Chapter 3, Part B
Descriptive Statistics: Numerical Measures

- Measures of Distribution Shape, Relative Location, and Detecting Outliers
- Five-Number Summaries and Box Plots
- Measures of Association Between Two Variables
- Data Dashboards: Adding Numerical Measures to Improve Effectiveness

Measures of Distribution Shape, Relative Location, and Detecting Outliers

- Distribution Shape
- z-Scores
- Chebyshev’s Theorem
- Empirical Rule (經驗法則)
- Detecting Outliers

Distribution Shape: Skewness

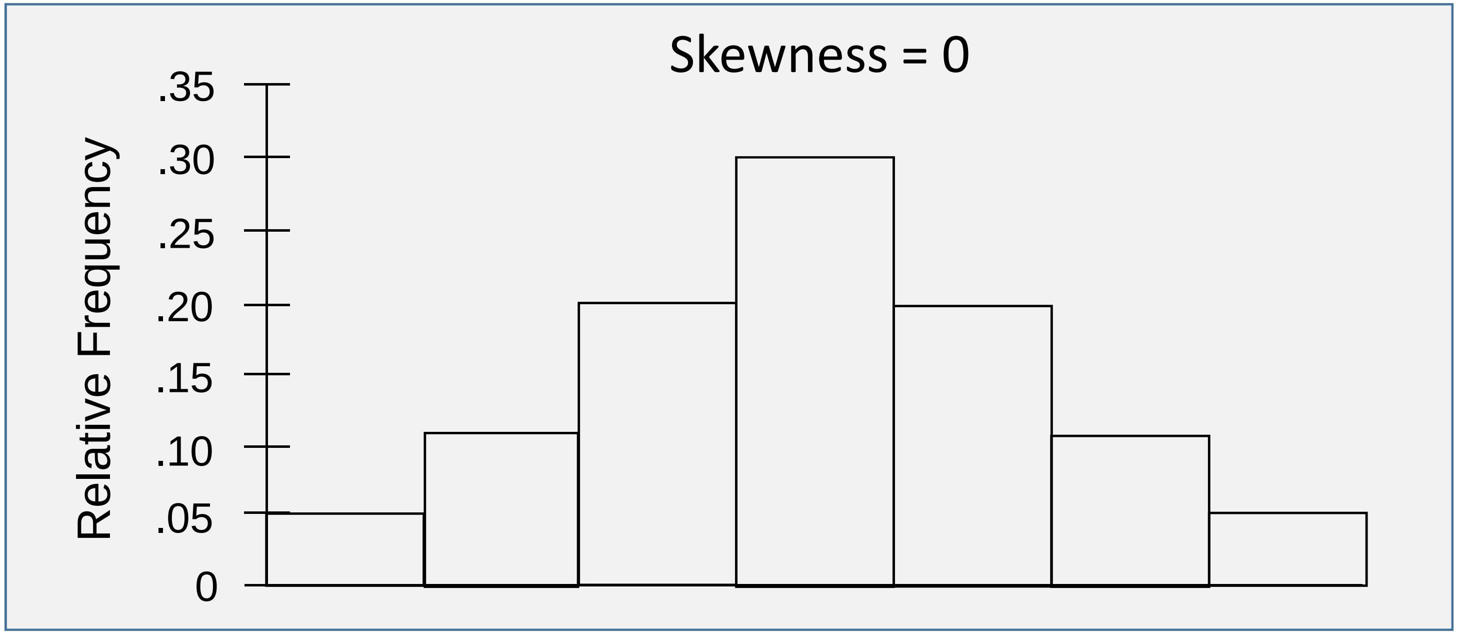
- An important numerical measure of the shape of a distribution is called <u>skewness</u>.
- The formula for the skewness of sample data is

$$\text{Skewness} = \frac{n}{(n-1)(n-2)} \sum \left[\frac{x_i - \bar{x}}{s}\right]^3$$

- Skewness can be easily computed using statistical software.

Distribution Shape: Skewness

- Symmetric (not skewed)
  - Skewness is zero.
  - Mean and median are equal.

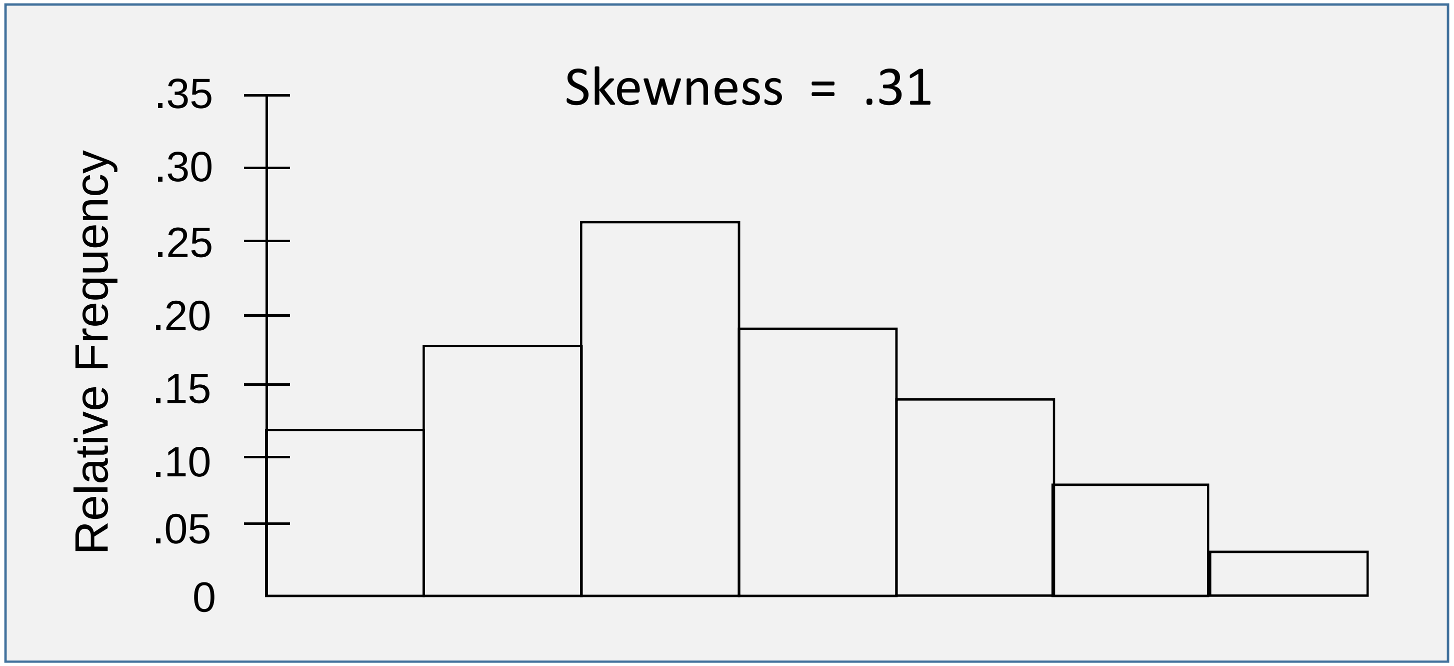


Distribution Shape: Skewness

- Moderately Skewed Left
  - Skewness is negative.
  - Mean will usually be less than the median.

