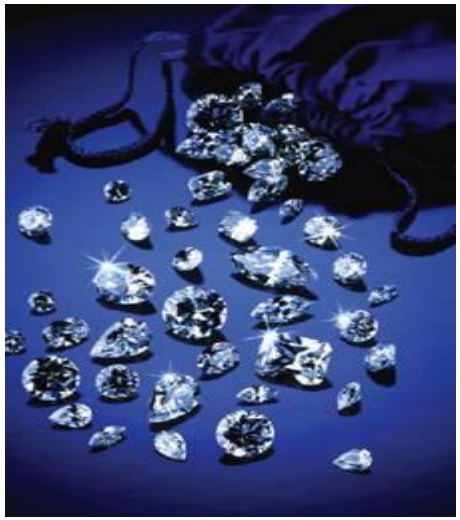
Describing Data:
Displaying and Exploring Data

Chapter 04



LEARNING OBJECTIVES

LO 4-1 Construct and interpret a *dot plot*.
LO 4-2 Identify and compute measures of position.
LO 4-3 Construct and analyze a *box plot*.
LO 4-4 Compute and describe the *coefficient of skewness*.
LO 4-5 Create and interpret a scatter diagram.
LO 4-6 Develop and explain a *contingency table*.

LO 4-1 Construct and interpret a *dot plot*.

Dot Plots

- A **dot plot** groups the data as little as possible and the identity of an individual observation is not lost.
- To develop a dot plot, each observation is simply displayed as a dot along a horizontal number line indicating the possible values of the data.
- If there are identical observations or the observations are too close to be shown individually, the dots are “piled” on top of each other.

Dot Plots – Examples

The Service Departments at Tionesta Ford Lincoln Mercury and Sheffield Motors, Inc., two of the four Applewood Auto Group Dealerships, were both open 24 days last month. Listed below is the number of vehicles serviced during the 24 working days at the two dealerships. Construct dot plots and report summary statistics to compare the two dealerships.

Tionesta Ford Lincoln Mercury

Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
23	33	27	28	39	26
30	32	28	33	35	32
29	25	36	31	32	27
35	32	35	37	36	30

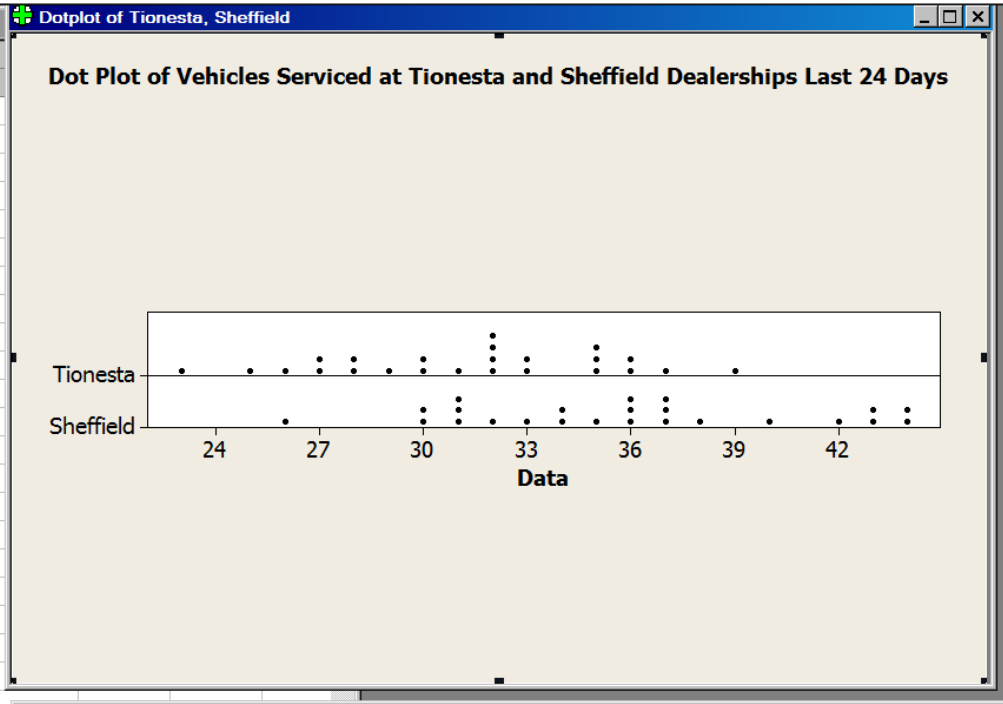
Sheffield Motors Inc.

Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
31	35	44	36	34	37
30	37	43	31	40	31
32	44	36	34	43	36

Dot Plot – Minitab Example

DOTPLOTS.MTW ***

	C1	C2
	Tionesta	Sheffield
1	23	31
2	27	44
3	30	30
4	27	36
5	32	37
6	31	34
7	32	43
8	32	38
9	35	37
10	33	35
11	28	36
12	39	34
13	32	31
14	29	32
15	35	40
16	36	36
17	33	31
18	25	44
19	35	26
20	37	30
21	26	37
22	28	43
23	36	42
24	30	33
25		
26		
27		
28		
29		
30		
31		



Session

Descriptive Statistics: Tionesta, Sheffield

Variable	N	Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Tionesta	24	31.292	4.112	23.000	28.000	32.000	35.000	39.000
Sheffield	24	35.83	4.96	26.00	31.25	36.00	39.50	44.00

LO 4-2 Identify and compute measures of position.

