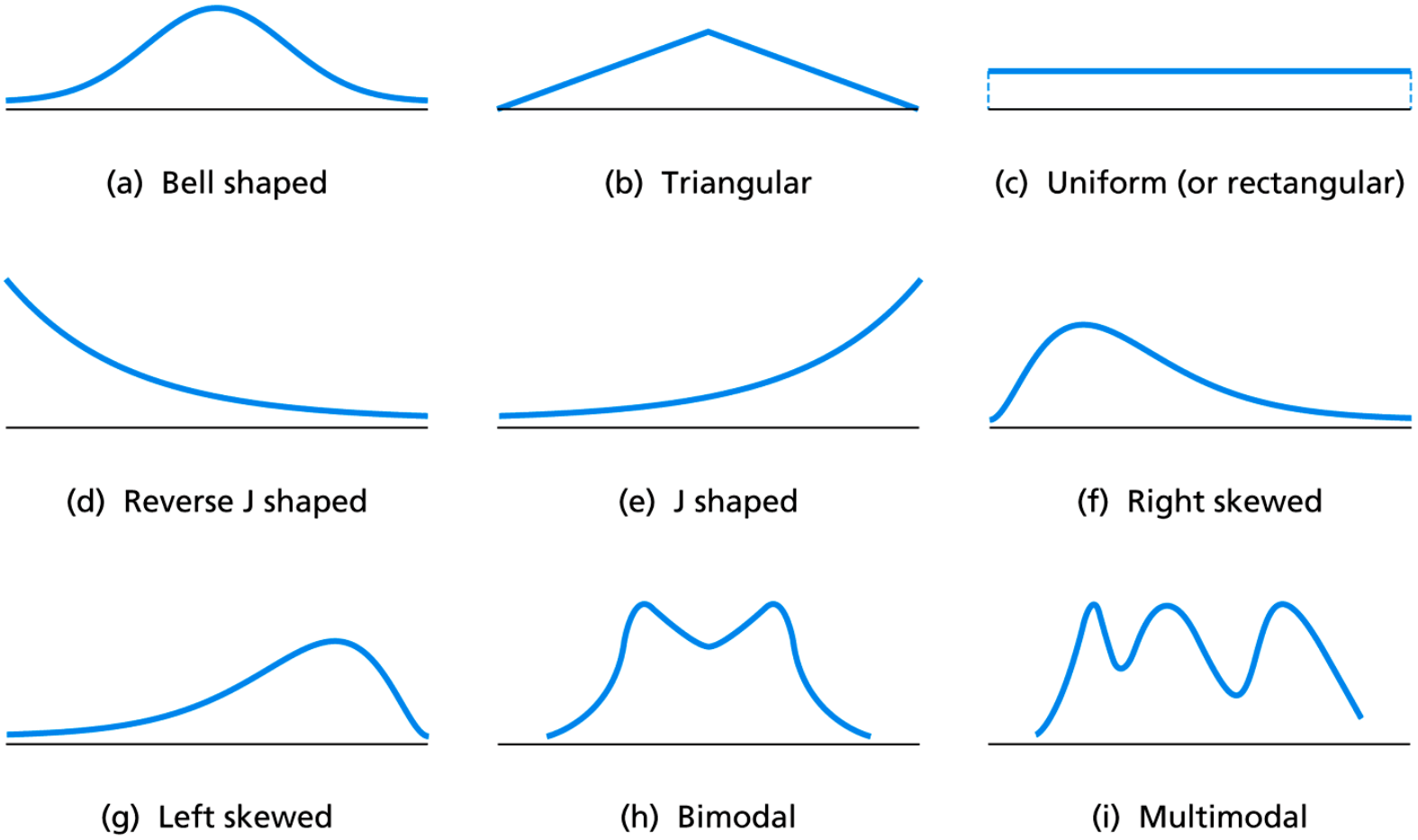
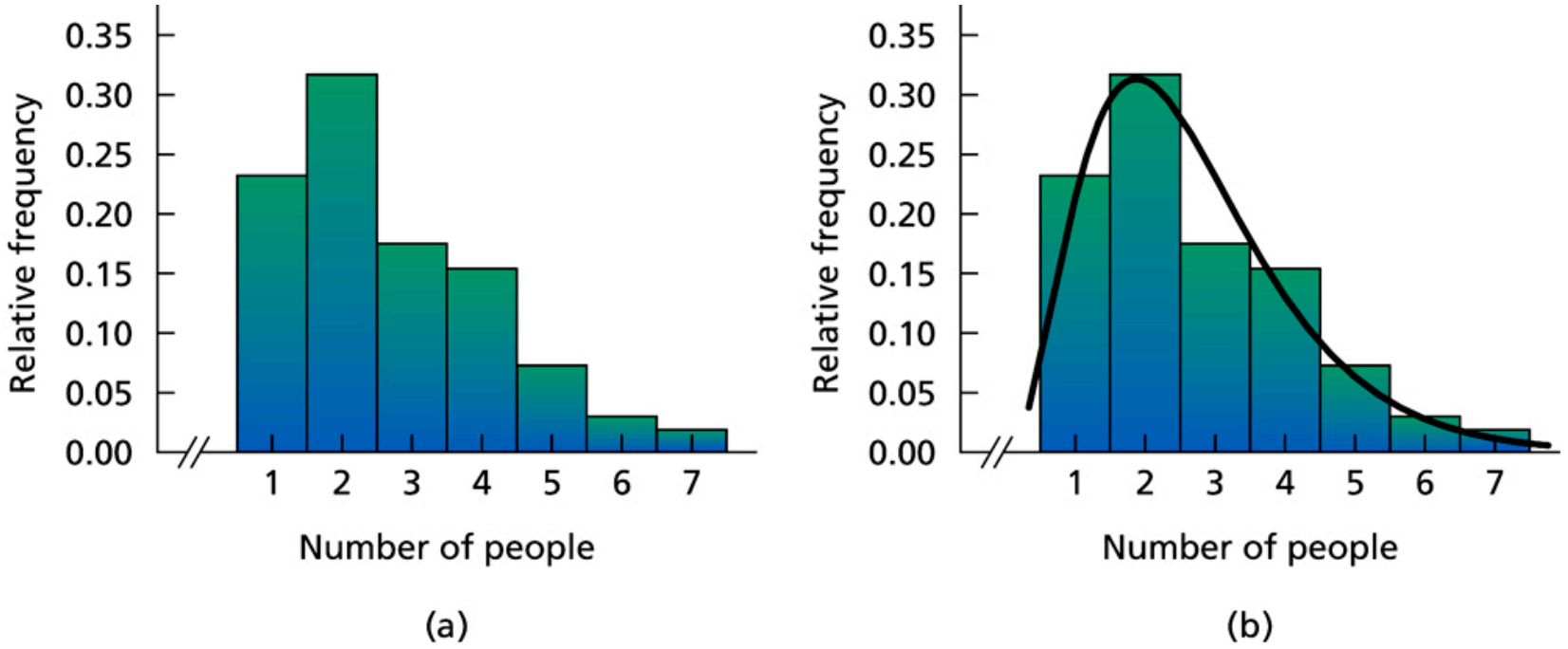