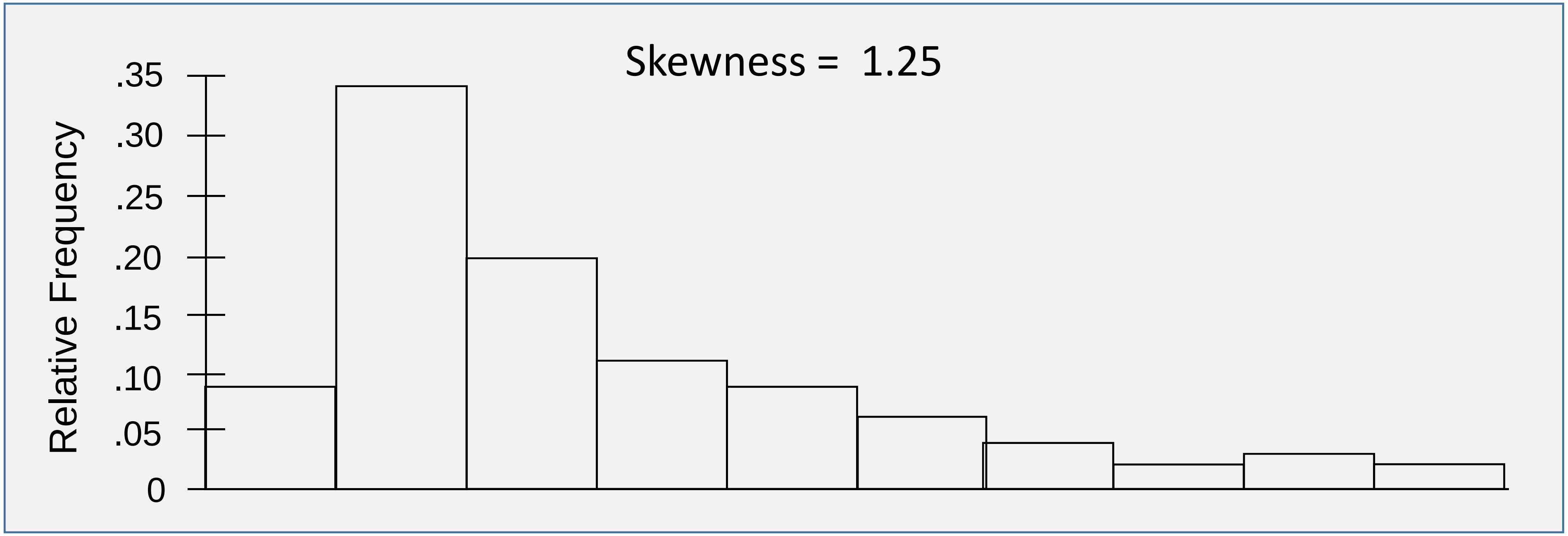
Distribution Shape: Skewness

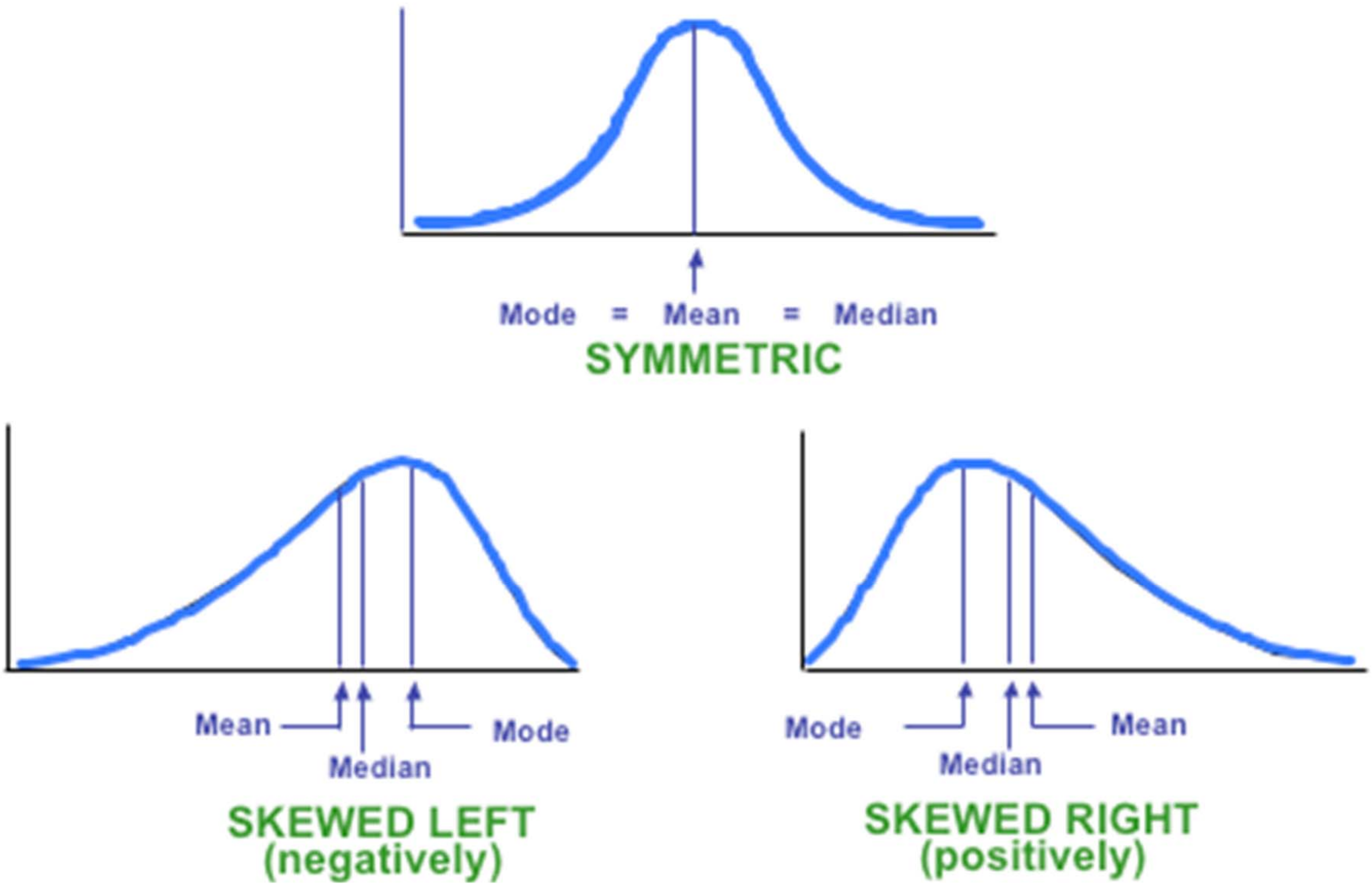
- Moderately Skewed Right
  - Skewness is positive.
  - Mean will usually be more than the median.



Distribution Shape: Skewness

- Highly Skewed Right
  - Skewness is positive (often above 1.0).
  - Mean will usually be more than the median.




Mode = Mean = Median
SYMMETRIC
Mean
Median
Mode
SKEWED LEFT
(negatively)
Mode
Median
Mean
SKEWED RIGHT
(positively)

Distribution Shape: Skewness

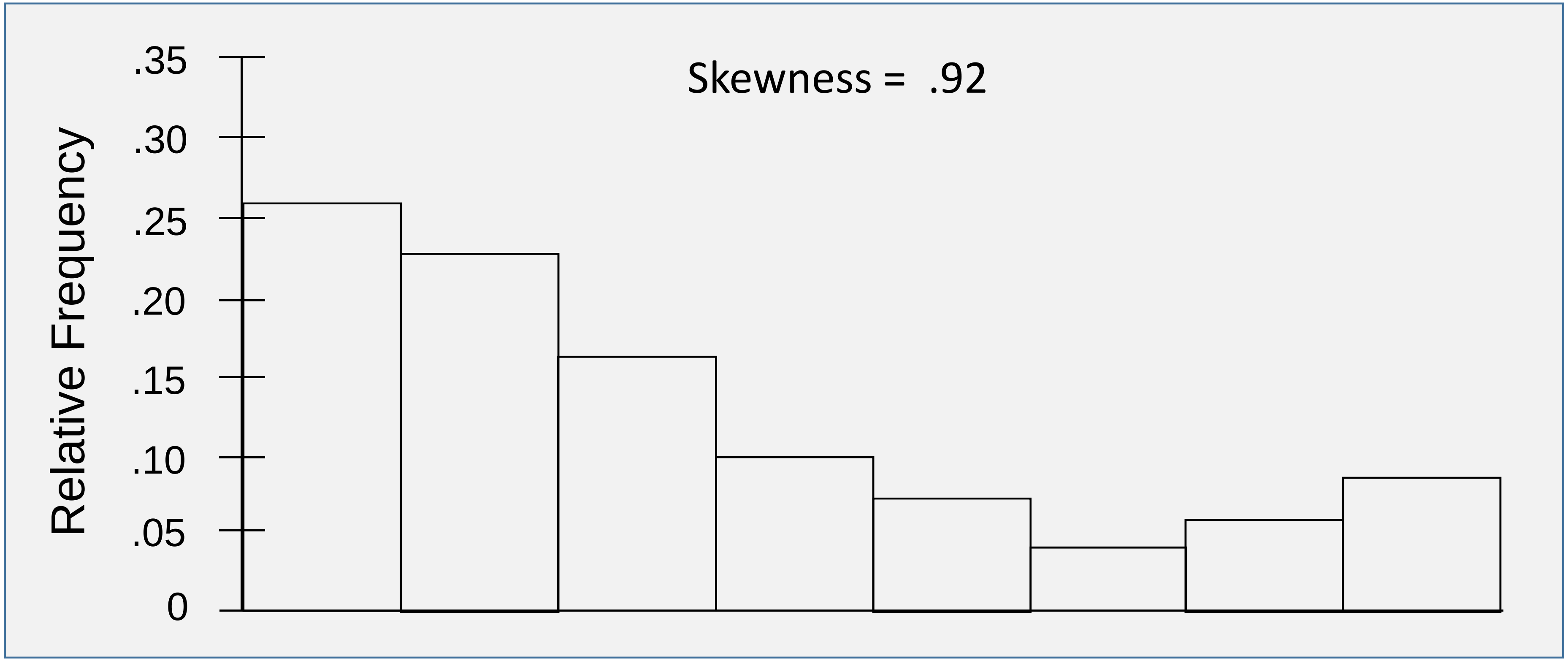
- Example: Apartment Rents

Seventy efficiency apartments were randomly sampled in a college town. The monthly rent prices for the apartments are listed below in ascending order.