Measures of Position

- The standard deviation is the most widely used measure of dispersion.
- Alternative ways of describing spread of data include determining the *location* of values that divide a set of observations into equal parts.

LOCATION OF A PERCENTILE

$$L_p = (n + 1)\frac{P}{100} \qquad [4–1]$$

- These measures include **quartiles, deciles,** and **percentiles.**

Percentile Computation

- To formalize the computational procedure, let L_p refer to the location of a desired percentile. So if we wanted to find the 33rd percentile, we would use L_{33}, and if we wanted the median, the 50th percentile, then L_{50}.

LOCATION OF A PERCENTILE

$$L_p = (n + 1)\frac{P}{100} \quad [4\text{–}1]$$

- The number of observations is *n,* so if we want to locate the median, its position is at $(n + 1)/2$, or we could write this as $(n + 1)(P/100)$, where P is the desired percentile.

Percentiles – Example

Listed below are the commissions earned last month by a sample of 15 brokers at Salomon Smith Barney's Oakland, California, office.

$2,038	$1,758	$1,721	$1,637
$2,097	$2,047	$2,205	$1,787
$2,287	$1,940	$2,311	$2,054
$2,406	$1,471	$1,460	

Locate the median, the first quartile, and the third quartile for the commissions earned.

Percentiles – Example (cont.)

Step 1: Organize the data from lowest to largest value.

$1,460	$1,471	$1,637	$1,721
$1,758	$1,787	$1,940	$2,038
$2,047	$2,054	$2,097	$2,205
$2,287	$2,311	$2,406	

Percentiles – Example (cont.)

Step 2: Compute the first and third quartiles.

Locate L_{25} and L_{75} using:

LOCATION OF A PERCENTILE $$L_p = (n + 1)\frac{P}{100}$$ **[4–1]**

$$L_{25} = (15+1)\frac{25}{100} = 4 \qquad L_{75} = (15+1)\frac{75}{100} = 12$$

Therefore, the first and third quartiles are located at the 4th and 12th positions, respectively:

$L_{25} = \$1{,}721$

$L_{75} = \$2{,}205$

Percentiles – Example (cont.)

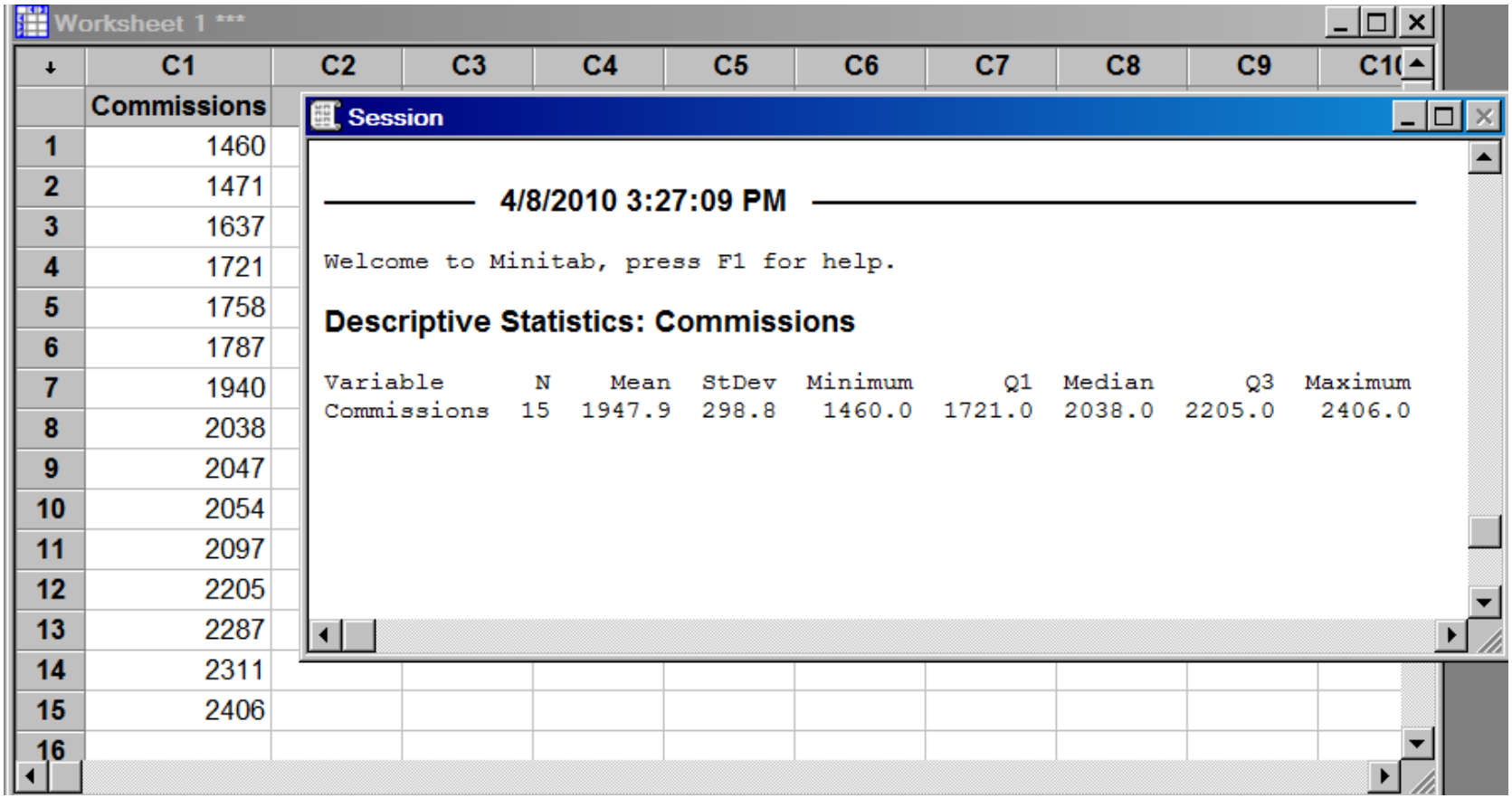
In the previous example, the location formula yielded a whole number. What if there were 6 observations in the sample with the following ordered observations: 43, 61, 75, 91, 101, and 104 , that is *n=6, and we wanted to locate the first quartile?*

