CHAPTER 3

Descriptive Statistics: Numerical Methods

Outline

In this chapter we study numerical methods for describing the important aspects of a set of measurements. If the measurements are values of a quantitative variable, we often describe
(1) what a typical measurement might be and
(2) how the measurements vary, or differ, from each other.

For example, in the car mileage case we might estimate
(1)a typical EPA gas mileage for the new midsize model
(2) how the EPA mileages vary from car to car. Or, in the marketing research case,

Taken together, the graphical displays of Chapter 2 and the numerical methods of this chapter give us a basic understanding of the important aspects of a set of measurements.

Chapter Outline

3.1 Describing Central Tendency

3.2 Measures of Variation

3.3 Percentiles, Quartiles and Box-and-Whiskers Displays

3.4 Covariance, Correlation, and the Least Square Line (Optional)

3.5 Weighted Means and Grouped Data (Optional)

3.6 The Geometric Mean (Optional)

Chapter Outline

3.1 Describing Central Tendency

3.1 Describing Central Tendency

- In addition to describing the shape of a distribution, want to describe the data set's central tendency
- A measure of central tendency represents the center or middle of the data
- May or may not be a typical value

Parameters and Statistics

- *A **population parameter*** is a number calculated using the population measurements that describes some aspect of the population
- A ***sample statistic*** is a number calculated using the sample measurements that describes some aspect of the sample

Point Estimates and Sample Statistics

A *point estimate* is a one-number estimate of the value of a population parameter

A *sample statistic* is a number calculated using sample measurements that describes some aspect of the sample

- Use sample statistics as point estimates of the population parameters

The sample mean, denoted $\bar{x}$, is a sample statistic and is the average of the sample measurements

- The sample mean is a point estimate of the population mean

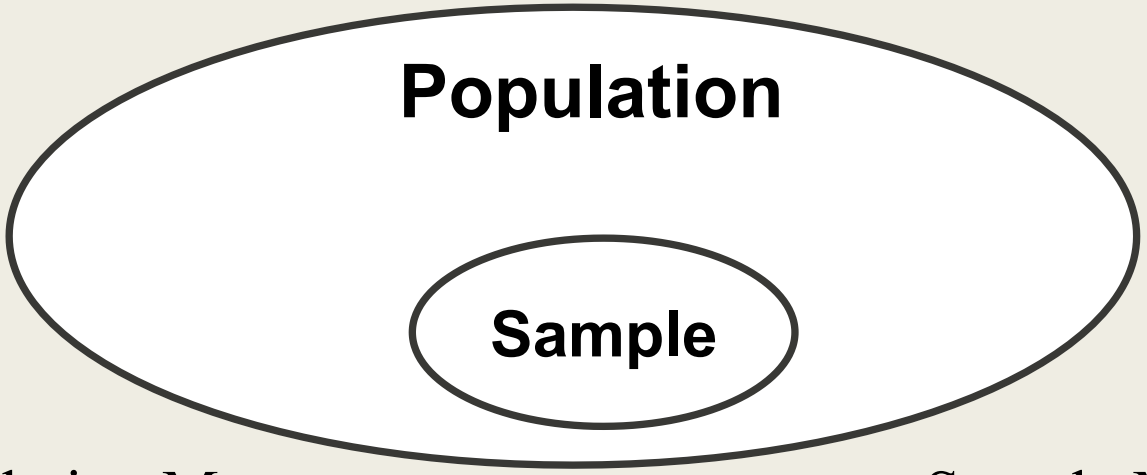
Measures of Central Tendency

Mean, μ	The average or expected value
Median, M_d	The value of the middle point of the ordered measurements
Mode, M_o	The most frequent value

The Mean

Population $X_1, X_2, \ldots, X_N$

Sample $x_1, x_2, \ldots, x_n$



Population Mean

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

Sample Mean

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Example: Car Mileage Case

Example 3.1: **Sample mean** for first five car mileages from Table 3.1

30.8, 31.7, 30.1, 31.6, 32.1

$$\bar{x} = \frac{\sum_{i=1}^{5} x_i}{5} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5}$$

$$\bar{x} = \frac{30.8 + 31.7 + 30.1 + 31.6 + 32.1}{5} = \frac{156.3}{5} = 31.26$$

the government will offer its tax credit to any automaker selling a midsize model equipped with an automatic transmission that achieves a mean EPA combined mileage of at least 31 mpg

Example: Car Mileage Case

Example 3.1: **Population mean** for all 50 car mileages from Table 3.1

TABLE 3.1 A Sample of 50 Mileages GasMiles

30.8	30.8	32.1	32.3	32.7
31.7	30.4	31.4	32.7	31.4
30.1	32.5	30.8	31.2	31.8
31.6	30.3	32.8	30.7	31.9
32.1	31.3	31.9	31.7	33.0
33.3	32.1	31.4	31.4	31.5
31.3	32.5	32.4	32.2	31.6
31.0	31.8	31.0	31.5	30.6
32.0	30.5	29.8	31.7	32.3
32.4	30.5	31.1	30.7	31.4

$$\sum_{i=1}^{50} x_i = x_1 + x_2 + \cdots + x_{50} = 30.8 + 31.7 + \cdots + 31.4 = 1578$$

$$\bar{x} = \frac{\sum_{i=1}^{50} x_i}{50} = \frac{1578}{50} = 31.56$$

Usually speaking, there is a difference between the sample mean (31.26) and the population mean (31.56), known as the sampling error.

The Median

- The median M_d is a value such that 50% of all measurements, after having been arranged in numerical order, lie above (or below) it
- The median divides a population or sample into two roughly equal parts.
  1. If the number of measurements is odd, the median is the middlemost measurement in the ordering
  2. If the number of measurements is even, the median is the average of the two middlemost measurements in the ordering

Example: Median

- **Example 1:** Car Mileage Case: First five observations from Table 3.1:

  30.8, 31.7, 30.1, 31.6, 32.1

- In order: 30.1, 30.8, 31.6, 31.7, 32.1
- There is an odd so median is one in middle, or 31.6

- **Example 2**: Chris's five classes have sizes 60, 41, 15, 30, and 34. Arrange them in increasing order

  15, 30, 30, 34, 41, 60

- Median is the average of the two in the middle or (30+34)/2=32

The Mode

- The mode M_o of a population or sample of measurements is the measurement that occurs most frequently
- Modes are the values that are observed “most typically”
- Sometimes higher frequencies at two or more values
  - *If there are two modes, the data is bimodal*
  - *If more than two modes, the data is multimodal*
- When data are in classes, the class with the highest frequency is the modal class

The Mode

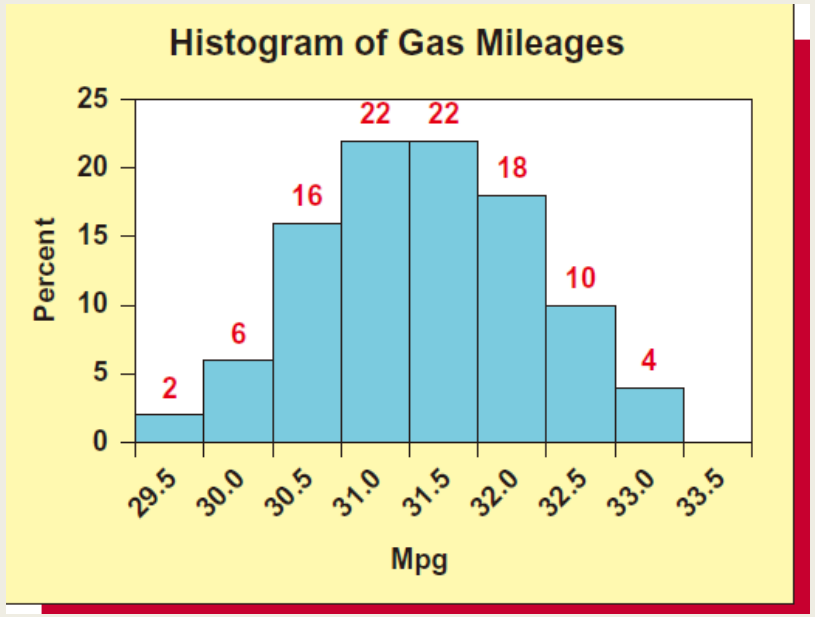
TABLE 3.1 A Sample of 50 Mileages GasMiles

30.8	30.8	32.1	32.3	32.7
31.7	30.4	31.4	32.7	31.4
30.1	32.5	30.8	31.2	31.8
31.6	30.3	32.8	30.7	31.9
32.1	31.3	31.9	31.7	33.0
33.3	32.1	31.4	31.4	31.5
31.3	32.5	32.4	32.2	31.6
31.0	31.8	31.0	31.5	30.6
32.0	30.5	29.8	31.7	32.3
32.4	30.5	31.1	30.7	31.4

The mode is 31.4, which occurs five times.

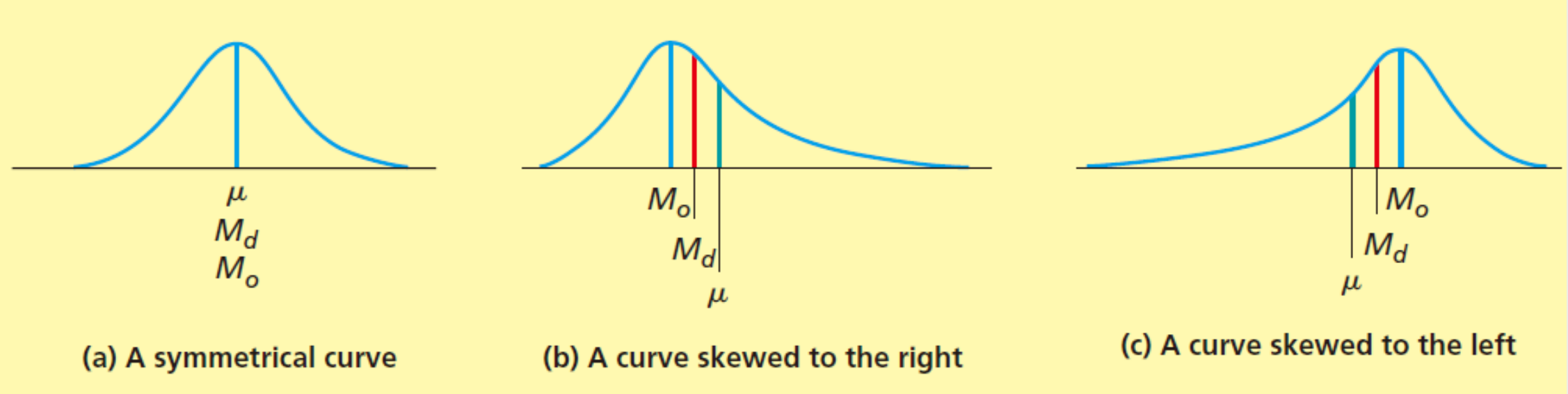
FIGURE 3.1 Excel Output of Statistics Describing the 50 Mileages

	Mileage
Mean	31.56
Standard Error	0.1128
Median	31.55
Mode	31.4



Relationships Among Mean, Median and Mode

Figure 3.3



Mean (μ), median (M_d), and mode (M_o) are all equal

$M_o < M_d < \mu$

$\mu < M_d < M_o$

Examples

EXAMPLE 3.2 Household Incomes

An economist wishes to study the distribution of household incomes in a Midwestern city. To do this, the economist randomly selects a sample of $n = 12$ households from the city and determines last year's income for each household.[1] The resulting sample of 12 household incomes—arranged in increasing order—is as follows (the incomes are expressed in dollars):

7,524	11,070	18,211	26,817	36,551	41,286
49,312	57,283	72,814	90,416	135,540	190,250

Number of incomes is even,

The median is (41286+49312)/2=45299

The sum of the incomes is 737074.
The mean is 737074/12 = 61423 (rounded)

The median is said to be resistant to these large incomes. The median is resistant to extreme values. Therefore, median is often used for salaries.

