Chapter 3, Part A
Descriptive Statistics: Numerical Measures

- Measures of Location
- Measures of Variability

CENGAGE

Numerical Measures

- If the measures are computed for data from a sample, they are called sample statistics.
- If the measures are computed for data from a population, they are called population parameters.
- A sample statistic is referred to as the point estimator of the corresponding population parameter.

Measures of Location

- Mean
- Median
- Mode
- Weighted Mean
- Geometric Mean
- Percentiles
- Quartiles

Mean

Perhaps the most important measure of location is the mean.

- The mean provides a measure of central location.
- The mean of a data set is the average of all the data values.
- The sample mean $\bar{x}$ is the point estimator of the population mean μ.

Sample Mean $\bar{x}$

$$\bar{x} = \frac{\sum x_i}{n}$$

where:

$\sum x_i$ = sum of the values of the n observations

n = number of observations in the sample

Population Mean μ

$$\mu = \frac{\sum x_i}{n}$$

where:

$\sum x_i$ = sum of the values of the n observations

n = number of observations in the population

Sample Mean $\bar{x}$ (1 of 2)

Example: Monthly Starting Salary

A placement office wants to know the average starting salary of business graduates. Monthly starting salaries for a sample of 12 business school graduates is provided here.

Graduate	Monthly Starting Salary ($)	Graduate	Monthly Starting Salary ($)
1	5850	7	5890
2	5950	8	6130
3	6050	9	5940
4	5880	10	6325
5	5755	11	5920
6	5710	12	5880

Sample Mean $\bar{x}$ (2 of 2)

Example: Monthly Starting Salary

$$\bar{x} = \frac{\sum x_i}{n} = \frac{71{,}280}{12} = 5{,}940$$

Median (1 of 4)

- The median of a data set is the value in the middle when the data items are arranged in ascending order.
- Whenever a data set has extreme values, median is the preferred measure of central location.
- The median is the measure of location most often reported for annual income and property value data.
- A few extremely large incomes or property values can inflate the mean.

Median (2 of 4)

For an odd number of observations:

7 observations

26 18 27 12 14 27 19

12 14 18 19 26 27 27

In ascending order

Median is the middle value

Median = 19

Median (3 of 4)

For an even number of observations:

8 observations

26	18	27	12	14	27	19	30

12	14	18	19	26	27	27	30

In ascending order

Median is the average of the middle two values.

$$\text{Median} = \frac{(19 + 26)}{2} = 22.5$$

Median (4 of 4)

Example: Monthly Starting Salary
Averaging the 6th and 7th data values:

5,710	5,755
5,850	5,880
5,880	5,890
5,920	5,940
5,950	6,050
6,130	6,325

Note: The data is in ascending order.

$$\text{Median} = \frac{(5{,}890 + 5{,}920)}{2} = 5{,}905$$

Trimmed Mean

- Another measure sometimes used when extreme values are present is the trimmed mean.
- It is obtained by deleting a percentage of the smallest and largest values from a data set and then computing the mean of the remaining values.
- For example, the 5% trimmed mean is obtained by removing the smallest 5% and the largest 5% of the data values and then computing the mean of the remaining values.

Mystery about Mean-E(x)

- when the data is realistic and handable, why is the mean/E(x) the best estimate?

  We can prove doing a little math:

Mode (1 of 2)

- The mode of a data set is the value that occurs with greatest frequency.
- The greatest frequency can occur at two or more different values.
- If the data have exactly two modes, the data are bimodal.
- If the data have more than two modes, the data are multimodal.

Mode (2 of 2)

Example: Monthly Starting Salary

The only monthly starting salary that occurs more than once is $5,880.

Mode = 5,880

5,710	5,755
5,850	5,880
5,880	5,890
5,920	5,940
5,950	6,050
6,130	6,325

Note: The data is in ascending order.

Weighted Mean (1 of 4)

- In some instances the mean is computed by giving each observation a weight that reflects its relative importance.
- The choice of weights depends on the application.
- The weights might be the number of credit hours earned for each grade, as in GPA.
- In other weighted mean computations, quantities such as pounds, dollars, or volume are frequently used.

Weighted Mean (2 of 4)

$$\bar{x} = \frac{\sum w_i x_i}{\sum w_i}$$

where: x_i = value of observation i

w_i = weight for observation i

Numerator: sum of the weighted data values

Denominator: sum of the weights

If data is from a population, μ replaces $\bar{x}$.