525	530	530	535	535	535	535	535	540	540
540	540	540	545	545	545	545	545	550	550
550	550	550	550	550	560	560	560	565	565
565	570	570	572	575	575	575	580	580	580
580	585	590	590	590	600	600	600	600	610
610	615	625	625	625	635	649	650	670	670
675	675	680	690	700	700	700	700	715	715

Distribution Shape: Skewness

- Example: Apartment Rents



z-Scores

- The z-score is often called the standardized value.
- It denotes the number of standard deviations a data value x_i is from the mean.

$$z_i = \frac{x_i - \bar{x}}{s}$$

- An observation's z-score is a measure of the relative location of the observation in a data set.
- A data value less than the sample mean will have a z-score less than zero.
- A data value greater than the sample mean will have a z-score greater than zero.
- A data value equal to the sample mean will have a z-score of zero.

z-Scores

- Example: Apartment Rents
  - z-Score of Smallest Value (525)

$$z_i = \frac{x_i - \bar{x}}{s} = \frac{525 - 590.80}{54.74} = -1.20$$

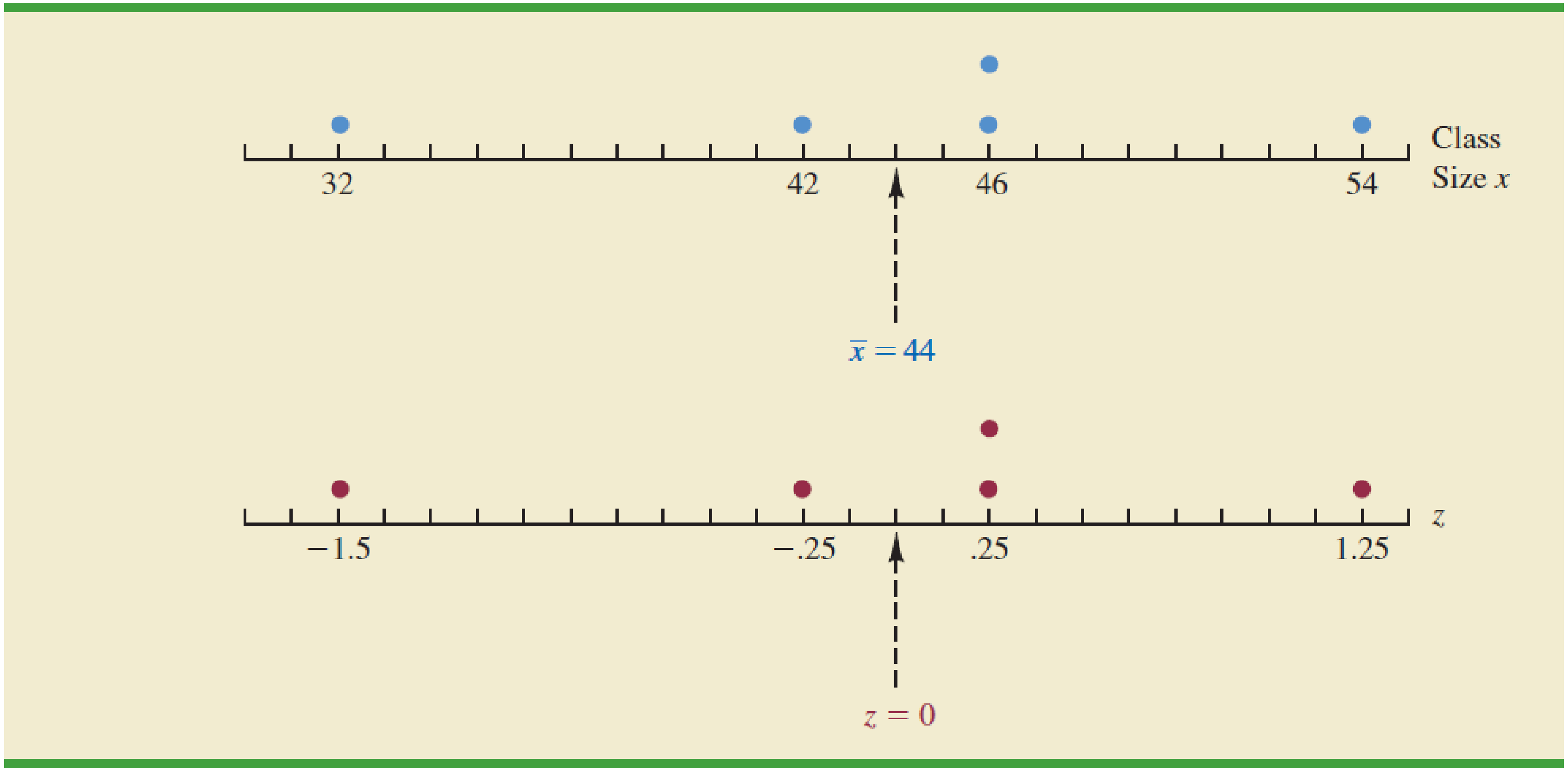
Standardized Values for Apartment Rents

-1.20	-1.11	-1.11	-1.02	-1.02	-1.02	-1.02	-1.02	-0.93	-0.93
-0.93	-0.93	-0.93	-0.84	-0.84	-0.84	-0.84	-0.84	-0.75	-0.75
-0.75	-0.75	-0.75	-0.75	-0.75	-0.56	-0.56	-0.56	-0.47	-0.47
-0.47	-0.38	-0.38	-0.34	-0.29	-0.29	-0.29	-0.20	-0.20	-0.20
-0.20	-0.11	-0.01	-0.01	-0.01	0.17	0.17	0.17	0.17	0.35
0.35	0.44	0.62	0.62	0.62	0.81	1.06	1.08	1.45	1.45
1.54	1.54	1.63	1.81	1.99	1.99	1.99	1.99	2.27	2.27

TABLE 3.5 z-SCORES FOR THE CLASS SIZE DATA

Number of Students in Class (x_i)	Deviation About the Mean ($x_i - \bar{x}$)	z-Score $\left(\frac{x_i - \bar{x}}{s}\right)$
46	2	$2/8 = .25$
54	10	$10/8 = 1.25$
42	-2	$-2/8 = -.25$
46	2	$2/8 = .25$
32	-12	$-12/8 = -1.50$

FIGURE 3.4 DOT PLOT SHOWING CLASS SIZE DATA AND z-SCORES



Chebyshev's Theorem

- At least $(1 - 1/z^2)$ of the data values must be within z standard deviations of the mean, where z is any value greater than 1.

- Chebyshev's theorem requires $z > 1$; but z need not be an integer.

- At least 75% of the data values must be within $z = 2$ standard deviations of the mean.
- At least 89% of the data values must be within $z = 3$ standard deviations of the mean.
- At least 94% of the data values must be within $z = 4$ standard deviations of the mean.

Chebyshev's Theorem

- Example: Apartment Rents

Let z = 1.5 with $\bar{x}$ = 590.80 and s = 54.74

- -

At least $(1 - 1/(1.5)^2) = 1 - 0.44 = 0.56$ or 56%

of the rent values must be between

$\bar{x} - z(s) = 590.80 - 1.5(54.74) = 509$

and

$\bar{x} + z(s) = 590.80 + 1.5(54.74) = 673$

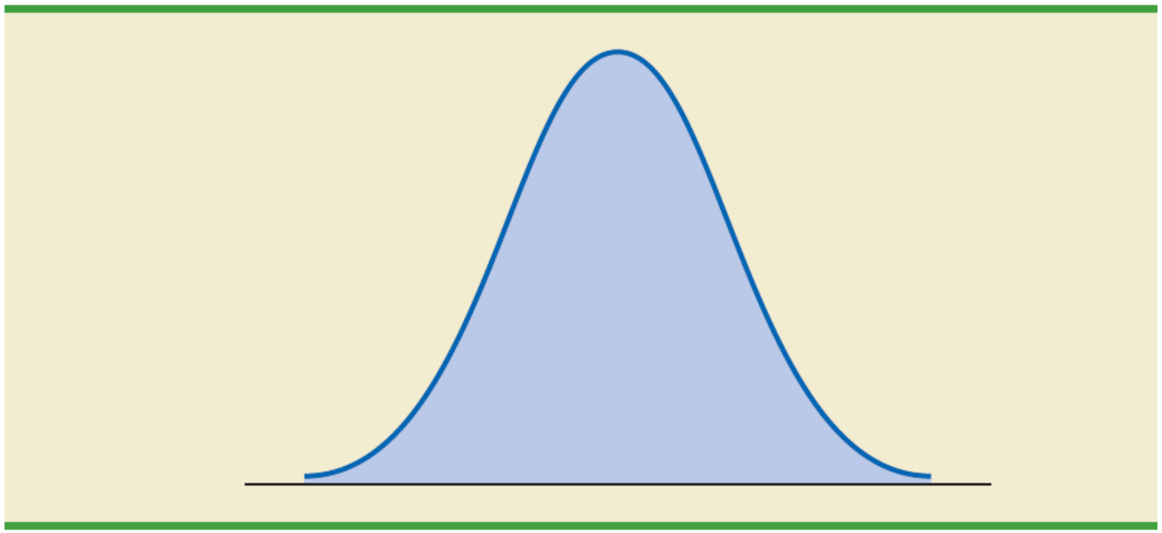
- -

(Actually, 86% of the rent values are between 509 and 673.)