Examples

EXAMPLE 3.3 The Marketing Research Case: Rating A Bottle Design

BI

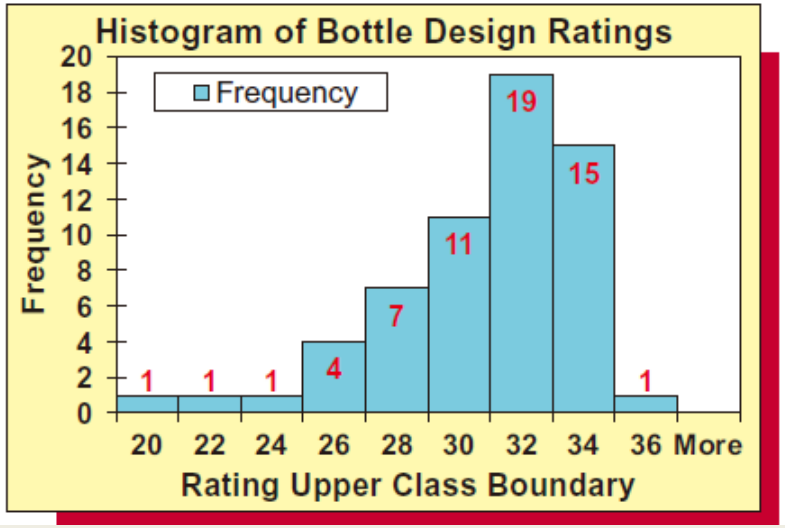
C

The Excel output in Figure 3.4 tells us that the mean and the median of the sample of 60 bottle design ratings are 30.35 and 31, respectively. Because the histogram of the bottle design ratings in Figure 3.5 is not highly skewed to the left, the sample mean is not much less than the sample median. Therefore, using the mean as our measure of central tendency, we estimate that the mean rating of the new bottle design that would be given by all consumers is 30.35. This is considerably higher than the minimum standard of 25 for a successful bottle design.

FIGURE 3.4 Excel Output of Statistics Describing the 60 Bottle Design Ratings

STATISTICS	
Mean	30.35
Standard Error	0.401146
Median	31
Mode	32
Standard Deviation	3.107263
Sample Variance	9.655085
Kurtosis	1.423397
Skewness	−1.17688
Range	15
Minimum	20
Maximum	35
Sum	1821
Count	60

FIGURE 3.5 Excel Frequency Histogram of the 60 Bottle Design Ratings



Payment Time Case

EXAMPLE 3.4 The e-billing Case: Reducing Bill Payment Times

BI

C

The MINITAB output in Figure 3.6 gives a histogram of the 65 payment times, and the MINITAB output in Figure 3.7 tells us that the mean and the median of the payment times are 18.108 days and 17 days, respectively. Because the histogram is not highly skewed to the right, the sample mean is not much greater than the sample median. Therefore, using the mean as our measure of central tendency, we estimate that the mean payment time of all bills using the new billing system is 18.108 days. This is substantially less than the typical payment time of 39 days that had been experienced using the old billing system.

- Mean=18.108 days
- Median=17.000 days
- Mode=16.000 days
- Expect the mean payment time to be 18.108 days
- A long payment time would be > 17 days and a short payment time would be < 17 days
- The typical payment time is 16 days

Payment Time Case

FIGURE 3.6 MINITAB Frequency Histogram of the 65 Payment Times

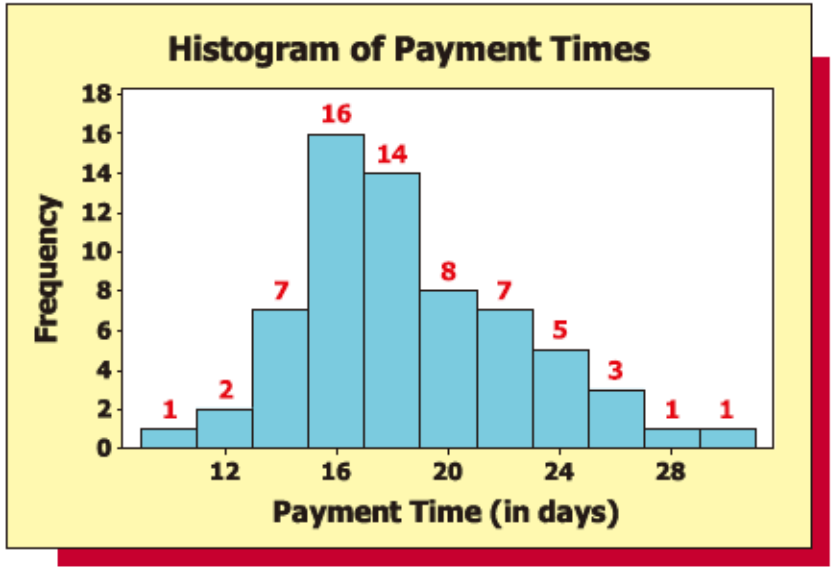


FIGURE 3.7 MINITAB Output of Statistics Describing the 65 Payment Times

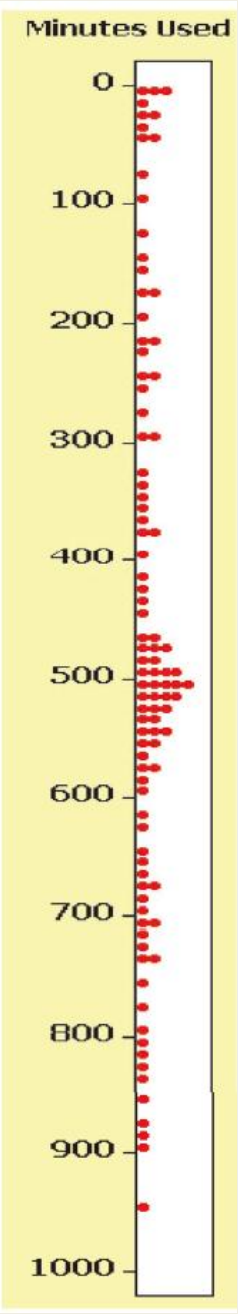
Variable	Count	Mean	StDev	Variance
PayTime	65	18.108	3.961	15.691

Variable	Minimum	Q1	Median	Q3	Maximum	Range
PayTime	10.000	15.000	17.000	21.000	29.000	19.000

Example 3.5

EXAMPLE 3.5 The Cell Phone Case: Reducing Cellular Phone Costs

Suppose that a cellular management service tells the bank that if its cellular cost per minute for the random sample of 100 bank employees is over 18 cents per minute, the bank will benefit from automated cellular management of its calling plans. Last month's cellular usages for the 100 randomly selected employees are given in Table 1.4 (page 9), and a dot plot of these usages is given in the page margin. If we add together the usages, we find that the 100 employees used a total of 46,625 minutes. Furthermore, the total cellular cost incurred by the 100 employees is found to be \$9,317 (this total includes base costs, overage costs, long distance, and roaming). This works out to an average of \$9,317/46,625 = \$.1998, or 19.98 cents per minute. Because this average cellular cost per minute exceeds 18 cents per minute, the bank will hire the cellular management service to manage its calling plans.