Weighted Mean (3 of 4)

Example: Purchase of Raw Material

Consider the following sample of five purchases of a raw material over a period of three months:

Purchase	Cost per Pound ($)	Number of Pounds
1	3.00	1200
2	3.4	500
3	2.8	2750
4	2.9	1000
5	3.25	800

Weighted Mean (4 of 4)

Example: Purchase of raw material

Purchase	Cost per Pound ($) x_i	Number of Pounds w_i	$w_i x_i$
1	3.00	1200	3600
2	3.4	500	1700
3	2.8	2750	7700
4	2.9	1000	2900
5	255755	800	2600

$$\bar{x} = \frac{\sum w_i x_i}{\sum w_i} = \frac{18{,}500}{6{,}250} = 2.96 = \$2.96$$

FYI, equally weighted (simple) mean = $3.07

Geometric Mean (1 of 4)

- The geometric mean is calculated by finding the *n*th root of the product of *n* values.
- It is often used in analyzing growth rates in financial data (where using the arithmetic mean will provide misleading results).
- It should be applied anytime you want to determine the mean rate of change over several successive periods (be it years, quarters, weeks, . . .).
- Other common applications include changes in populations of species, crop yields, pollution levels, and birth and death rates.

Geometric Mean (2 of 4)

$$\bar{x}_g = \sqrt[n]{(x_1)(x_2)\ldots(x_n)}$$

$$= \left[(x_1)(x_2)\ldots(x_n)\right]^{1/n}$$

Geometric Mean (3 of 4)

Example: Mutual fund

Year	Annual Return %	Growth Factor
1	−22.1	0.779
2	28.7	1.287
3	10.9	1.109
4	4.9	1.049
5	15.8	1.158
6	5.5	1.055
7	−37.0	0.630
8	26.5	1.265
9	15.1	1.151
10	2.1	1.021

Geometric Mean (4 of 4)

$$\bar{x}_g = \sqrt[10]{(0.779)(1.287)(1.109)(1.049)(1.158)(1.055)(0.630)(1.265)(1.151)(1.021)}$$

$$= \sqrt[10]{1.334493}$$

$$= 1.029275$$

Average growth rate per period is $(1.029275 - 1)(100) = 2.9\%$

Percentiles (1 of 2)

- A percentile provides information about how the data are spread over the interval from the smallest value to the largest value.
- Admission test scores for colleges and universities are frequently reported in terms of percentiles.
- The *p*th percentile of a data set is a value such that at least *p*% of the items take on this value or less and at least (100 – *p*)% of the items take on this value or more.

Percentiles (2 of 2)

Arrange the data in ascending order.

Compute L_p, the location of the pth percentile.

$$L_p = \left(\frac{p}{100}\right)(n+1)$$

80th Percentile (1 of 2)

Example: Monthly Starting Salary

$$L_p = (p/100)(n+1) = (80/100)(12+1) = 10.4$$

(the 10^{th} value plus .4 times the difference between the 11^{th} and 10^{th} values)

$$80\text{th Percentile} = 6{,}050 + 0.4(6{,}130 - 6{,}050) = 6{,}082$$

5,710	5,755
5,850	5,880
5,880	5,890
5,920	5,940
5,950	6,050
6,130	6,325

80th Percentile (2 of 2)

Example: Monthly Starting Salary

At least 80% of the items take on a value of 6082 or less.
10/12 = .833 or 83%

At least 20% of the items take on a value of 6082 or more.
2/12 = .167 or 16.7%

5,710	5,755
5,850	5,880
5,880	5,890
5,920	5,940
5,950	6,050
6,130	6,325

Quartiles

Quartiles are specific percentiles.

- First Quartile = 25th Percentile
- Second Quartile = 50th Percentile = Median
- Third Quartile = 75th Percentile

Third Quartile (75th Percentile)

Example: Monthly Starting Salary

$$L_p = (p/100)(n+1) = (75/100)(12+1) = 9.75$$

(the 9th value plus .75 times the difference between the 10th and 9th values)

$$\text{Third quartile} = 5{,}950 + .75(6{,}050 - 5{,}950) = 6{,}025$$

5,710	5,755
5,850	5,880
5,880	5,890
5,920	5,940
5,950	6,050
6,130	6,325

Measures of Variability (1 of 2)

- It is often desirable to consider measures of variability (dispersion), as well as measures of location.
- For example, in choosing supplier A or supplier B we might consider not only the average delivery time for each but also the variability in delivery time for each.

Measures of Variability (2 of 2)

- **Range**
- Interquartile Range
- **Variance**
- **Standard Deviation**
- Coefficient of Variation

Range (1 of 2)

- The range of a data set is the difference between the largest and smallest data values.

$$\text{Range} = \text{Largest value} - \text{Smallest value}$$

- It is the simplest measure of variability.
- It is very sensitive to the smallest and largest data values.

Range (2 of 2)

Example: Monthly Starting Salary

$$\text{Range} = \text{largest value} - \text{smallest value}$$

$$\text{Range} = 6{,}325 - 5{,}710 = 615$$

5,710	5,755
5,850	5,880
5,880	5,890
5,920	5,940
5,950	6,050
6,130	6,325

Interquartile Range

- The interquartile range of a data set is the difference between the third quartile and the first quartile.
- It is the range for the middle 50% of the data.
- It overcomes the sensitivity to extreme data values.