Empirical Rule (經驗法則)

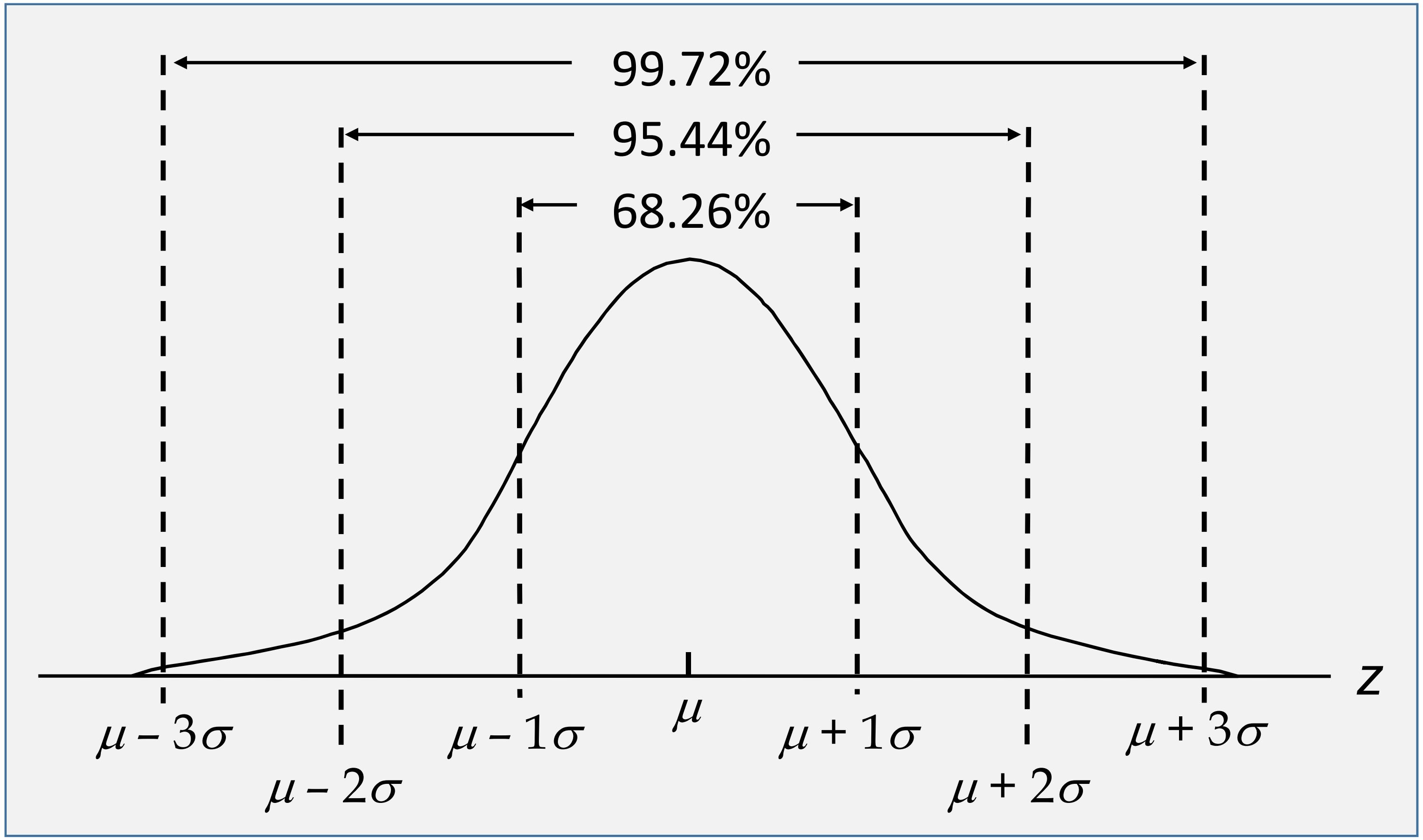
- When the data are believed to approximate a **bell-shaped distribution**:
  - The empirical rule can be used to determine the percentage of data values that must be within a specified number of standard deviations of the mean.
  - The empirical rule is based on the normal distribution, which is covered in Chapter 6.

FIGURE 3.5 A SYMMETRIC MOUND-SHAPED OR BELL-SHAPED DISTRIBUTION



Empirical Rule

- Approximately 68% of the data values will be within one standard deviation of the mean.
- Approximately 95% of the data values will be within two standard deviations of the mean.
- Almost all of the data values will be within three standard deviations of the mean.



For example, liquid detergent cartons are filled automatically on a production line. Filling weights frequently have a bell-shaped distribution. If the mean filling weight is 16 ounces and the standard deviation is .25 ounces, we can use the empirical rule to draw the following conclusions.

- Approximately of the filled cartons will have weights between and ounces (within one standard deviation of the mean).
- Approximately of the filled cartons will have weights between and ounces (within two standard deviations of the mean).
- Almost all filled cartons will have weights between and ounces (within three standard deviations of the mean).

Exercise

The results of a national survey showed that on average, adults sleep 6.9 hours per night. Suppose that the standard deviation is 1.2 hours.

a. Use Chebyshev's theorem to calculate the percentage of individuals who sleep between 4.5 and 9.3 hours.
b. Use Chebyshev's theorem to calculate the percentage of individuals who sleep between 3.9 and 9.9 hours.
c. Assume that the number of hours of sleep follows a bell-shaped distribution. Use the empirical rule to calculate the percentage of individuals who sleep between 4.5 and 9.3 hours per day. How does this result compare to the value that you obtained using Chebyshev's theorem in part (a)?

Detecting Outliers

- An outlier is an unusually small or unusually large value in a data set.
- A data value with a z-score less than -3 or greater than +3 might be considered an outlier.
- It might be:
  - an incorrectly recorded data value
  - a data value that was incorrectly included in the data set
  - a correctly recorded data value that belongs in the data set

Empirical Rule

- Example: Apartment Rents
  - The most extreme z-scores are -1.20 and 2.27.
  - Using $|z| \geq 3$ as the criterion for an outlier, there are no outliers in this data set.

Standardized Values for Apartment Rents

-1.20	-1.11	-1.11	-1.02	-1.02	-1.02	-1.02	-1.02	-0.93	-0.93
-0.93	-0.93	-0.93	-0.84	-0.84	-0.84	-0.84	-0.84	-0.75	-0.75
-0.75	-0.75	-0.75	-0.75	-0.75	-0.56	-0.56	-0.56	-0.47	-0.47
-0.47	-0.38	-0.38	-0.34	-0.29	-0.29	-0.29	-0.20	-0.20	-0.20
-0.20	-0.11	-0.01	-0.01	-0.01	0.17	0.17	0.17	0.17	0.35
0.35	0.44	0.62	0.62	0.62	0.81	1.06	1.08	1.45	1.45
1.54	1.54	1.63	1.81	1.99	1.99	1.99	1.99	2.27	2.27

NCAA college basketball game scores

a. Compute the mean and standard deviation for the points scored by the winning team.
b. Assume that the points scored by the winning teams for all NCAA games follow a bell-shaped distribution. Using the mean and standard deviation found in part (a), estimate the percentage of all NCAA games in which the winning team scores 84 or more points. Estimate the percentage of NCAA games in which the winning team scores more than 90 points.
c. Compute the mean and standard deviation for the winning margin. Do the data contain outliers? Explain.

Winning Team	Points		Losing Team	Points		Winning Margin
Arizona	90		Oregon	66		24
Duke	85		Georgetown	66		19
Florida State	75		Wake Forrest	70		5
Kansas	78		Colorado	57		21
Kentucky	71		Notre Dame	63		8
Louisville	65		Tennessee	62		3
Oklahoma State	72		Texas	66		6
Purdue	76		Michigan State	70		6
Stanford	77		Southern Cal	67		10
Wisconsin	76		Illinois	56		20

Five-Number Summaries and Box Plots

- Summary statistics and easy-to-draw graphs can be used to quickly summarize large quantities of data.
- Two tools that accomplish this are five-number summaries and box plots.