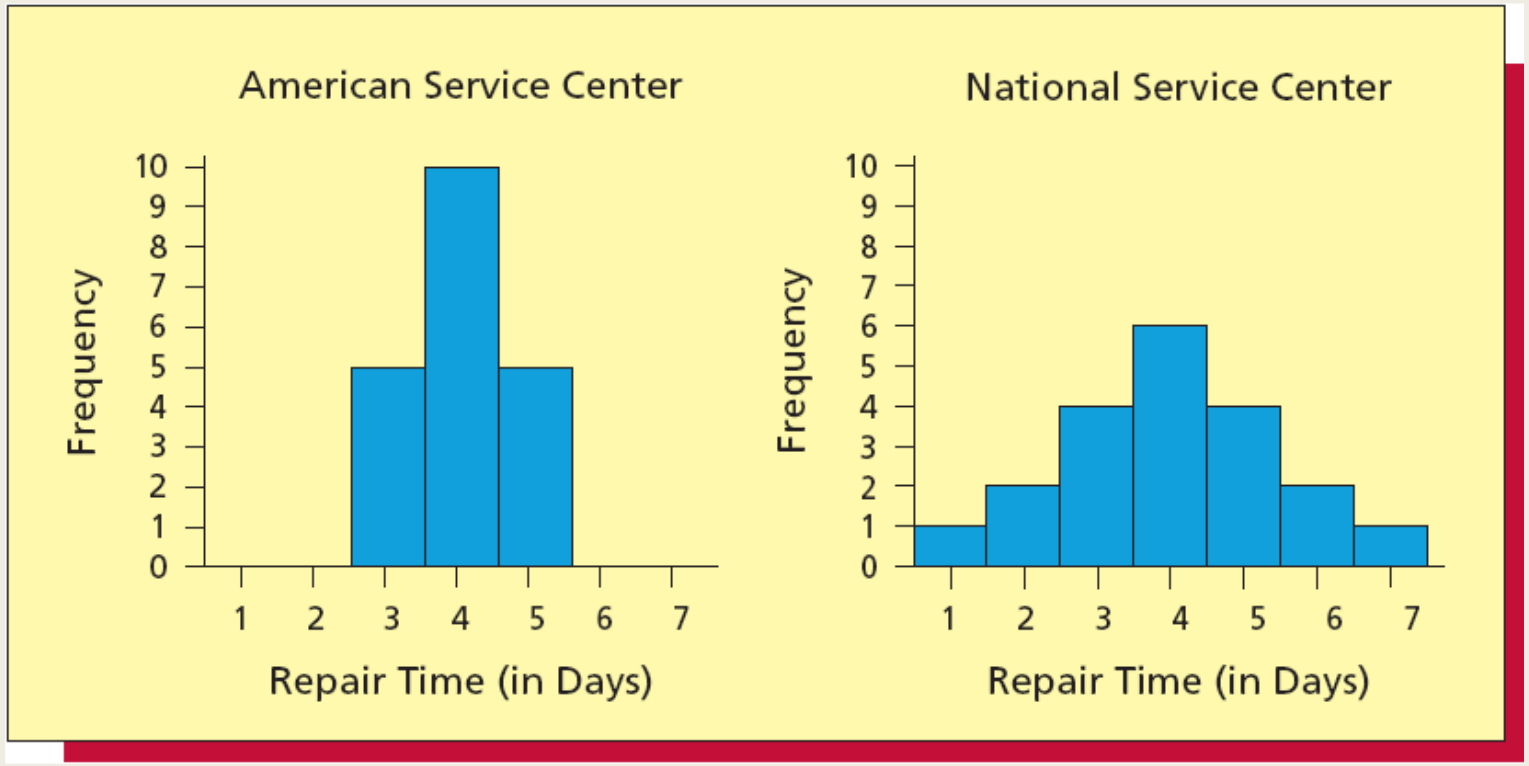


Chapter Outline

3.2 Measures of Variation

3.2 Measures of Variation

In addition to estimating a population's central tendency (which does not reflect everything), it is important to estimate the variation of the population's individual values.



The measures of central tendency do not indicate any difference between the American and National Service Centers.
But the two distributions significantly differ.

Measures of Variation

Range Largest minus the smallest measurement

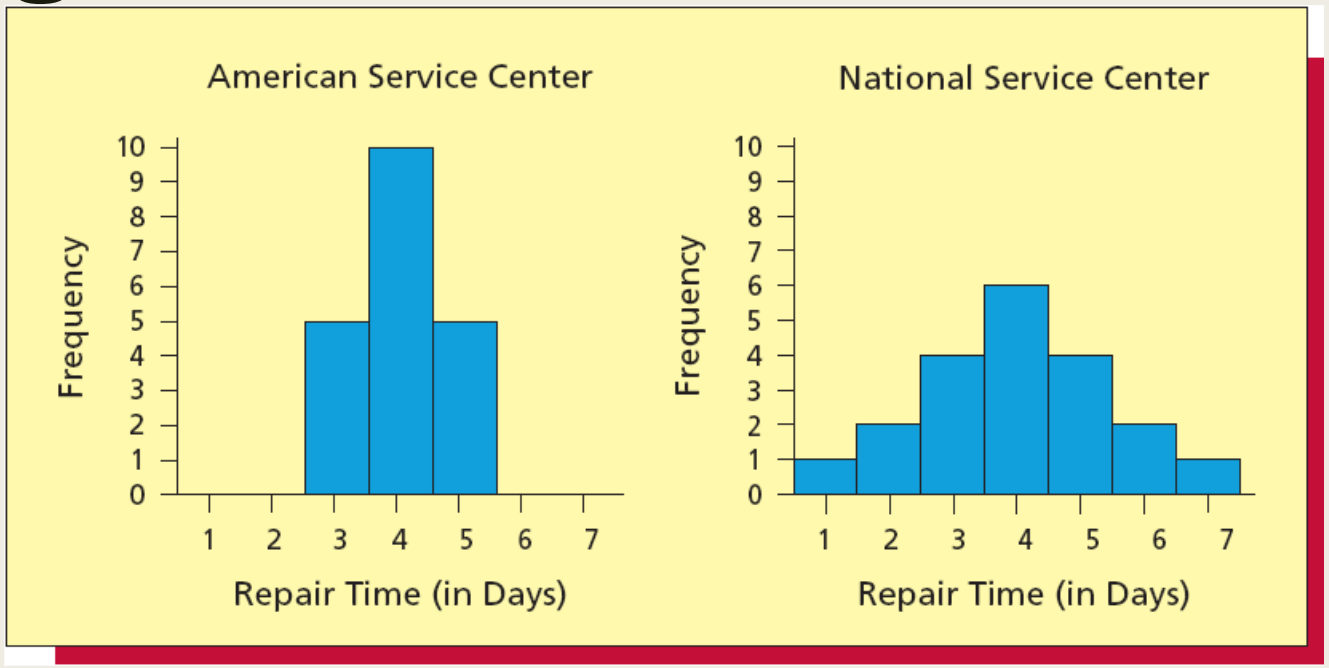
Variance (σ^2, pronounced sigma squared)

The average of the squared deviations of all the population measurements from the population mean μ

Standard Deviation (σ, pronounced sigma)

The positive square root of the variance

The Range



- Largest minus smallest
- Measures the interval spanned by all the data
- For American Service Center, largest is 5 and smallest is 3
  - *Range is 5 – 3 = 2 days*
- For National Service Center, range is 6

Variance

Population variance

Population Of Size N

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_N - \mu)^2}{N}$$

Sample variance

Sample Of Size n

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}$$

Standard Deviation

Population Standard Deviation

$$\sigma = \sqrt{\sigma^2}$$

Sample Standard Deviation

$$s = \sqrt{s^2}$$

Standard deviation should be positive values

Example: The Car Mileage Case

Example 1: Car Mileage Case: First five observations from Table 3.1:

30.8, 31.7, 30.1, 31.6, 32.1

$$\bar{x} = \frac{\sum_{i=1}^{5} x_i}{5} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5}$$

$$\bar{x} = \frac{30.8 + 31.7 + 30.1 + 31.6 + 32.1}{5} = \frac{156.3}{5} = 31.26$$

$$s^2 = \frac{\sum_{i=1}^{5} (x_i - \bar{x})^2}{5 - 1}$$

$$= \frac{(30.8 - 31.26)^2 + (31.7 - 31.26)^2 + (30.1 - 31.26)^2 + (31.6 - 31.26)^2 + (32.1 - 31.26)^2}{4}$$

$$= \frac{2.572}{4} = 0.643$$

$$s = \sqrt{s^2} = \sqrt{0.643} = 0.8019$$

Sample variance

Method 1

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}$$

Method 2

The **sample variance** can be calculated using the *computational formula*

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right]$$

Example: The Payment Time Case

TABLE 2.4 A Sample of Payment Times (in Days) for 65 Randomly Selected Invoices DS PayTime

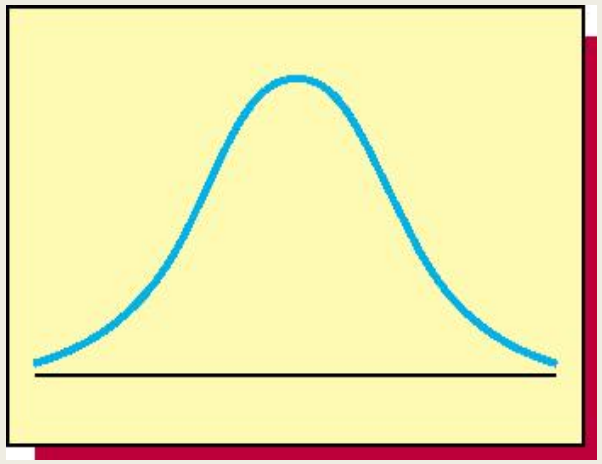
22	29	16	15	18	17	12	13	17	16	15
19	17	10	21	15	14	17	18	12	20	14
16	15	16	20	22	14	25	19	23	15	19
18	23	22	16	16	19	13	18	24	24	26
13	18	17	15	24	15	17	14	18	17	21
16	21	25	19	20	27	16	17	16	21	

$$\sum_{i=1}^{65} x_i = x_1 + x_2 + \cdots + x_{65} = 22 + 19 + \cdots + 21 = 1{,}177$$

$$\sum_{i=1}^{65} x_i^2 = x_1^2 + x_2^2 + \cdots + x_{65}^2 = (22)^2 + (19)^2 + \cdots + (21)^2 = 22{,}317$$

- Sample variance $$s^2 = \frac{1}{(65-1)}\left[22{,}317 - \frac{(1{,}177)^2}{65}\right] = \frac{1{,}004.2464}{64} = 15.69135$$
- Sample standard deviation $$s = \sqrt{s^2} = \sqrt{15.69135} = 3.9612$$