Interquartile Range (IQR)

Example: Monthly Starting Salary

- 3rd Quartile (Q_3) = 6,000
- 1st Quartile (Q_1) = 5,865
- IQR = $Q_3 - Q_1$ = 6,000 – 5,865 = 135

5,710	5,755
5,850	5,880
5,880	5,890
5,920	5,940
5,950	6,050
6,130	6,325

Variance (1 of 2)

- The variance is a measure of variability that utilizes all the data.
- It is based on the difference between the value of each observation (x_i) and the mean ($\bar{x}$ for a sample, μ for a population).
- The variance is useful in comparing the variability of two or more variables.

Variance (2 of 2)

- The variance is the average of the squared differences between each data value and the mean.
- The variance is computed as follows:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

for a sample

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

for a population

Variance (General)

- The variance is the average of the squared differences between each data value and the mean.
- Please think what if the sample size is given by frequency:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

still work?

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

for a population

Standard Deviation (1 of 2)

- The standard deviation of a data set is the positive square root of the variance.
- It is measured in the same units as the data, making it more easily interpreted than the variance.

Standard Deviation (2 of 2)

The standard deviation is computed as follows:

For a sample	For a population
$s = \sqrt{s^2}$	$\sigma = \sqrt{\sigma^2}$

Coefficient of Variation

- The coefficient of variation indicates how large the standard deviation is in relation to the mean.

The coefficient of variation is computed as follows:

$$\left[\frac{s}{\bar{x}} \times 100\right]\% \qquad \left[\frac{\sigma}{\mu} \times 100\right]\%$$

for a sample

for a population

Sample Variance, Standard Deviation, And Coefficient of Variation

Example: Monthly Starting Salary

Variance

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = 27,440.91$$

Standard Deviation

$$s = \sqrt{s^2} = \sqrt{27,440.91} = 165.65$$

Coefficient of Variation

$$\left[\frac{s}{\bar{x}} \times 100\right]\% = \left[\frac{165.65}{3,940} \times 100\right]\% = 4.2\%$$

Chapter 3, Part B
Descriptive Statistics: Numerical Measures

- Measures of Distribution Shape, Relative Location, and Detecting Outliers.
- Five-Number Summaries and Boxplots
- 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
- Detecting Outliers

Distribution Shape: Skewness (1 of 5)

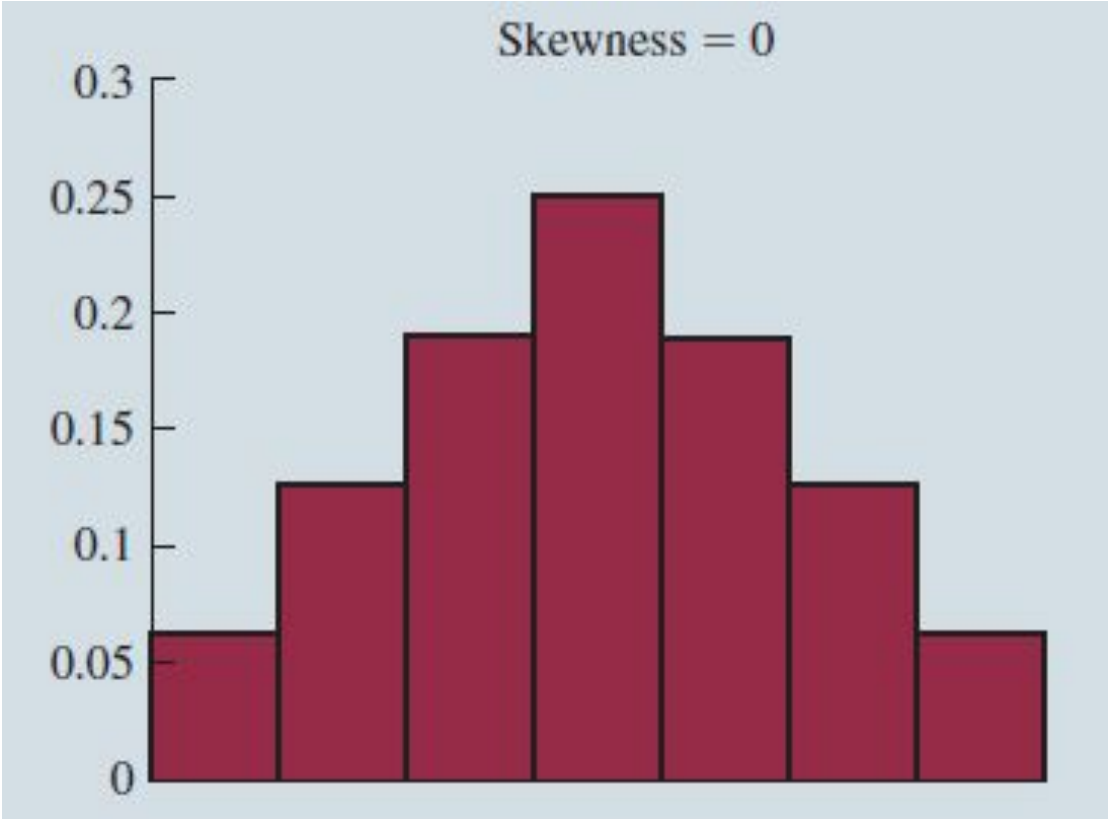
- An important measure of the shape of a distribution is called skewness.
- The formula for the skewness of sample data is

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

- Skewness can be easily computed using statistical software.