Figure 2.12

Relative-frequency histogram for household size



Definition 2.12

Population and Sample Distributions; Distribution of a Variable

The distribution of population data is called the **population distribution,** or the **distribution of the variable.**

The distribution of sample data is called a **sample distribution.**

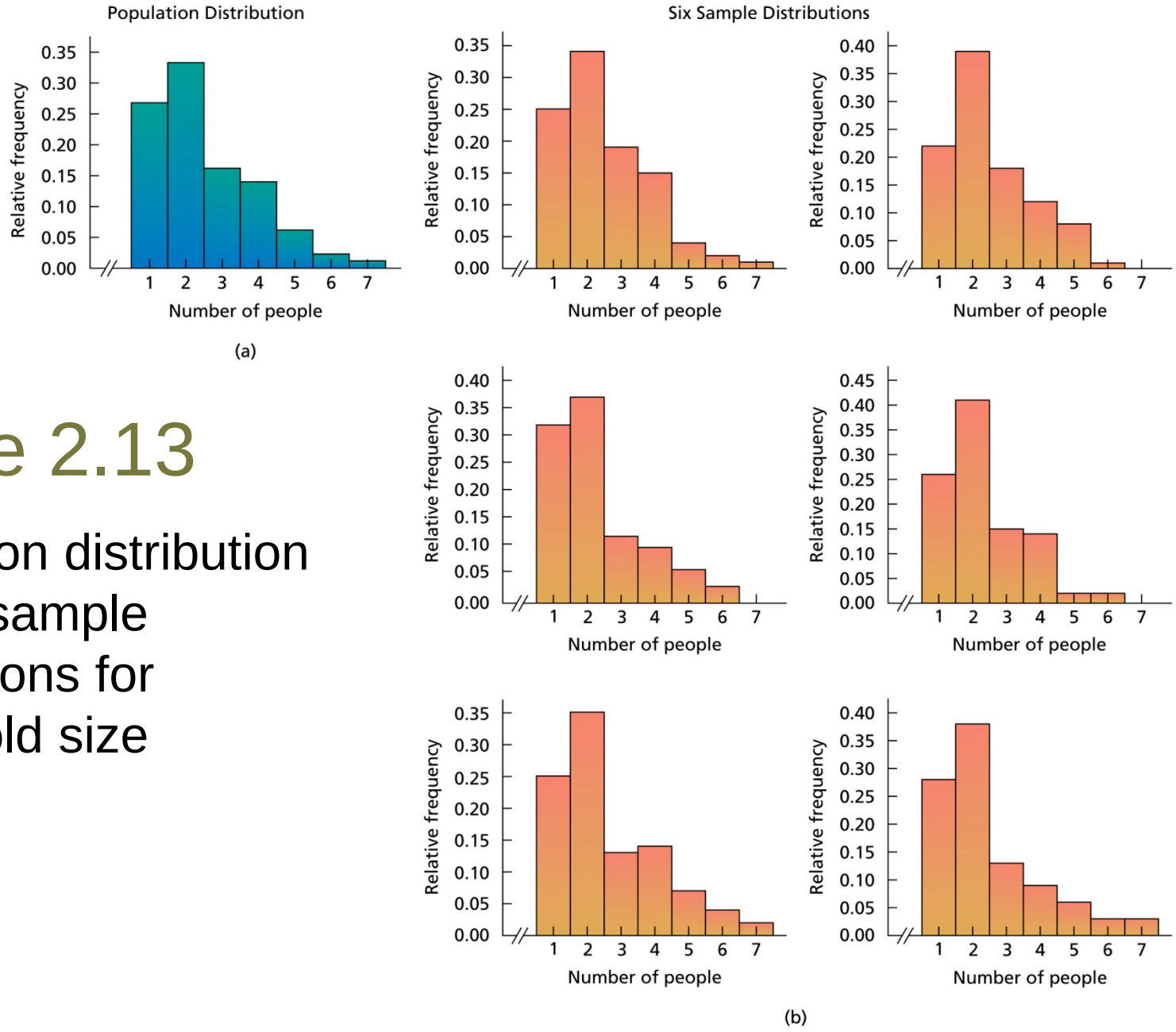


Figure 2.13

Population distribution and six sample distributions for household size

Key Fact 2.1

Population and Sample Distributions