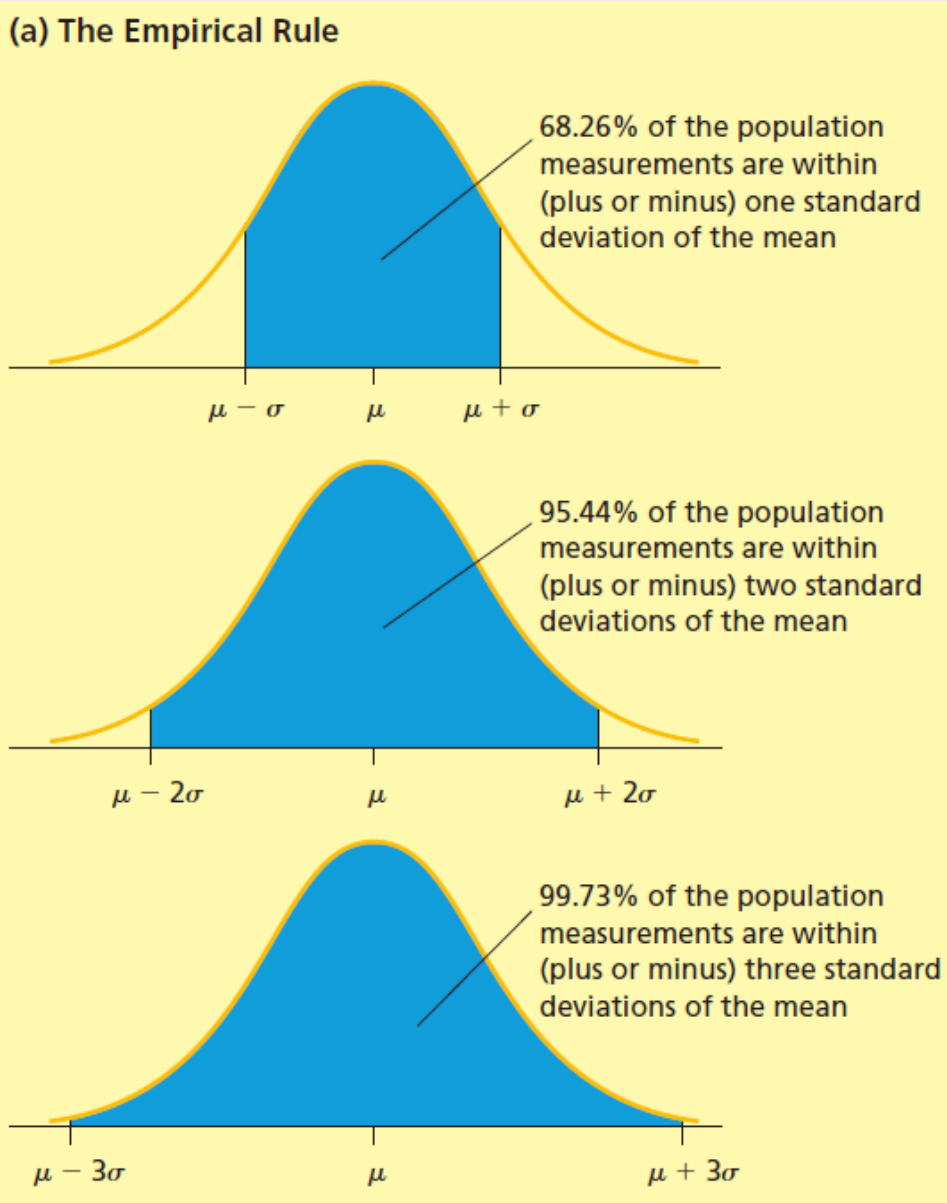
The Normal Curve



One type of relative frequency curve describing a population is the normal curve,

- Symmetrical and bell-shaped curve for a normally distributed population
- The height of the normal over any point represents the relative proportion of values near that point
- If a population is described by a normal curve, we say that the population is normally distributed

The Empirical Rule for Normal Populations



An interval that contains a specified percentage of the individual measurements in a population is called a tolerance interval

1. **68.26%** of the population measurements lie within one standard deviation of the mean: $[\mu-\sigma, \mu+\sigma]$
2. **95.44%** of the population measurements lie within two standard deviations of the mean: $[\mu-2\sigma, \mu+2\sigma]$
3. **99.73%** of the population measurements lie within three standard deviations of the mean: $[\mu-3\sigma, \mu+3\sigma]$

Three-sigma interval is a tolerance interval that contains almost all of the measurements in a normally distributed population

The Empirical Rule (经验准则) for Normal Populations

If a population has mean μ and standard deviation σ and is described by a normal curve, then

1. **68.26%** of the population measurements lie within one standard deviation of the mean: $[\mu-\sigma, \mu+\sigma]$

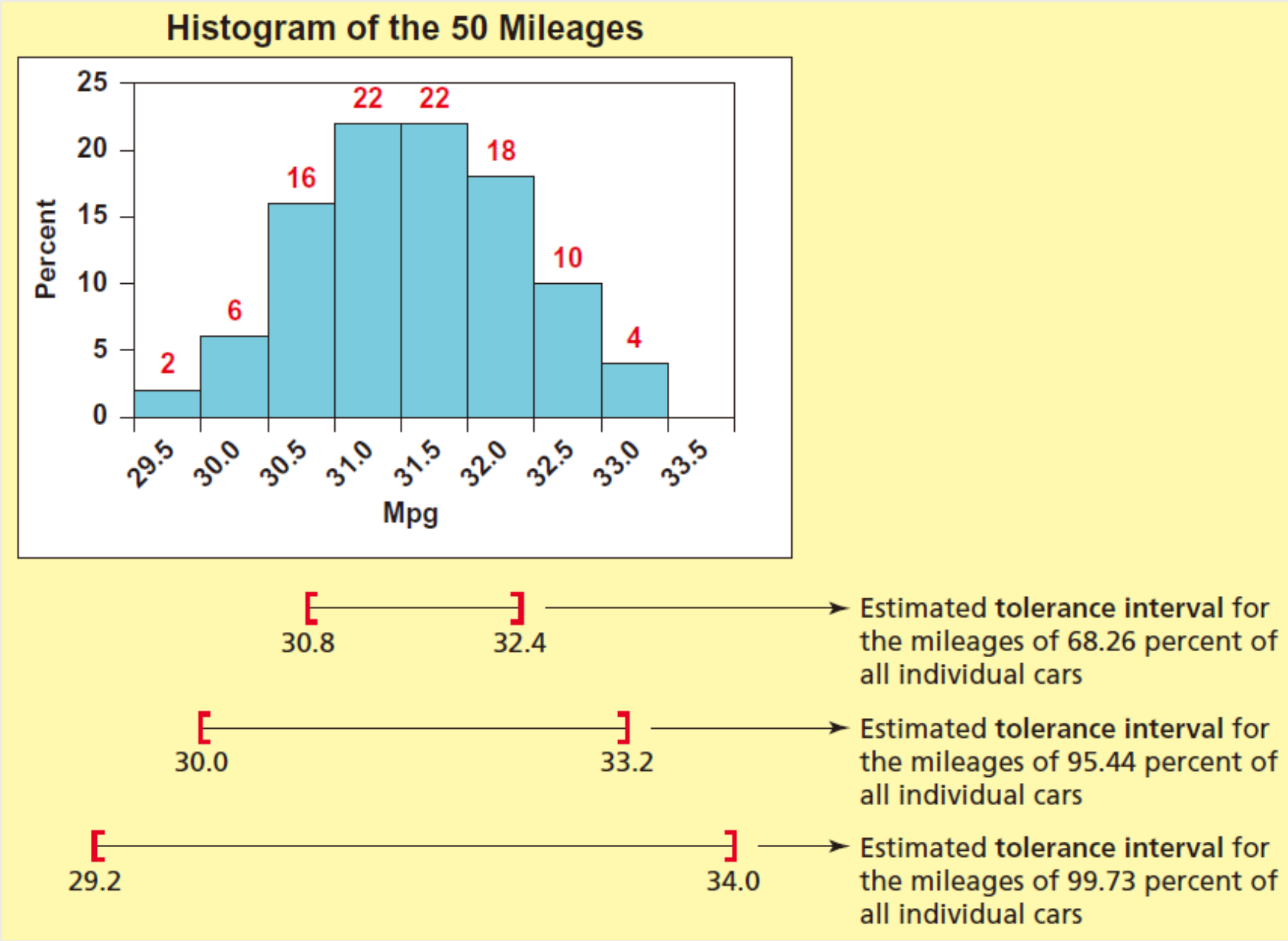
2. **95.44%** of the population measurements lie within two standard deviations of the mean: $[\mu-2\sigma, \mu+2\sigma]$

3. **99.73%** of the population measurements lie within three standard deviations of the mean: $[\mu-3\sigma, \mu+3\sigma]$

The Empirical Rule

- The Empirical Rule holds for normally distributed populations.
- This rule also approximately holds for populations having mound-shaped (single-peaked) distributions that are not very skewed to the right or left.
- For example, recall that the distribution of 65 payment times, it indicates that the empirical rule holds.

Estimated Tolerance Intervals in Care Mileage Case



Estimated Tolerance Intervals in Care Mileage Case

- 68.26% of all individual cars will have mileages in the range

$$[\bar{x} \pm s] = [31.6 \pm 0.8] = [30.8, 32.4] \text{ mpg}$$

- 95.44% of all individual cars will have mileages in the range

$$[\bar{x} \pm 2s] = [31.6 \pm 1.6] = [30.0, 33.2] \text{ mpg}$$

- 99.73% of all individual cars will have mileages in the range

$$[\bar{x} \pm 3s] = [31.6 \pm 2.4] = [29.2, 34.0] \text{ mpg}$$

Because the difference between the upper and lower limits of each estimated tolerance interval is fairly small, we might conclude that the variability of the individual car mileages around the estimated mean mileage of 31.6 mpg is fairly small. Furthermore, the interval [29.2, 34.0] implies that almost any individual car that a customer might purchase this year will obtain a mileage between 29.2 mpg and 34.0 mpg

Skewness and the Empirical Rule

- The **Empirical Rule holds for a normally distributed population**
- It approximately holds for populations having **mound-shaped, single-peaked** distributions. As long as they are **not *very* skewed to the right or left**
- In some situations, skewness can make it tricky to know whether to use the Empirical Rule. **When a distribution seems to be too skewed** for the Empirical Rule to hold, it is probably best to **describe the distribution's variation by using percentiles**

Chebyshev's Theorem

If we fear that the Empirical Rule does not hold for a particular population, we can consider using Chebyshev's Theorem to find an interval that contains a specified percentage of the individual measurements in the population

- Let μ and σ be a population's mean and standard deviation, then for any value k> 1
- At least 100(1 - 1/k^2)% of the population measurements lie in the interval [μ-kσ, μ+kσ]
- Holds for any distribution.
- Only useful for non-mound-shaped distribution population that is not very skewed to left or right

$k = 2$, 100(1-1/2^2)%=100(3/4)%=75% of population lie in the interval $[\mu \pm 2\sigma]$
$k = 3$, 100(1-1/3^2)%=100(8/9)%=88.89% of population lie in the interval $[\mu \pm 3\sigma]$
$k = 19.25$, 100(1-1/19.25^2)%=99.73% of population lie in the interval $[\mu \pm 19.25\sigma]$

Chebyshev's Theorem

Although Chebyshev's Theorem technically applies to any population, it is only of practical use when analyzing a non-mound-shaped (for example, a double-peaked) population that is not very skewed to the right or left.

- Not for a mound-shaped population that is not very skewed because we can use the Empirical Rule to do this.
- On the other hand, if we use Chebyshev's Theorem, the interval $[\mu \pm 19.25\sigma]$ is needed to include 99.73% of population
- Not appropriate to use Chebyshev's Theorem—or any other result making use of the population standard deviation σ—to describe a population that is very skewed. For very skewed, it is best to measure variation by using percentiles.

z-scores

- Determine the relative location of any value in a population or sample
- For any x in a population or sample, the associated z score is

$$z = \frac{x - \text{mean}}{\text{standard deviation}}$$

- The z score is the number of standard deviations that x is from the mean
  - *A positive z score is for x above the mean*
  - *A negative z score is for x below the mean*
  - *The mean has a z score of zero*

z-scores

z-score, also called the standardized value, is the number of standard deviations that x is from the mean.