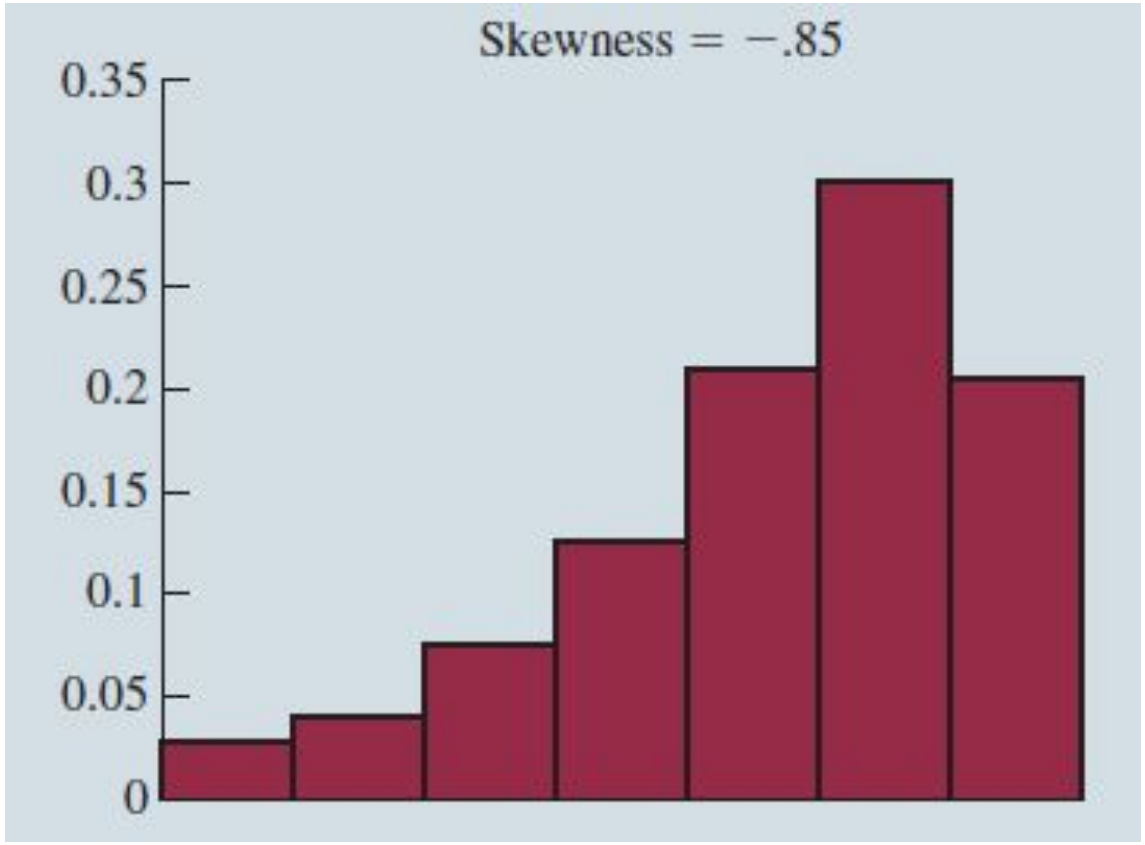
Distribution Shape: Skewness (2 of 5)