Five-Number Summary

1. Smallest Value
2. First Quartile
3. Median
4. Third Quartile
5. Largest Value

Five-Number Summary

- Example: Apartment Rents

Lowest Value = 525 First Quartile = 545

Median = 575

Third Quartile = 625 Largest Value = 715

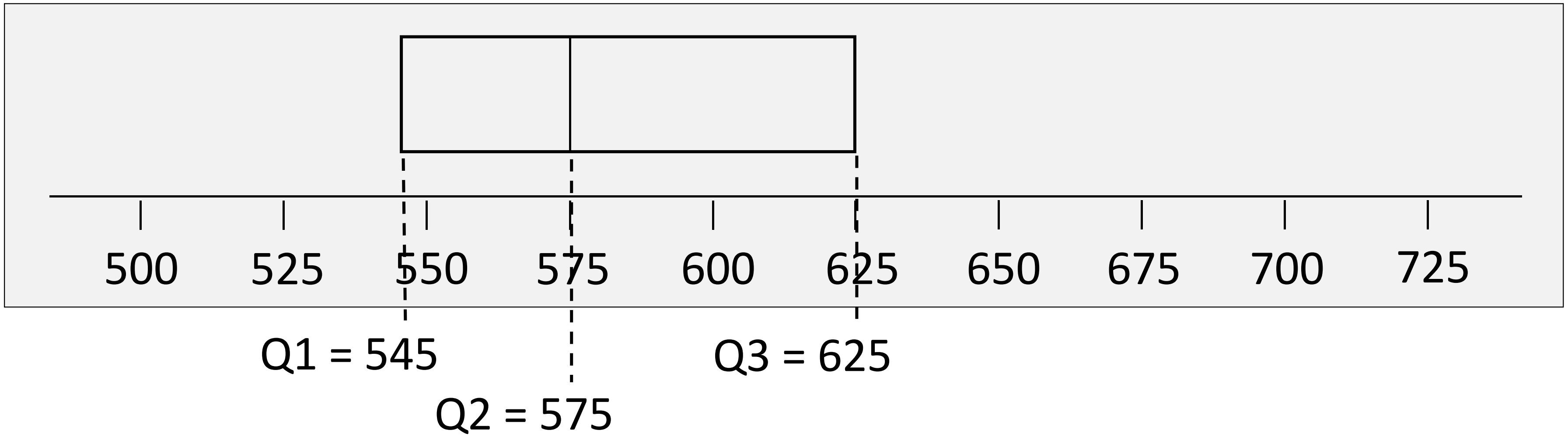
$\boxed{525}$	530	530	535	535	535	535	535	540	540
540	540	540	545	545	545	$\boxed{545}$	$\boxed{545}$	550	550
550	550	550	550	550	560	560	560	565	565
565	570	570	572	$\boxed{575}$	$\boxed{575}$	575	580	580	580
580	585	590	590	590	600	600	600	600	610
610	615	$\boxed{625}$	$\boxed{625}$	625	635	649	650	670	670
675	675	680	690	700	700	700	700	715	$\boxed{715}$

Box Plot

- A box plot is a graphical display of data that is based on a five-number summary.
- A key to the development of a box plot is the computation of the median and the quartiles Q_1 and Q_3.
- Box plots provide another way to identify outliers.

Box Plot

- Example: Apartment Rents
  - A box is drawn with its ends located at the first and third quartiles.
  - A vertical line is drawn in the box at the location of the median (second quartile).



Box Plot

- Limits are located (not drawn) using the interquartile range (IQR).
- Data outside these limits are considered <u>outliers</u>.
- The location of each outlier is shown with the symbol $*$.

Box Plot

- Example: Apartment Rents
  - The lower limit is located 1.5(IQR) below $Q1$.

Lower Limit: Q1 - 1.5(IQR) = 545 - 1.5(80) = 425

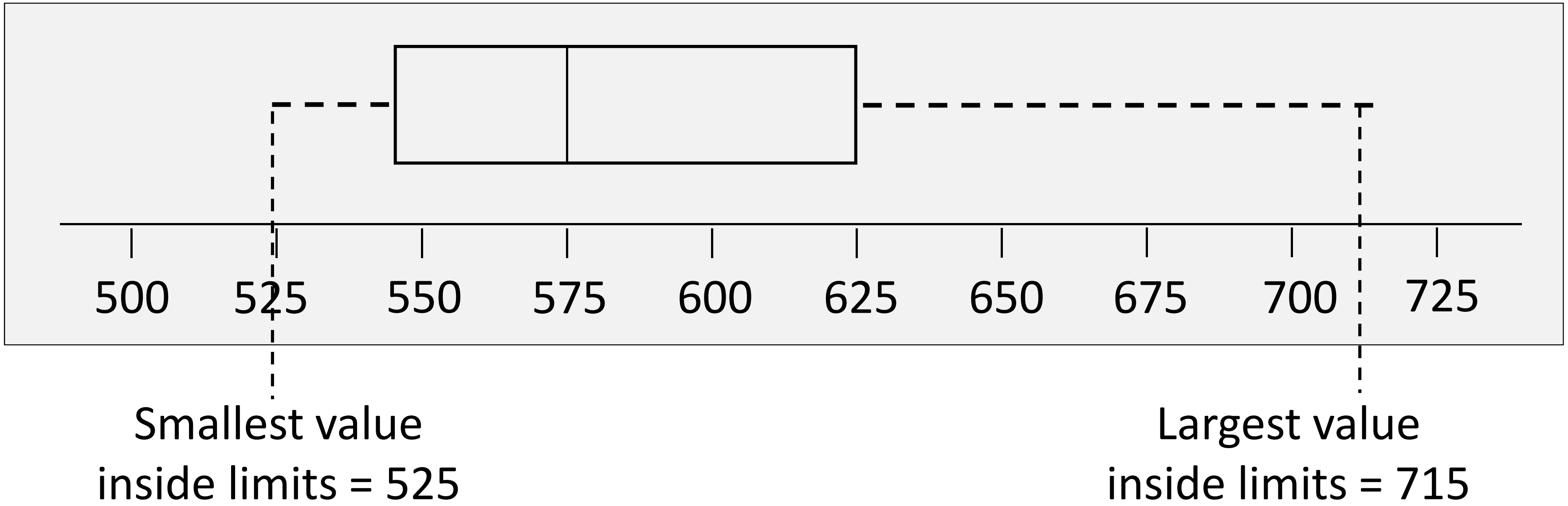
  - The upper limit is located 1.5(IQR) above $Q3$.

Upper Limit: Q3 + 1.5(IQR) = 625 + 1.5(80) = 745

  - There are no outliers (values less than 425 or greater than 745) in the apartment rent data.

Box Plot

- Example: Apartment Rents
  - Whiskers (dashed lines) are drawn from the ends of the box to the smallest and largest data values inside the limits.



25% 的觀察值			25% 的觀察值			25% 的觀察值			25% 的觀察值		
3710	3755	3850	3880	3880	3890	3920	3940	3950	4050	4130	4325