Company	Profit margin, x	x − mean	z-score
1	8%	$8 - 10 = -2$	$-2/3.406 = -.59$
2	10	$10 - 10 = 0$	$0/3.406 = 0$
3	15	$15 - 10 = 5$	$5/3.406 = 1.47$
4	12	$12 - 10 = 2$	$2/3.406 = .59$
5	5	$5 - 10 = -5$	$-5/3.406 = -1.47$

Mean = 10% Standard deviation = 3.406%

- These z-scores tell us that the profit margin for Company 3 is the farthest above the mean, 1.47 standard deviations above the mean
- Company 5 is the farthest below the mean—it is 1.47 standard deviations below the mean.
- Because the z-score for Company 2 equals zero, its profit margin equals the mean

z-scores

Values in two different populations or samples having the same z-score are the same number of standard deviations from their respective means and, therefore, have the same relative locations.

Example
Suppose that the mean score on the midterm exam for students in Section A of a statistics course is 65 and the standard deviation of the scores is 10. Meanwhile, the mean score on the same exam for students in Section B is 80 and the standard deviation is 5.

A student in Section A who scores an 85 and a student in Section B who scores a 90 have the same relative locations within their respective sections because their *z*-scores, (85-65)/10=2 and (90-80)/5=2, are equal.

Coefficient of Variation

- Measures the size of the standard deviation relative to the size of the mean

$$\text{coefficient of variation} = \frac{\text{standard deviation}}{\text{mean}} \times 100$$

- Used to:
  - *Compare the variability of values to the mean*
  - *Compare variability of populations or samples with different means and standard deviations*
  - *Measure risk.* Larger Coefficient of Variation means higher risk (especially for investment).

Coefficient of Variation

The coefficient of variation compares populations or samples having different means and different standard deviations.

Example
Mean yearly return for Stock Fund 1 is 10.39 percent with a standard deviation of 16.18 percent
Mean yearly return for Stock Fund 2 is 7.7 percent with a standard deviation of 13.82 percent

$(16.18/10.39)*100\%=155.73\%$
$(13.82/7.7)*100\%=179.48\%$

Stock Fund 2 has a higher coefficient of variation than does Stock Fund 1. Investing in Stock Fund 2 is riskier than investing in Stock Fund 1.

Chapter Outline

3.3 Percentiles, Quartiles and Box-and-Whiskers Displays

3.3 Percentiles, Quartiles, and Box-and-Whiskers Displays

For a set of measurements arranged in increasing order, the p^{th} percentile is a value such that p percent of the measurements fall at or below the value and (100-p) percent of the measurements fall at or above the value

- The first quartile $\mathbf{Q_1}$ is the 25^{th} percentile
- The second quartile (or median) is the 50^{th} percentile
- The third quartile $\mathbf{Q_3}$ is the 75^{th} percentile
- The interquartile range IQR is $Q_3 - Q_1$

Steps Calculating Percentiles

1. Arrange the measurements in increasing order
2. Calculate the index i=(p/100)n where p is the percentile to find
3. Calculating the percentile

a) If i is not an integer, round up and the next integer greater than i denotes the p^{th} percentile

b) If i is an integer, the p^{th} percentile is the average of the measurements in the i and i+1 positions

Example (10^{th} Percentile, Q_1, Q_2, Q_3)

7,524	11,070	18,211	26,817	36,551	41,286
49,312	57,283	72,814	90,416	135,540	190,250

- i=(10/100)12=1.2
- Not an integer so round up to 2
- 10^{th} percentile is in the second position so 11,070
- Q_1 = (18,211+26,817)/2=22,514
- $Q_2 = M_d$ = (41,286+49,312)/2 = 45,299
- Q_3 = (72,814+90,416)/2= 81,615
- IQR= Q_3- Q_1=59,101

Example

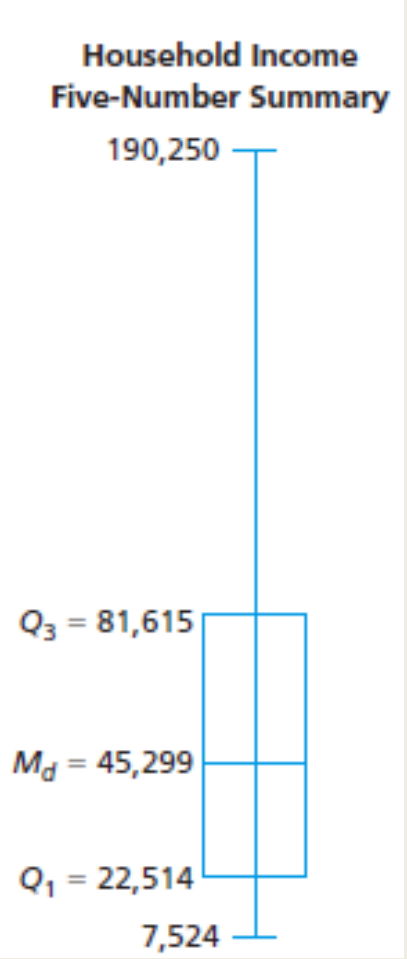
7,524	11,070	18,211	26,817	36,551	41,286
49,312	57,283	72,814	90,416	135,540	190,250

- 10^{th} percentile is in the second position so 11,070
  - *10% of the incomes are less than or equal to $11,070*
- Q_1 = (18,211+26,817)/2=22,514
  - *25% of the incomes are less than or equal to $22,514*
- $Q_2 = M_d$ = (41,286+49,312)/2 = 45,299
  - *25% of the incomes are between $22,514 and $45,299*
- Q_3 = (72,814+90,416)/2= 81,615
  - *25% of the incomes are greater than or equal to $81,615*
- IQR= Q_3- Q_1=59,101

  50 percent of all household incomes fall within a range that is $59,101 long

Five Number Summary

1. The smallest measurement
2. First quartile, Q_1
3. Median, M_d
4. Third quartile, Q_3
5. The largest measurement



- Box-and-whiskers display (sometimes called a box plot)

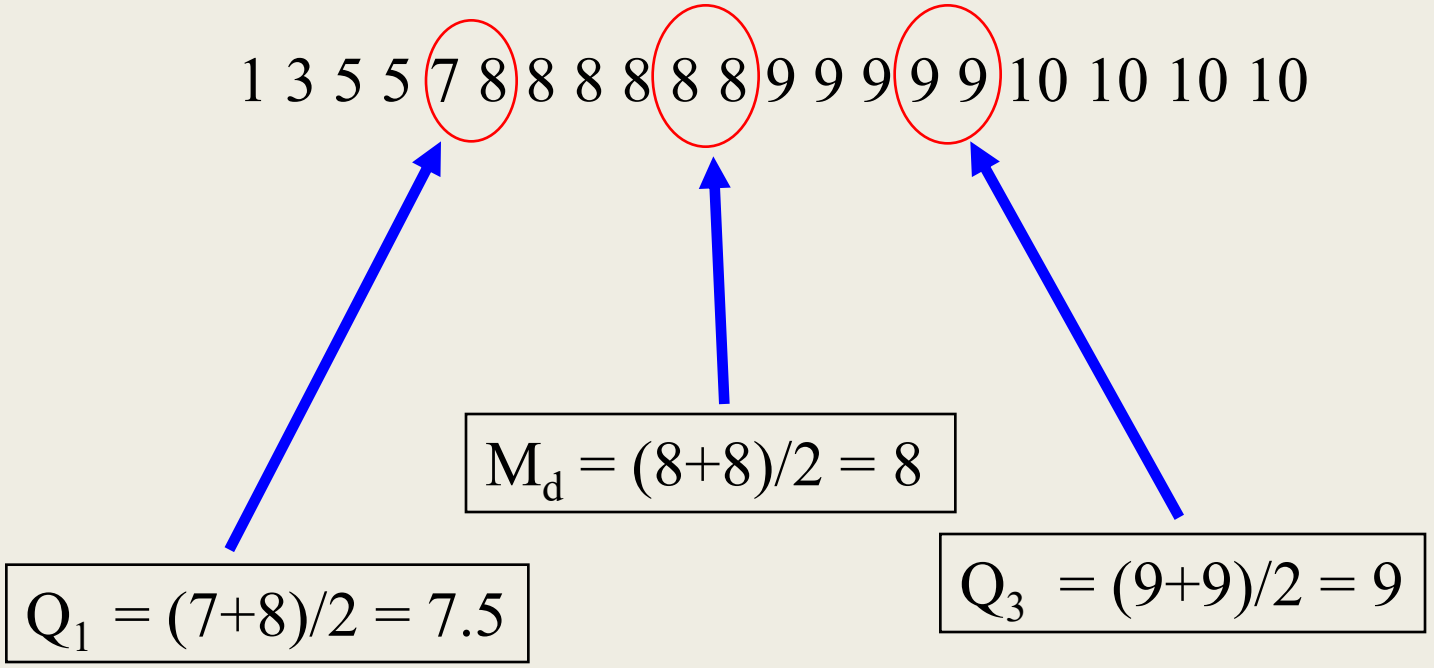
General rule for using percentiles or standard deviation

- In general, if a population is highly skewed to the right or left, it can be best to describe the variation of the population by using various percentiles. This is what we did when we estimated the variation of the household incomes in the city by using the 10th, 25th, 50th, 75th, and 90th percentiles of the 12 sampled incomes and when we depicted this variation by using the five-number summary.
- Using other percentiles can also be informative. For example, the Bureau of the Census sometimes assesses the variation of all household incomes in the United States by using the 20th, 40th, 60th, and 80th percentiles of these incomes.