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



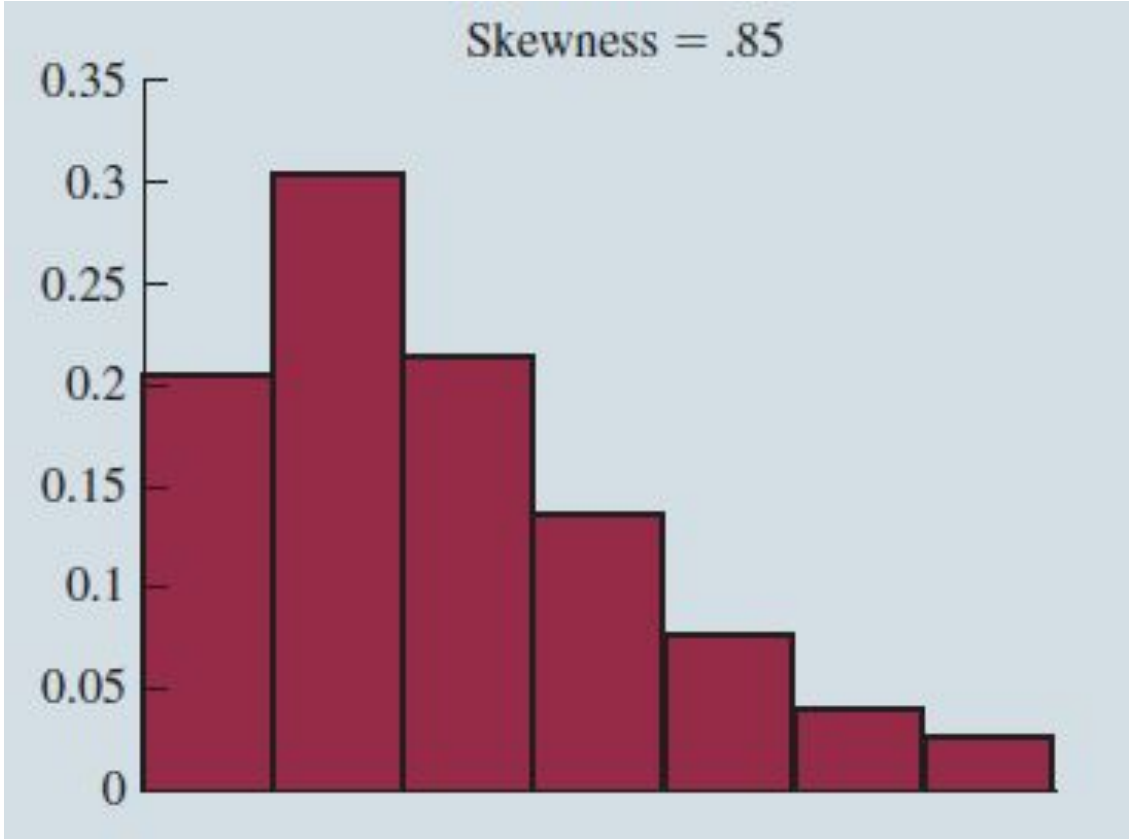
Distribution Shape: Skewness (3 of 5)

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



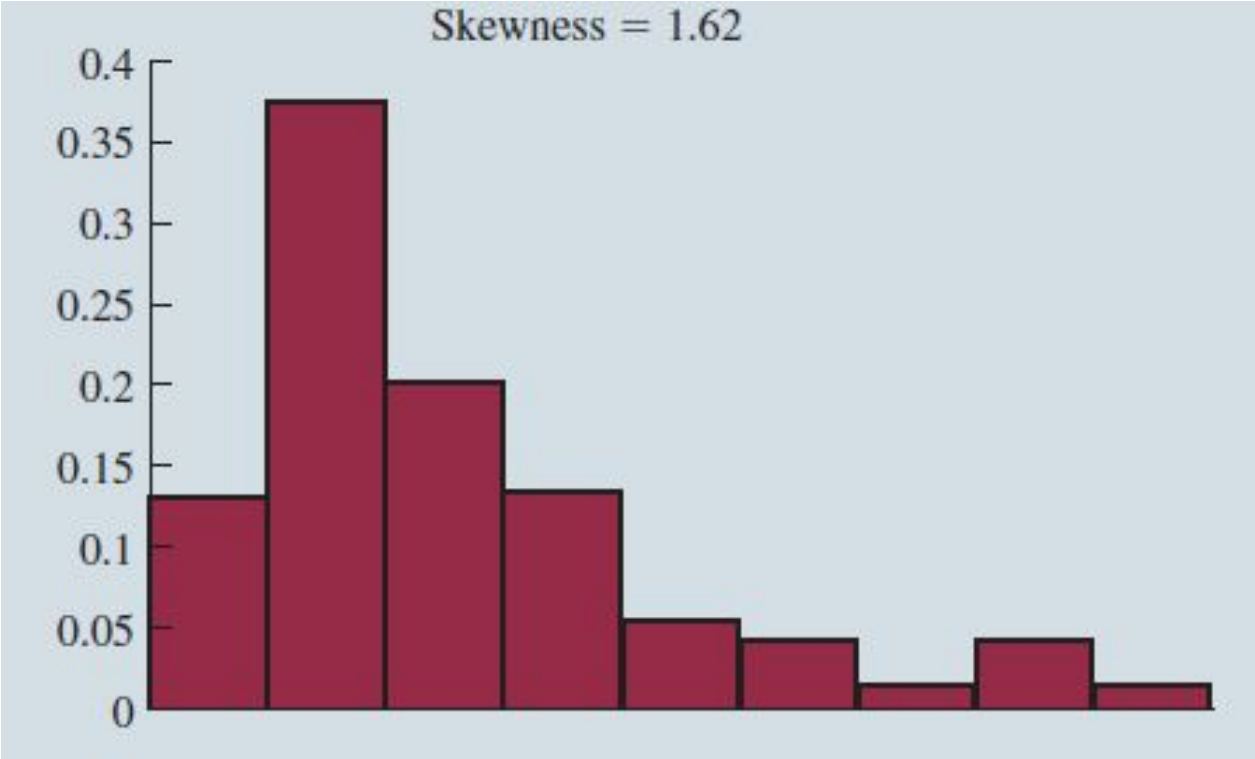
Distribution Shape: Skewness (4 of 5)

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



Distribution Shape: Skewness (5 of 5)

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



z-Scores (1 of 3)

- 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.