$Q_1 = 3865$ $Q_2 = 3905$ (中位數) $Q_3 = 4000$

FIGURE 3.6 BOX PLOT OF THE MONTHLY STARTING SALARY DATA WITH LINES SHOWING THE LOWER AND UPPER LIMITS

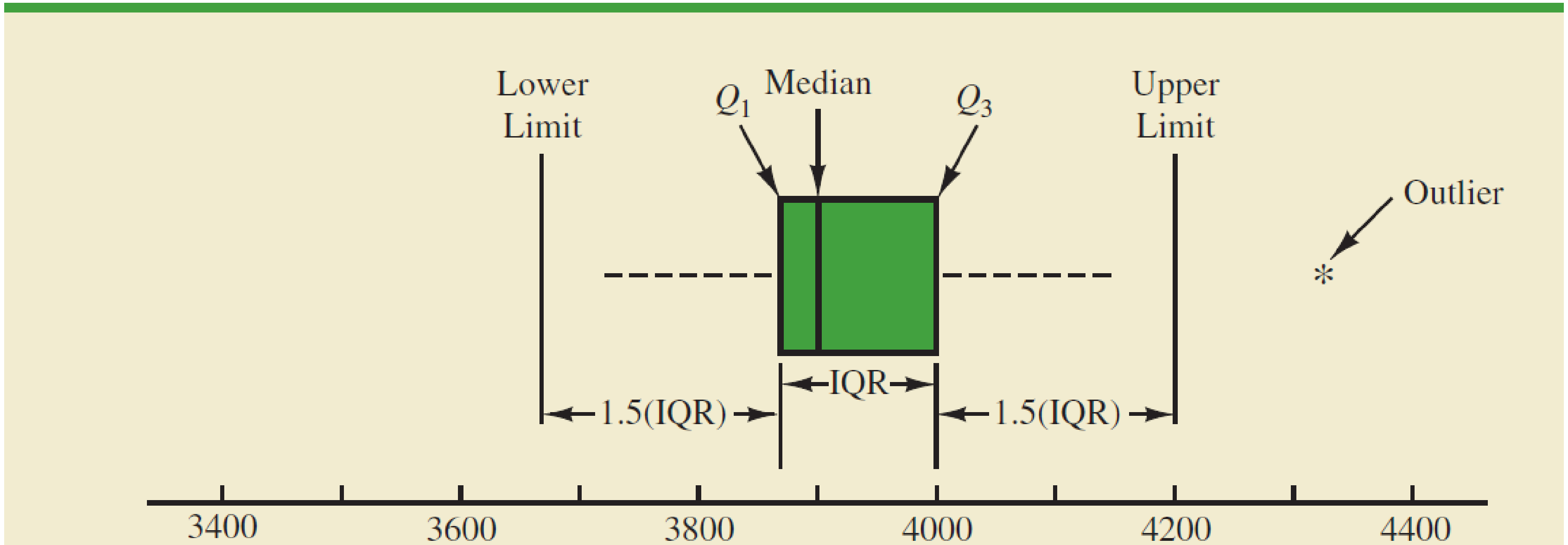


FIGURE 3.7 BOX PLOT OF THE MONTHLY STARTING SALARY DATA

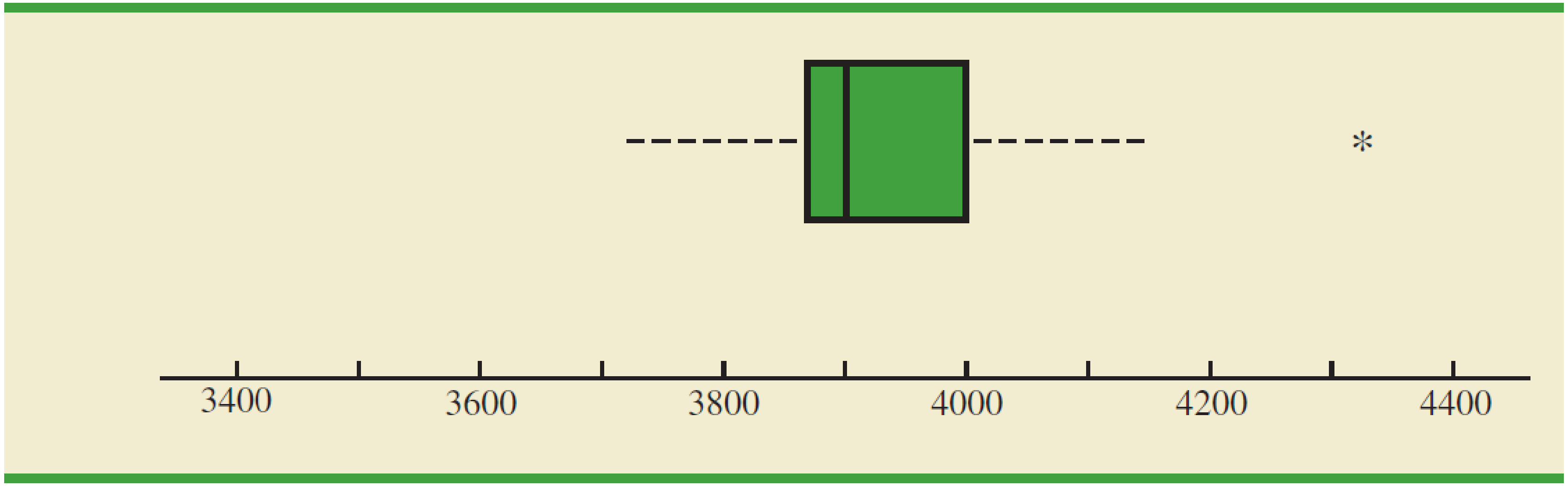
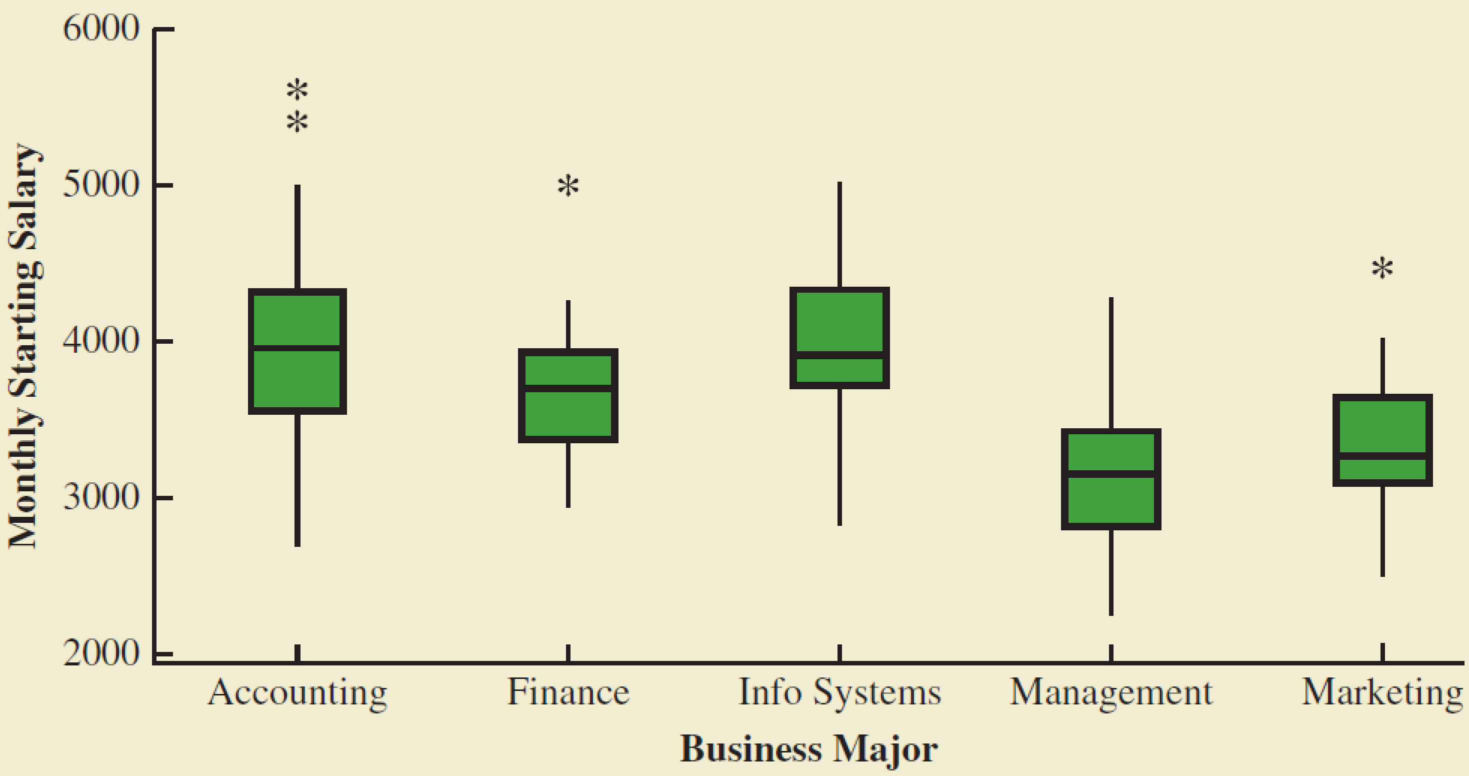


FIGURE 3.8 MINITAB BOX PLOTS OF MONTHLY STARTING SALARY BY MAJOR



Measures of Association Between Two Variables

- Thus far we have examined numerical methods used to summarize the data for one variable at a time.
- Often a manager or decision maker is interested in the relationship between two variables.
- Two descriptive measures of the relationship between two variables are covariance and correlation coefficient.

TABLE 3.6 SAMPLE DATA FOR THE STEREO AND SOUND EQUIPMENT STORE

Week	Number of Commercials x	Sales Volume ($100s) y
1	2	50
2	5	57
3	1	41
4	3	54
5	4	54
6	1	38
7	5	63
8	3	48
9	4	59
10	2	46