DVD Recorder Satisfaction

20 customer satisfaction ratings:

1 3 5 5 7 8 8 8 8 8 8 9 9 9 9 9 10 10 10 10



$$IQR = Q_3 - Q_1 = 9 - 7.5 = 1.5$$

The Box-and-Whiskers Plots

- The box plots the:
  - first quartile, Q_1
  - median, M_d
  - third quartile, Q_3
  - inner fences, located 1.5×IQR away from the quartiles:
    - $= Q_1 - (1.5 \times IQR)$
    - $= Q_3 + (1.5 \times IQR)$
  - outer fences, located 3×IQR away from the quartiles:
    - $= Q_1 - (3 \times IQR)$
    - $= Q_3 + (3 \times IQR)$

The Box-and-Whiskers Plots

$Q_1 = (7+8)/2 = 7.5$

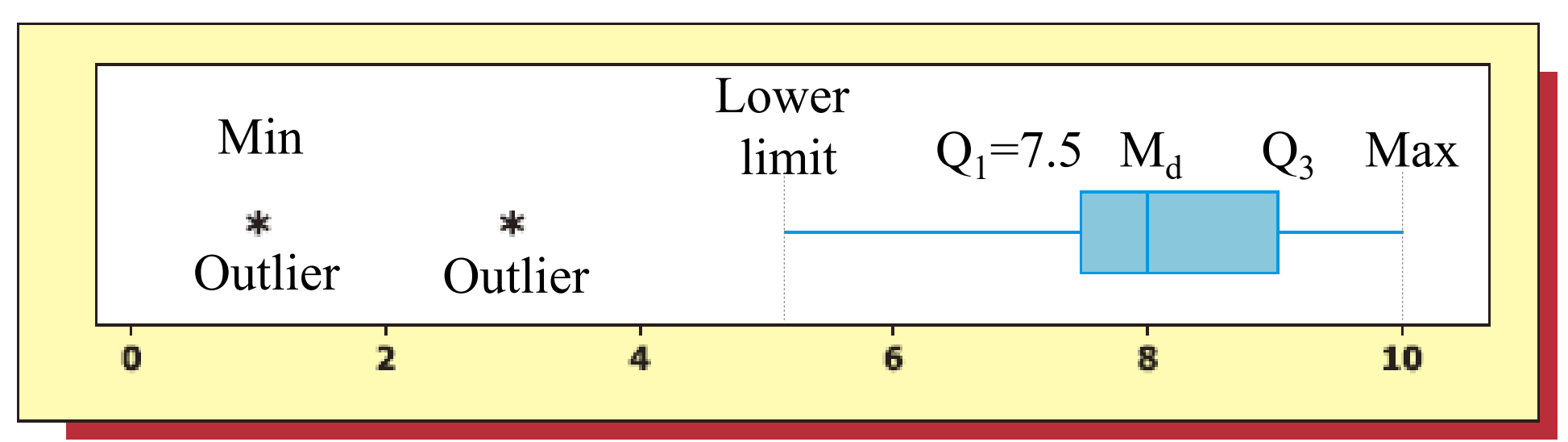
$M_d = (8+8)/2 = 8$

$Q_3 = (9+9)/2 = 9$

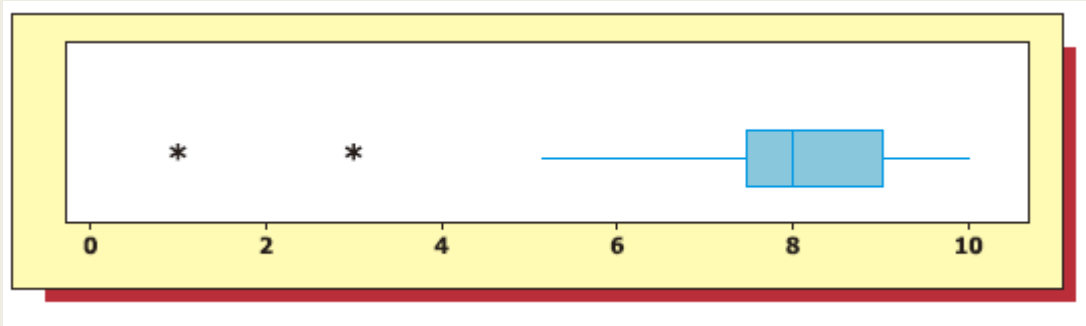
IQR $= Q_3 - Q_1 = 9 - 7.5 = 1.5$

Lower limit $= Q_1 - 1.5IQR = 5.25$

Upper limit $= Q_3 + 1.5IQR = 11.25$



Constructing a Box-and-Whiskers Display (Box Plot)



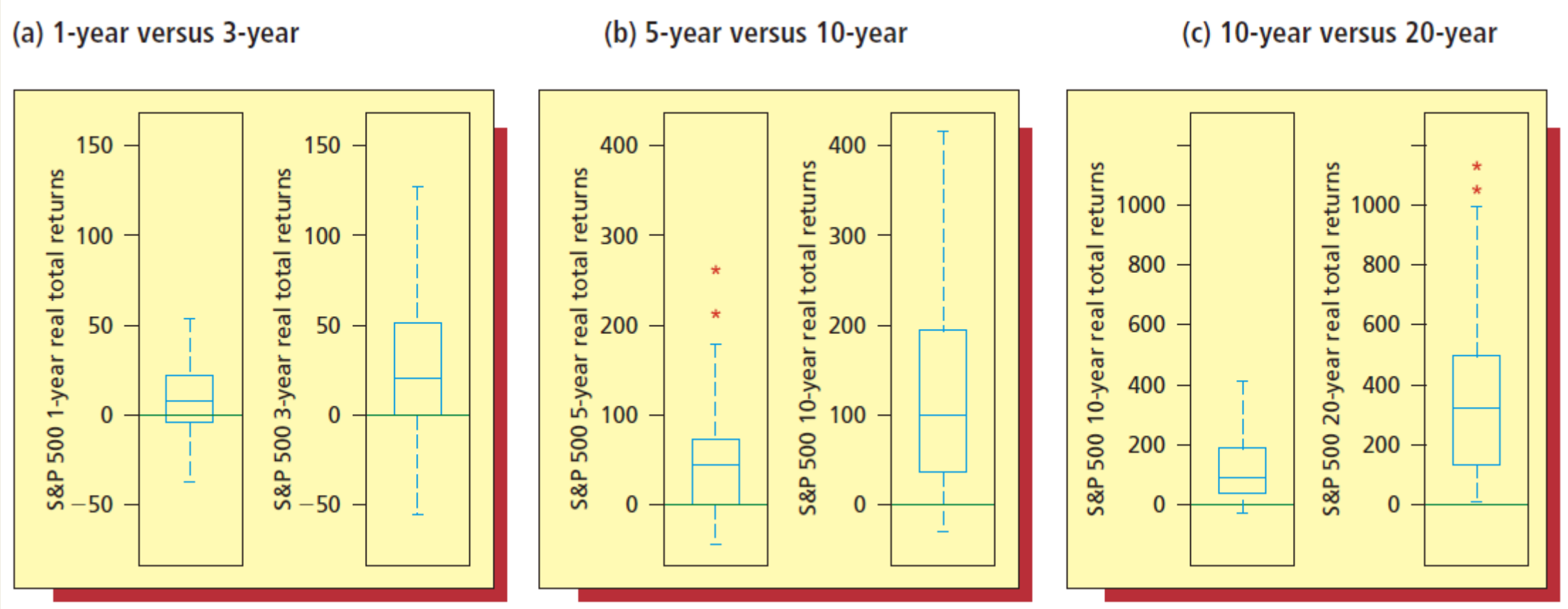
1. Draw a **box** that extends from the first quartile Q_1 to the third quartile Q_3. Also draw a vertical line through the box located at the median *Md*.
2. Determine the values of the **lower** and **upper limits.** The **lower limit** is located 1.5**IQR* below Q_1 and the **upper limit** is located 1.5**IQR* above Q_3. That is, the lower and upper limits are Q_1-1.5(*IQR*) and Q_3+1.5(*IQR*)
3. Draw **whiskers** as dashed lines that extend below Q_1 and above Q_3. Draw one whisker from Q_1 to the *smallest* measurement that is between the lower and upper limits. Draw the other whisker from Q_3 to the *largest* measurement that is between the lower and upper limits.
4. A measurement that is less than the lower limit or greater than the upper limit is an **outlier.** Plot each outlier using the symbol *.

- The “whiskers” are dashed lines that plot the range of the data
  - A dashed line drawn from the box below Q_1 down to the smallest measurement
  - Another dashed line drawn from the box above Q_3 up to the largest measurement
- Note: Q_1, M_d, Q_3, the smallest value, and the largest value are sometimes referred to as the five number summary

Outliers

- Outliers are measurements that are very different from other measurements
  - *They are either much larger or much smaller than most of the other measurements*
- Outliers lie beyond the fences of the box-and-whiskers plot
- Outliers are plotted with an “*”

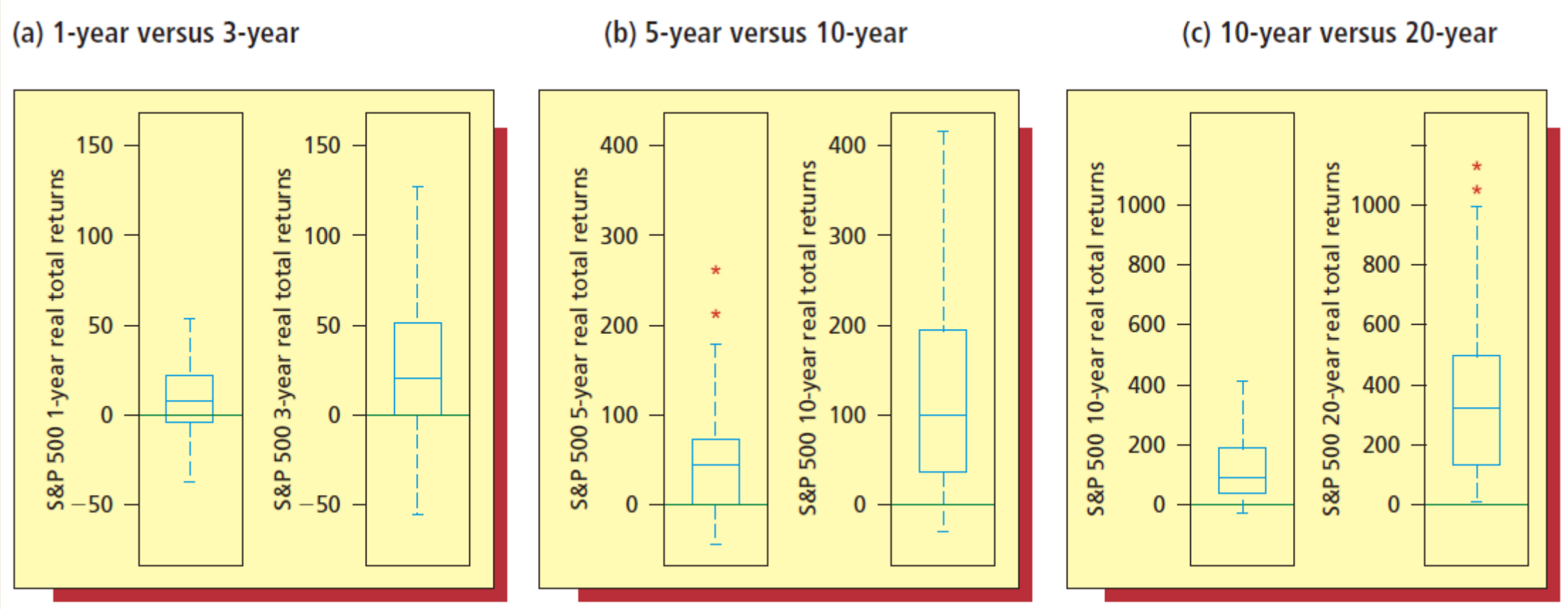
Example 3.9: S&P 500 Case



Box plots of the percentage returns of stocks on the Standard and Poor's 500 (S&P 500) for different time horizons of investment.