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

- Excel's STANDARDIZE function can be used to compute the z-score.

z-Scores (2 of 3)

- 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 (3 of 3)

Example: Class size data

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

Number of students In class	Deviation about the Mean	Z score $\left(\frac{x_i - \bar{x}}{s}\right)$
46	2	2/8 = 0.25
54	10	10/8 = 1.25
42	−2	−2/8 = −0.25
46	2	2/8 = 0.25
32	−12	−12/8 = −1.5

Note: $\bar{x} = 44$ and $s = 8$ for the given data.

CENGAGE

Chebyshev's Theorem (1 of 4)

- At least $(1 - 1/z^2)$ of the items in any data set will 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.

Chebyshev's Theorem (2 of 4)

- 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 (3 of 4)

Example: Midterm scores of students

Suppose the midterm test scores of 100 students in a course had a mean of 70 and a standard deviation of 5. We want to know the number of students having test scores between 60 and 80.

60 and 80 are 2 standard deviations below and above the mean respectively.

$$60 = 70 - 2(5) \longrightarrow s$$

$$80 = 70 + 2(5)$$

$$Z = 75\%$$

Chebyshev's Theorem (4 of 4)

Example: Midterm scores of students

Number of students having test scores between 58 and 82:

$$(58 - 70)/5 = -2.4$$

$$(82 - 70)/5 = 2.4$$

$$z = 2.4$$

$$(1 - 1/z^2) = (1 - 1/(2.4)^2) = 0.826 = 82.6\%$$

Empirical Rule (1 of 3)

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.

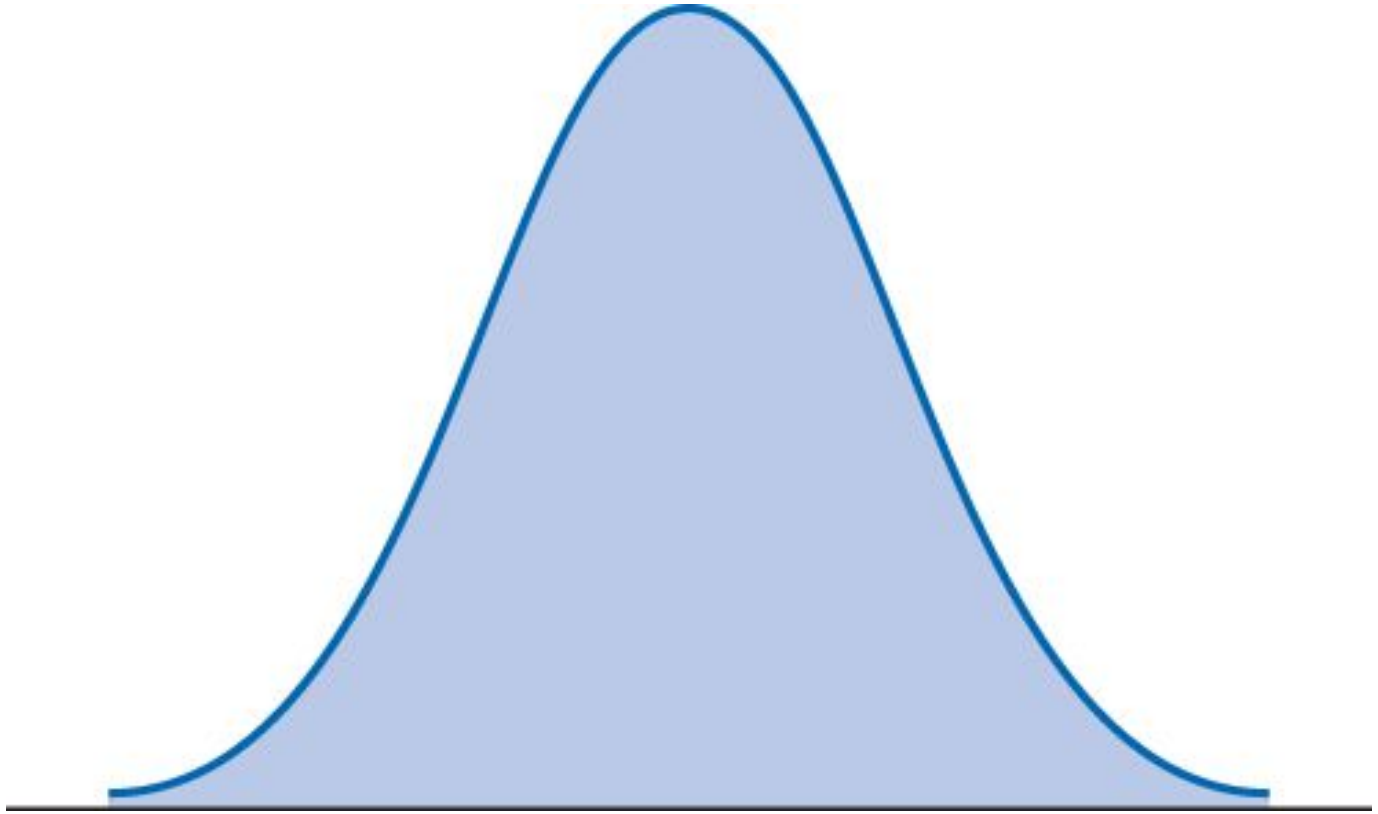
Empirical Rule (2 of 3)

For data having a bell-shaped distribution:

- Approximately 68% of the data values will be within +/−1 standard deviation of its mean.
- Approximately 95% of the data values will be within +/ −2 standard deviations of its mean.
- Almost all (approximately 99.7%) of the data values will be within +/ −3 standard deviations of its mean.