For a simple random sample, the sample distribution approximates the population distribution (i.e., the distribution of the variable under consideration). The larger the sample size, the better the approximation tends to be.

Section 2.5

Misleading Graphs

Figure 2.17

Unemployment rates: (a) truncated graph; (b) nontruncated graph

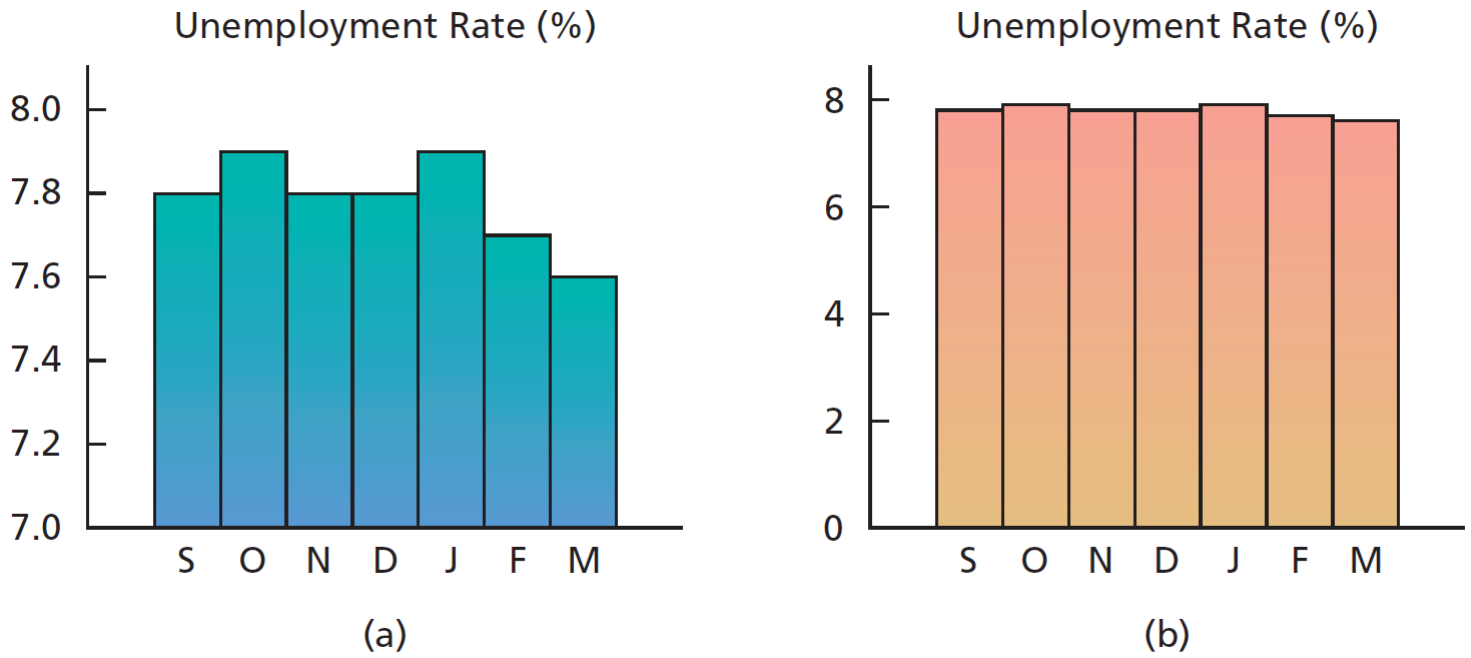


Figure 2.19

Improper scaling: Number of homes this year will be double last year, so the developer doubled the width and height, which makes it look like four times the number of homes will be built.