FIGURE 3.9 SCATTER DIAGRAM FOR THE STEREO AND SOUND EQUIPMENT STORE

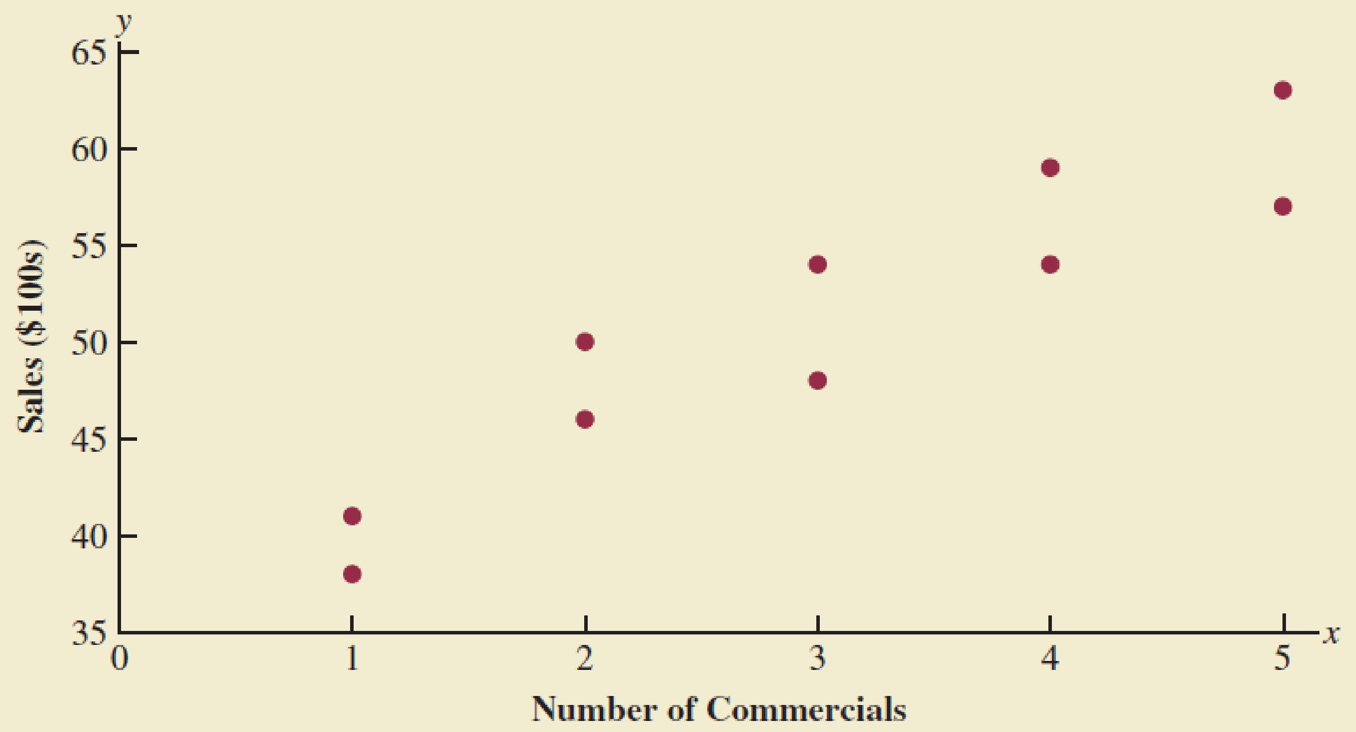


FIGURE 3.10 PARTITIONED SCATTER DIAGRAM FOR THE STEREO AND SOUND EQUIPMENT STORE

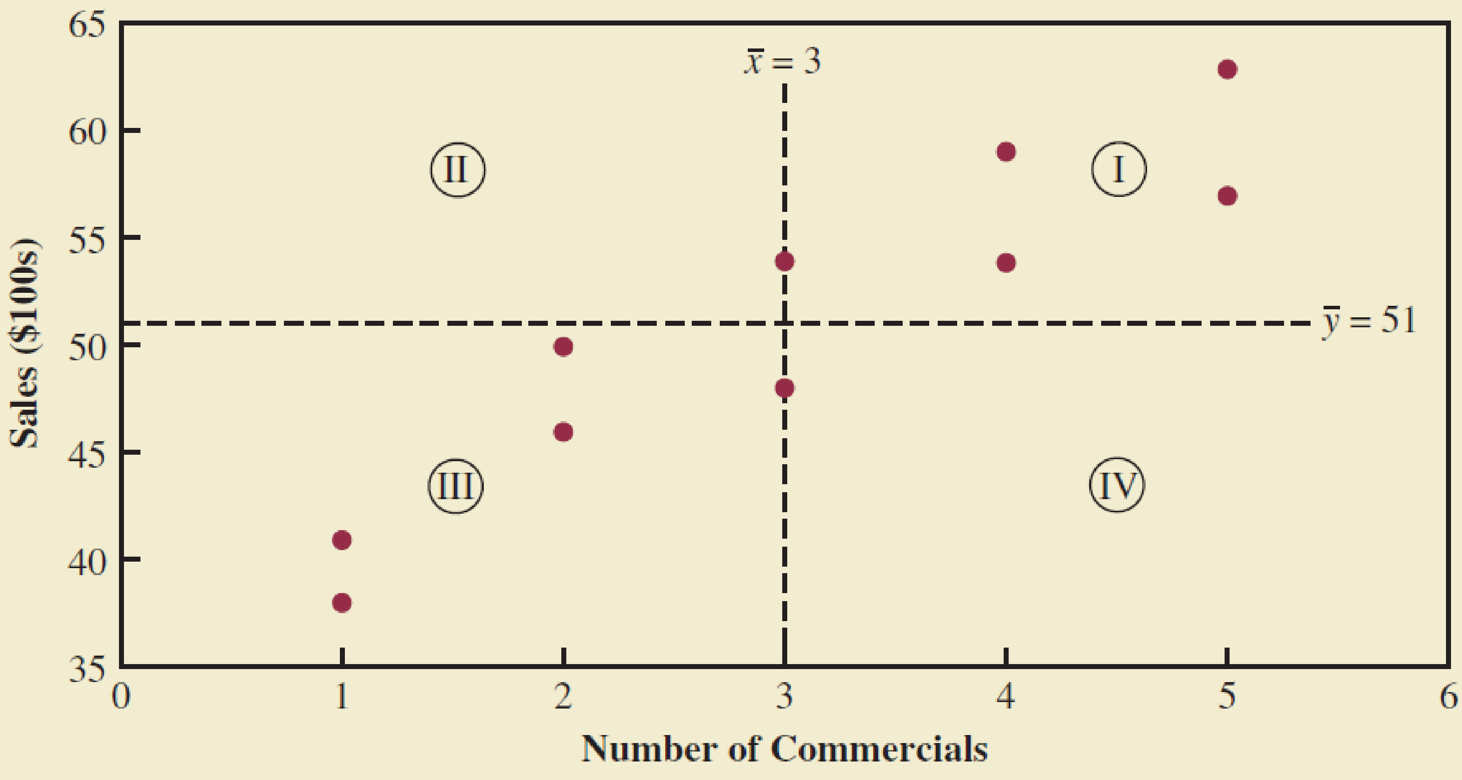
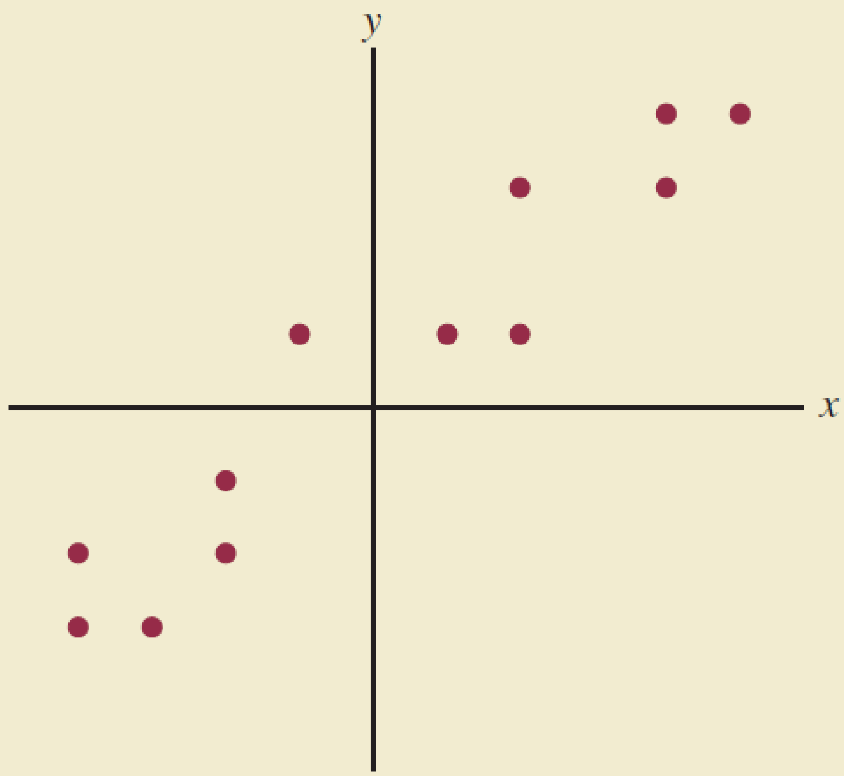
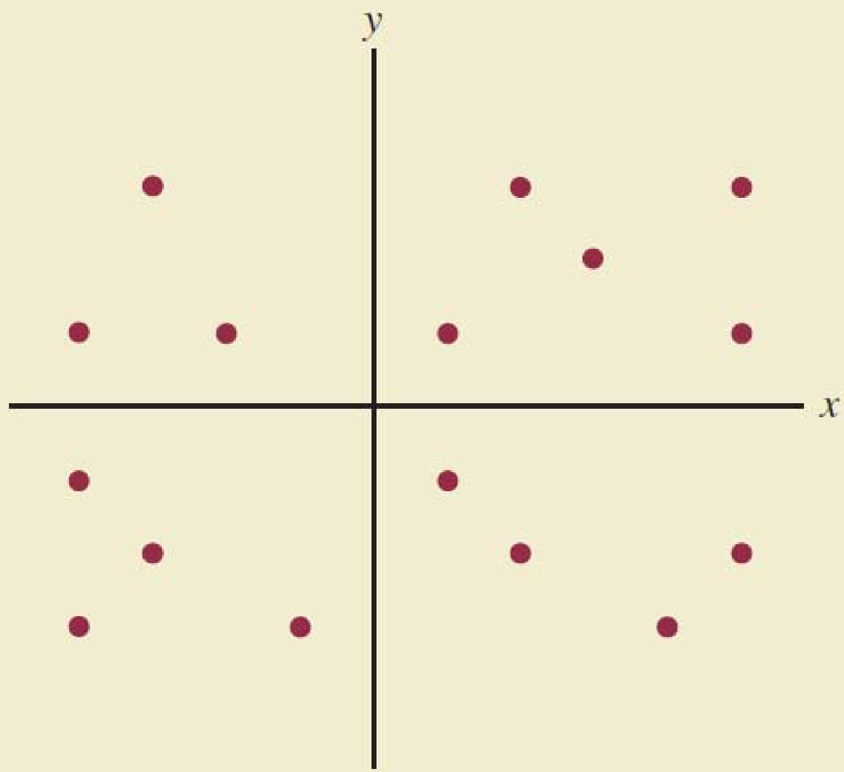