Empirical Rule (3 of 3)

- Bell shaped distribution



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 unusual data value that belongs in the data set

Outliers

Example: Class size data

Number of students In class	Deviation about the Mean	Z score $\left(\frac{x_i - \bar{x}}{s}\right)$
46	2	2/8 = 0.25
54	10	10/8 = 1.25
42	−2	−2/8 = −0.25
46	2	2/8 = 0.25
32	−12	−12/8 = −1.5

- −1.5 shows the fifth class size is farthest from the mean.
- No outliers are present as the z values are within the +/−3 guideline.

Five-Number Summaries and Boxplots

- 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 boxplots.

Five-Number Summary (1 of 2)

- Smallest Value
- First Quartile
- Median
- Third Quartile
- Largest Value

Five-Number Summary (2 of 2)

Example: Monthly starting salary

- Lowest Value = 5,710
- Third Quartile = 6,025
- Median = 5905
- First Quartile = 5,857.5
- Largest Value = 6,325

Monthly Starting Salary ($)	
5,710	5,755
5,850	5,880
5,880	5,890
5,920	5,940
5,950	6,050
6,130	6,325

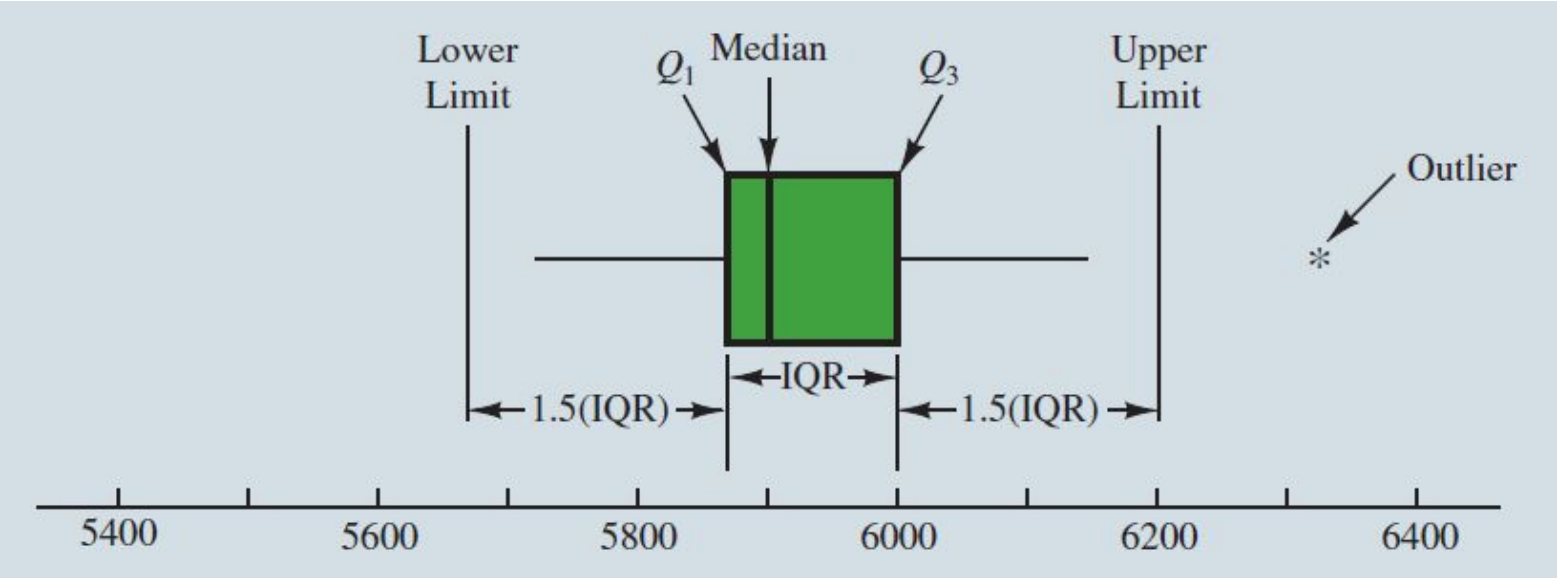
Boxplot (1 of 2)

- A boxplot is a graphical summary of data that is based on a five-number summary.
- A key to the development of a boxplot is the computation of the median and the quartiles, Q_1 and Q_3.
- Boxplots provide another way to identify outliers.