There is a 25 percent chance of a negative return (loss) for the 3-year horizon and a 25 percent chance of earning more than 50 percent on the principal during the three years.

Example 3.9: S&P 500 Case



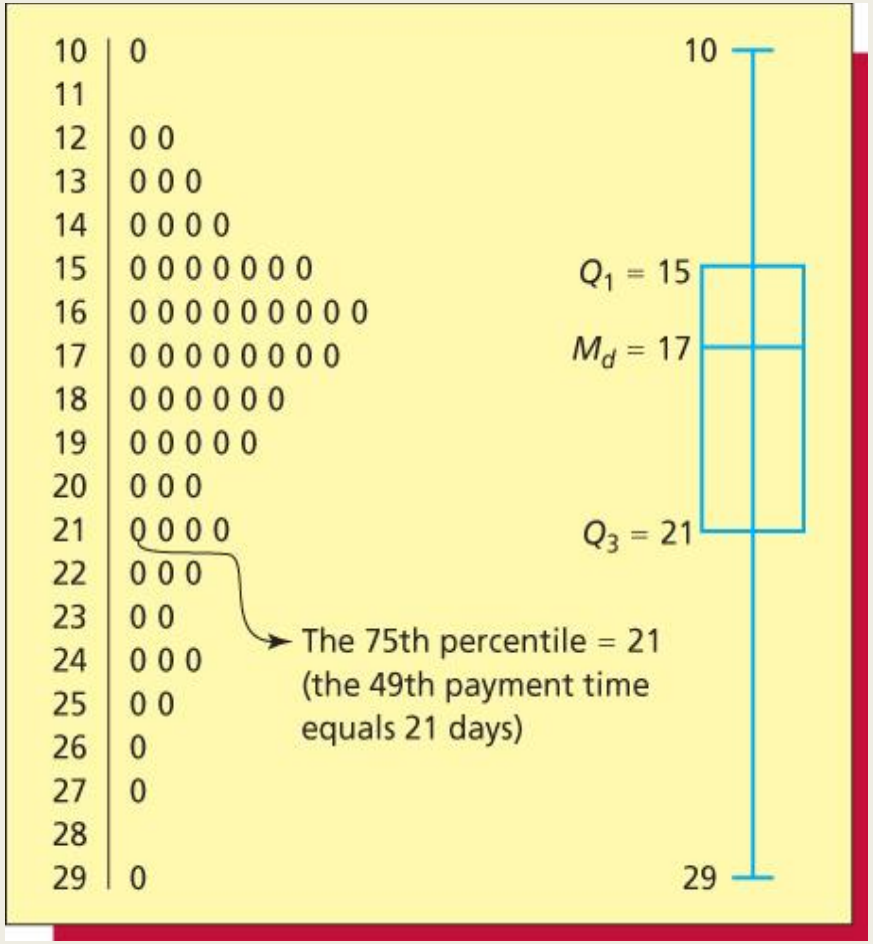
We see that there is still a positive chance of a loss for the 10-year horizon, but the median return for the 10-year horizon almost doubles the principal (a 100 percent return, which is about 8 percent per year compounded). With a 20-year horizon, there is virtually no chance of a loss, and there were two positive outlying returns of over 1000 percent (about 13 percent per year compounded).

Summary

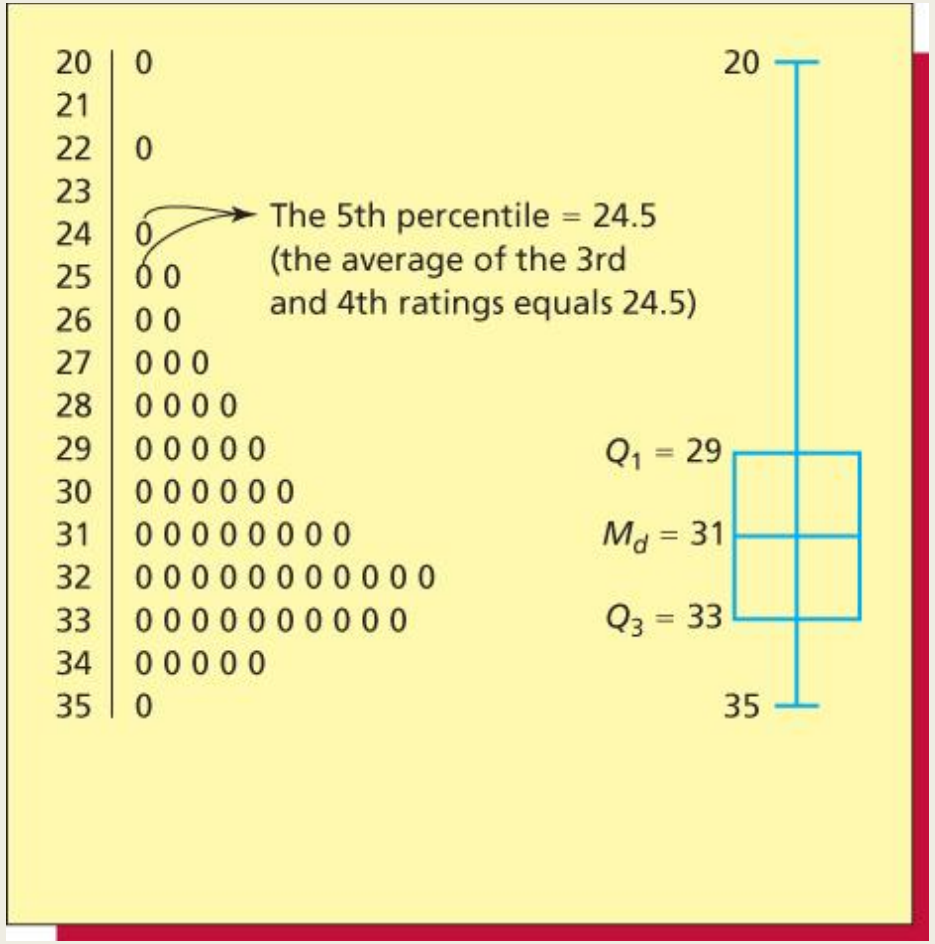
- Central tendency: **mean, median** and **mode.**
- Estimate the population mean by using a sample mean
- **Variation** (or *spread*): **range, variance,** and **standard deviation**
- When population is (approximately) **normally distributed** is to use the **Empirical Rule**
- **Chebyshev's Theorem** for skewed (but not highly) distribution
- For highly skewed data set, use **percentiles** and **quartiles** to measure variation, and construct a **box-and-whiskers plot**

Using stem-and-leaf displays to find percentiles.

(a) The 75th percentile of the 65 payment times, and a five-number summary



(b) The 5th percentile of the 60 bottle design ratings and a five-number summary



3.4 Covariance, Correlation, and the Least Squares Line (Optional)

- Sample covariance

$$s_{xy} = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{n-1}$$

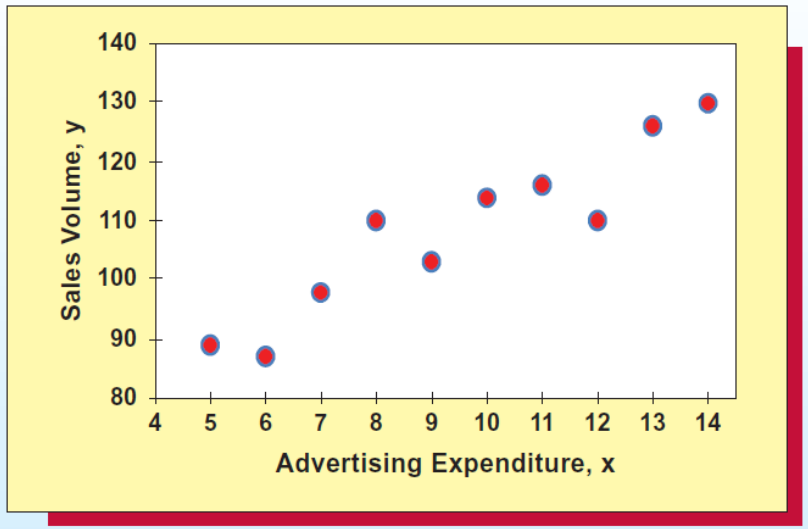
- A positive covariance indicates a positive linear relationship between x and y
  - *As x increases, y increases*
- A negative covariance indicates a negative linear relationship between x and y
  - *As x increases, y decreases*

FIGURE 3.22 The Sales Volume Data, and a Scatter Plot

(a) The sales volume data SalesPlot

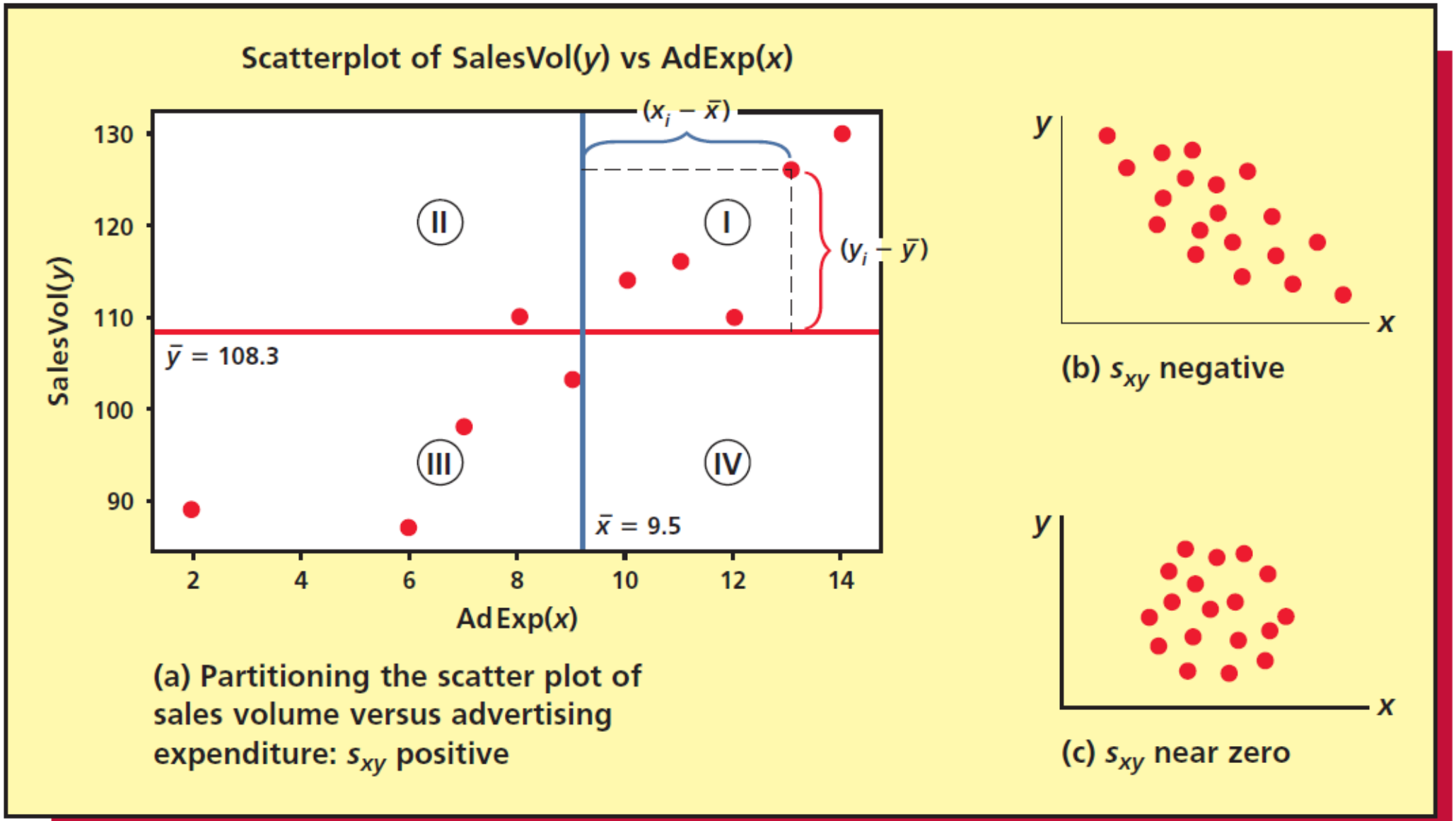
Sales Region	Advertising Expenditure, x	Sales Volume, y
1	5	89
2	6	87
3	7	98
4	8	110
5	9	103
6	10	114
7	11	116
8	12	110
9	13	126
10	14	130