FIGURE 3.11 INTERPRETATION OF SAMPLE COVARIANCE

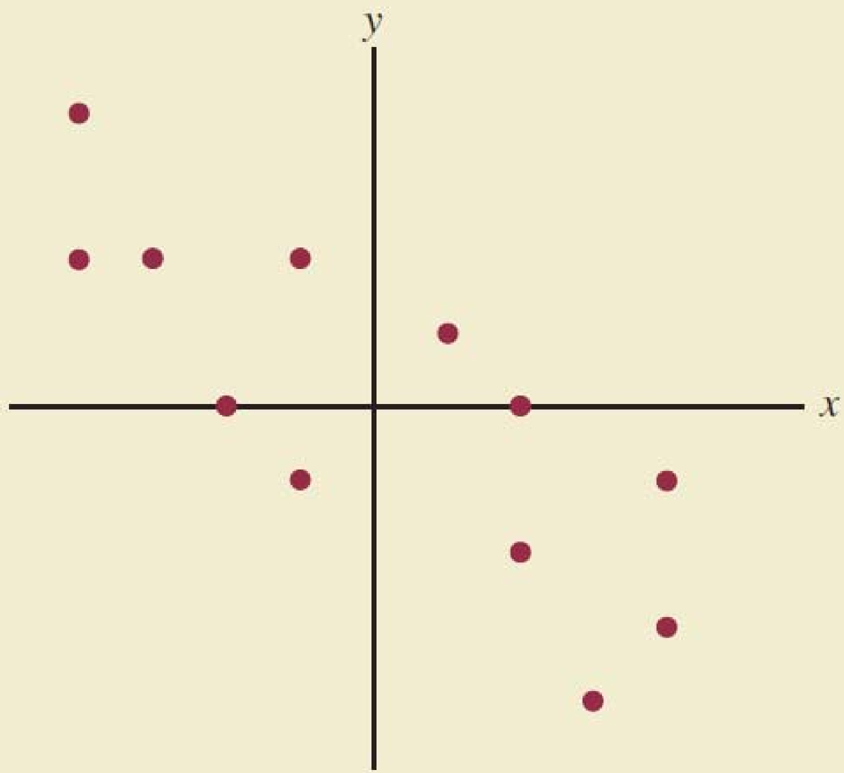
s_{xy} Positive:
(*x* and *y* are positively
linearly related)



s_{xy} Approximately 0:
(x and y are not linearly related)



s_{xy} **Negative:**
(x and y are negatively linearly related)



Covariance

- The covariance is a measure of the linear association between two variables.
- Positive values indicate a positive relationship.
- Negative values indicate a negative relationship.

For samples: $s_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1}$

For populations: $\sigma_{xy} = \frac{\sum(x_i - \mu_x)(y_i - \mu_y)}{N}$

TABLE 3.7 CALCULATIONS FOR THE SAMPLE COVARIANCE

	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
	2	50	−1	−1	1
	5	57	2	6	12
	1	41	−2	−10	20
	3	54	0	3	0
	4	54	1	3	3
	1	38	−2	−13	26
	5	63	2	12	24
	3	48	0	−3	0
	4	59	1	8	8
	2	46	−1	−5	5
Totals	30	510	0	0	99

$$s_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1} = \frac{99}{10 - 1} = 11$$

Correlation Coefficient

- Correlation is a measure of linear association and not necessarily causation.
- Just because two variables are highly correlated, it does not mean that one variable is the cause of the other.

For samples: $r_{xy} = \frac{s_{xy}}{s_x s_y}$

For populations: $\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$

Correlation Coefficient

- The coefficient can take on values between -1 and +1.
- Values near -1 indicate a strong negative linear relationship.
- Values near +1 indicate a strong positive linear relationship.
- The closer the correlation is to zero, the weaker the relationship.

Covariance and Correlation Coefficient

- Example: Golfing Study

  A golfer is interested in investigating the relationship, if any, between driving distance and 18-hole score.

Average Driving Distance (yds.)	Average 18-Hole Score
277.6	69
259.5	71
269.1	70
267.0	70
255.6	71
272.9	69

Covariance and Correlation Coefficient

- Example: Golfing Study

	x	y	$(x_i\text{-}\bar{x})$	$(y_i\text{-}\bar{y})$	$(x_i\text{-}\bar{x})(y_i\text{-}\bar{y})$
	277.6	69	10.65	-1.0	-10.65
	259.5	71	-7.45	1.0	-7.45
	269.1	70	2.15	0	0
	267.0	70	0.05	0	0
	255.6	71	-11.35	1.0	-11.35
	272.9	69	5.95	-1.0	-5.95
Average	267.0	70.0		Total	-35.40
Std. Dev.	8.2192	.8944			

Covariance and Correlation Coefficient

- Example: Golfing Study
  - Sample Covariance

$$s_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{-35.40}{6-1} = -7.08$$

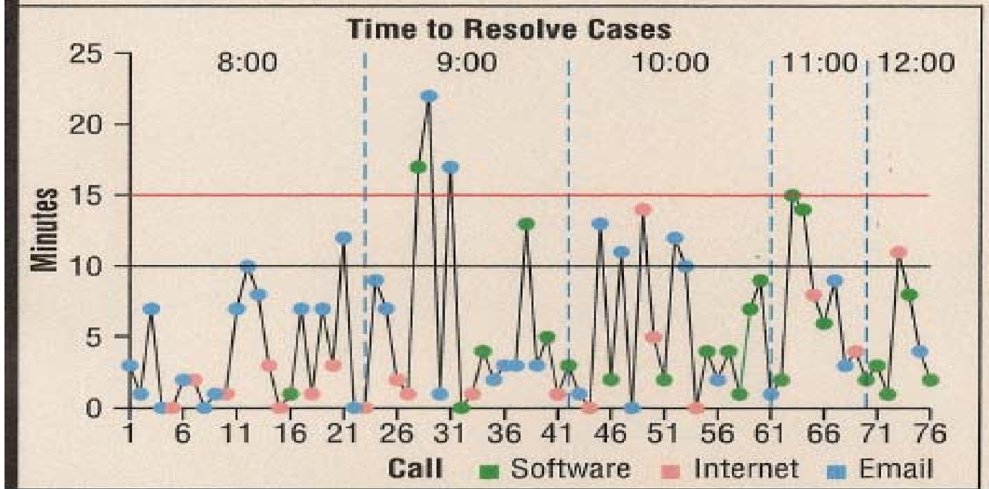
  - Sample Correlation Coefficient

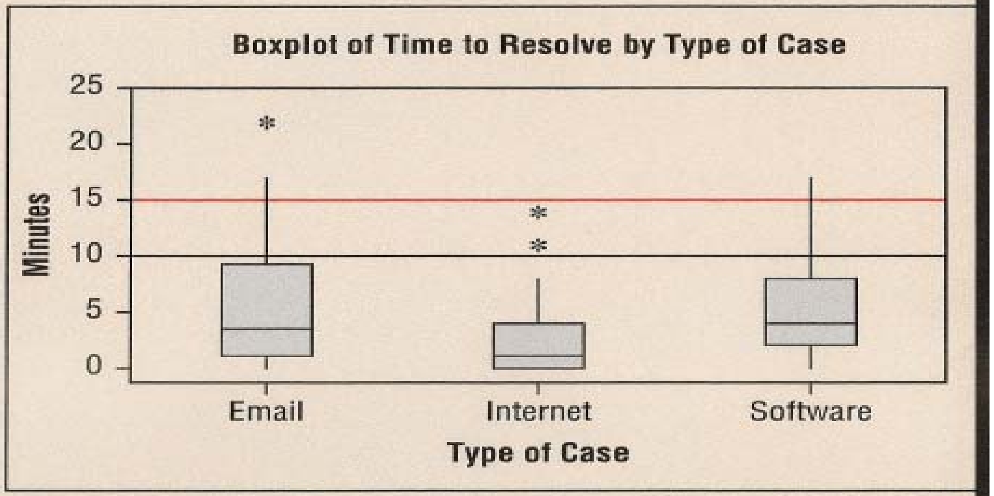
$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{-7.08}{(8.2192).8944)} = -.9631$$

Data Dashboards:
Adding Numerical Measures to Improve Effectiveness

- Data dashboards are not limited to graphical displays.
- The addition of numerical measures, such as the mean and standard deviation of KPIs, to a data dashboard is often critical.
- Dashboards are often interactive.
- Drilling down refers to functionality in interactive dashboards that allows the user to access information and analyses at an increasingly detailed level.

Data Dashboards:
Adding Numerical Measures to Improve Effectiveness





Summary Statistics - Resolved Cases

Type of Case	Cases	Mean	Median	Std. Dev
Email	34	4.6	2.0	5.6
Internet	19	5.4	3.0	4.9
Software	23	5.2	4.0	4.2

Hour	Cases	Mean	Median	Std. Dev
8:00	22	3.5	2.0	3.7
9:00	19	5.8	3.0	6.6
10:00	19	5.3	4.0	4.8
11:00	9	6.9	6.0	5.1
12:00	6	4.8	3.5	3.9