Boxplot (2 of 3)

Example Monthly starting salary

- 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).



Boxplots and Outliers

- Limits are located using the interquartile range (IQR).
- Data outside these limits are considered outliers.
- The locations of each outlier are shown with the symbol.

Boxplot (3 of 3)

Example: Monthly starting salary

- The lower limit is located 1.5(IQR) below Q_1.

$$\text{Lower Limit: } Q_1 - 1.5(\text{IQR}) = 5{,}857.5 - 1.5(167.5) = 5{,}606.25$$

- The upper limit is located 1.5(IQR) above Q_3.

$$\text{Upper Limit: } Q_3 + 1.5(\text{IQR}) = 6{,}025 + 1.5(167.5) = 6{,}276.25$$

- There is one outlier: 6,325.

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.

Covariance (1 of 2)

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

Covariance (2 of 2)

The covariance is computed as follows:

For samples:

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

For population:

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

Correlation Coefficient (1 of 3)

- 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.

Correlation Coefficient (2 of 3)

- The correlation coefficient is computed as follows:

For samples:

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

For population:

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Correlation Coefficient (3 of 3)

- 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 (1 of 4)

Example: San Francisco Electronics store

The store's manager wants to determine the relationship between the number of weekend television commercials shown and the sales at the store during the following week.

Covariance and Correlation Coefficient (2 of 4)

Example: San Francisco Electronics Store

Week	Number of Commercials	Sales ($100s)
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	49

Covariance and Correlation Coefficient (3 of 4)

	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

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

Covariance and Correlation Coefficient (4 of 4)

Example: San Francisco Electronics Store

Sample Covariance

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1} = 99/9 = 11$$

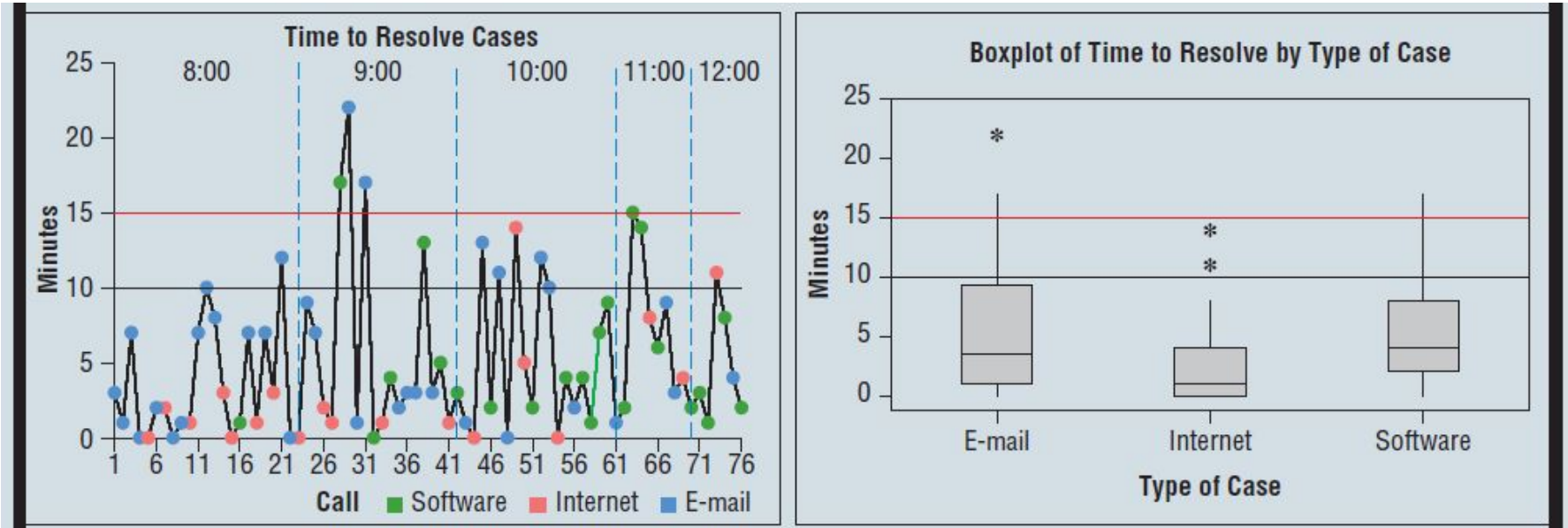
Sample Correlation Coefficient

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = 11/(1.49 \times 7.93) = 0.93$$

Data Dashboards: Adding Numerical Measures to Improve Effectiveness (1 of 2)

- 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 increasingly detailed level.

Data Dashboards: Adding Numerical Measures to Improve Effectiveness (2 of 2)



Summary Statistics—Resolved Cases

Type of Case	Cases	Mean	Median	Std. Dev
E-mail	34	5.8	3.5	5.4
Internet	19	3.0	1.0	4.0
Software	23	5.4	4.0	5.0

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