(b) A scatter plot of sales volume versus advertising expenditure



					$9.5)(y_i - 108.3)$
	7	98	−2.5	−10.3	25.75
	8	110	−1.5	1.7	−2.55
	9	103	−0.5	−5.3	2.65
	10	114	0.5	5.7	2.85
	11	116	1.5	7.7	11.55
	12	110	2.5	1.7	4.25
	13	126	3.5	17.7	61.95
	14	130	4.5	21.7	97.65
Totals	95	1083	0	0	365.50

FIGURE 3.23 Interpretation of the Sample Covariance



(a) Partitioning the scatter plot of sales volume versus advertising expenditure: s_{xy} positive

From the previous discussion, it might seem that a large positive value for the covariance indicates that x and y have a strong positive linear relationship and a very negative value for the covariance indicates that x and y have a strong negative linear relationship. However, one problem with using the covariance as a measure of the strength of the linear relationship between x and y is that the value of the covariance depends on the units in which x and y are measured. A measure of the strength of the linear relationship between x and y that does not depend on the units in which x and y are measured is the **correlation coefficient.**

Correlation Coefficient

- Magnitude of covariance does not indicate the strength of the relationship
- Correlation coefficient (r) is a measure of the strength of the relationship that does not depend on the magnitude of the data

$$r = \frac{S_{xy}}{S_x S_y}$$

Correlation Coefficient Continued

- Always between ±1
  - *Near -1 shows strong negative correlation*
  - *Near 0 shows no correlation*
  - *Near +1 shows strong positive correlation*
- Sample correlation coefficient is the point estimate for the population correlation coefficient ρ

It can be shown that the sample correlation coefficient r is always between -1 and 1. A value of r near 0 implies little linear relationship between x and y. A value of r close to 1 says that x and y have a strong tendency to move together in a straight-line fashion with a positive slope and, therefore, that x and y are highly related and **positively correlated.** A value of r close to -1 says that x and y have a strong tendency to move together in a straight-line fashion with a negative slope and, therefore, that x and y are highly related and **negatively correlated.** Note that if $r = 1$, the (x, y) points fall exactly on a positively sloped straight line, and, if $r = -1$, the (x, y) points fall exactly on a negatively sloped straight line. For example, since $r = .93757$ in the sales volume example, we conclude that advertising expenditure (x) and sales volume (y) have a strong tendency to move together in a straight-line fashion with a positive slope. That is, x and y have a strong positive linear relationship.

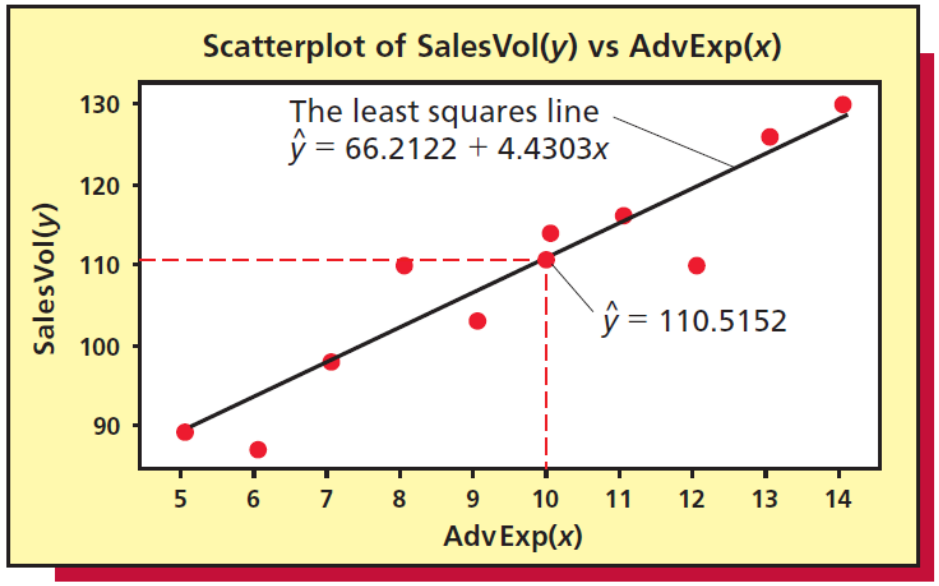
Least Squares Line

$$b_1 = \frac{s_{xy}}{s_x^2}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

- b_0 is the y-intercept
- b_1 is the slope

FIGURE 3.24 The Least Squares Line for the Sales Volume Data



3.5 Weighted Means and Grouped Data

(Optional)

- Sometimes, some measurements are more important than others
- Assign numerical “weights” to the data
- Weights measure relative importance of the value

$$\frac{\sum w_i x_i}{\sum w_i}$$

In order to illustrate the need for a weighted mean and the required calculations, suppose that an investor obtained the following percentage returns on different amounts invested in four stock funds:

Stock Fund	Amount Invested	Percentage Return
1	$50,000	9.2%
2	$10,000	12.8%
3	$10,000	−3.3%
4	$30,000	6.1%

If we wish to compute a mean percentage return for the total of $100,000 invested, we should use a weighted mean. This is because each of the four percentage returns applies to a different amount invested. For example, the return 9.2 percent applies to $50,000 invested and thus should count more heavily than the return 6.1 percent, which applies to $30,000 invested.

The percentage return measurements are $x_1 = 9.2$ percent, $x_2 = 12.8$ percent, $x_3 = -3.3$ percent, and $x_4 = 6.1$ percent, and the weights applied to these measurements are $w_1 = \$50{,}000$, $w_2 = \$10{,}000$, $w_3 = \$10{,}000$, and $w_4 = \$30{,}000$. That is, we are weighting the percentage returns by the amounts invested. The weighted mean is computed as follows:

$$\mu = \frac{50{,}000(9.2) + 10{,}000(12.8) + 10{,}000(-3.3) + 30{,}000(6.1)}{50{,}000 + 10{,}000 + 10{,}000 + 30{,}000}$$

$$= \frac{738{,}000}{100{,}000} = 7.38\%$$

In this case the unweighted mean of the four percentage returns is 6.2 percent. Therefore, the unweighted mean understates the percentage return for the total of $100,000 invested.

Descriptive Statistics for Grouped Data

- Data already categorized into a frequency distribution or a histogram is called grouped data
- Can calculate the mean and variance even when the raw data is not available
- Calculations are slightly different for data from a sample and data from a population

Descriptive Statistics for Grouped Data

(Continued)

Sample

$$\bar{x} = \frac{\sum f_i M_i}{\sum f_i} = \frac{\sum f_i M_i}{n}$$

$$s^2 = \frac{\sum f_i \left(M_i - \bar{x}\right)^2}{n-1}$$

Population

$$\mu = \frac{\sum f_i M_i}{\sum f_i} = \frac{\sum f_i M_i}{N}$$

$$\sigma^2 = \frac{\sum f_i \left(M_i - \bar{x}\right)^2}{N}$$

Table 3.8 and Table 3.8

Sample Mean and Sample Variance of the Satisfaction Rates

Calculating the Sample Mean Satisfaction Rating

Satisfaction Rating	Frequency (f_i)	Class Midpoint (M_i)	f_iM_i
36–38	4	37	4(37) = 148
39–41	15	40	15(40) = 600
42–44	25	43	25(43) = 1,075
45–47	19	46	19(46) = 874
48–50	2	49	2(49) = 98
	$n = 65$		2,795

$$\bar{x} = \frac{\sum f_i M_i}{n} = \frac{2{,}795}{65} = 43$$

Calculating the Sample Variance of the Satisfaction Ratings

Satisfaction Rating	Frequency f_i	Class Midpoint M_i	Deviation $(M_i - \bar{x})$	Squared Deviation $(M_i - \bar{x})^2$	$f_i(M_i - \bar{x})^2$
36–38	4	37	37 − 43 = −6	36	4(36) = 144
39–41	15	40	40 − 43 = −3	9	15(9) = 135
42–44	25	43	43 − 43 = 0	0	25(0) = 0
45–47	19	46	46 − 43 = 3	9	19(9) = 171
48–50	2	49	49 − 43 = 6	36	2(36) = 72
	65				$\sum f_i(M_i - \bar{x})^2 = 522$

$$s^2 = \text{sample variance} = \frac{\sum f_i(M_i - \bar{x})^2}{n - 1} = \frac{522}{65 - 1} = 8.15625$$

3.6 The Geometric Mean (Optional)

- For rates of return of an investment, use the geometric mean
- Suppose the rates of return are R_1, R_2, …, R_n for periods 1, 2, …, n
- The mean of all these returns is the calculated as the geometric mean:

$$R_g = \sqrt[n]{(1+R_1)\times(1+R_2)\times\cdots\times(1+R_n)} - 1$$

THANK YOU!