$$L_{25} = (6+1)\frac{25}{100} = 1.75$$

Locate the first value in the ordered array and then move .75 of the distance between the first and second values and report that as the first quartile. Like the median, the quartile does not need to be one of the actual values in the data set.

The 1st and 2nd values are 43 and 61. Moving 0.75 of the distance between these numbers, the 25th percentile is **56.5**, obtained as **43 + 0.75*(61 – 43)**

Percentiles – Example (Minitab)



LO 4-3 Construct and analyze a *box plot*.

Box Plot

- A *box plot* is a graphical display, based on quartiles, that helps us picture a set of data.
- To construct a box plot, we need only five statistics:
  - The minimum value
  - *Q1* (the first quartile)
  - The median
  - *Q3 (the third quartile)*
  - *The maximum value.*

Box Plot – Example

- Alexander's Pizza offers free delivery of its pizza within 15 miles. Alex, the owner, wants some information on the time it takes for delivery. How long does a typical delivery take? Within what range of times will most deliveries be completed? For a sample of 20 deliveries, he determined the following information:
  - Minimum value = 13 minutes
  - Q1 = 15 minutes
  - Median = 18 minutes
  - Q3 = 22 minutes
  - Maximum value = 30 minutes

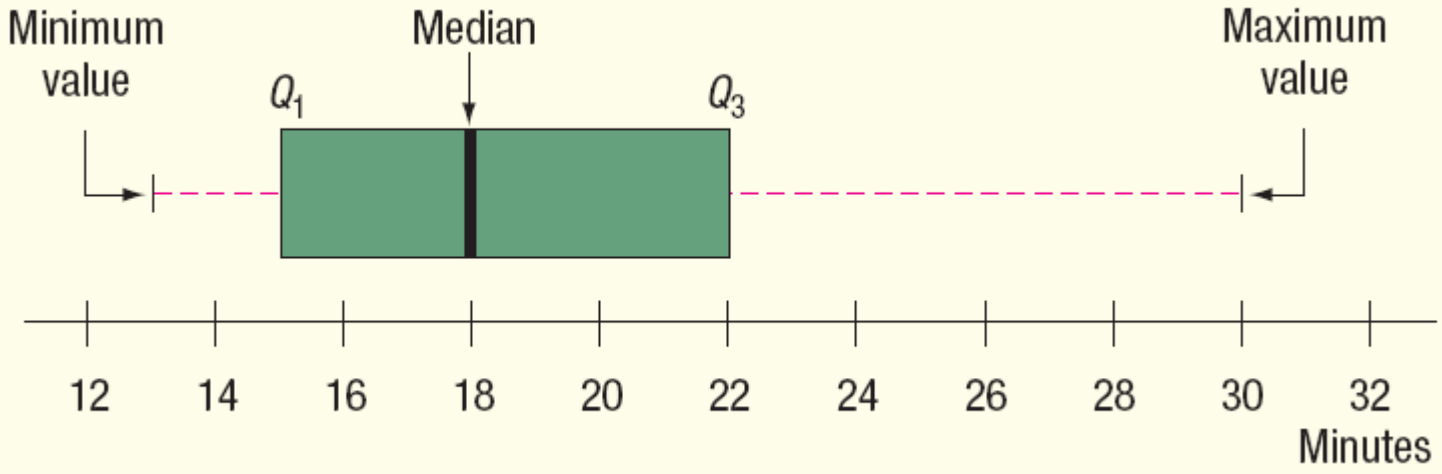
- Develop a box plot for the delivery times. What conclusions can you make about the delivery times?

Box Plot – Example

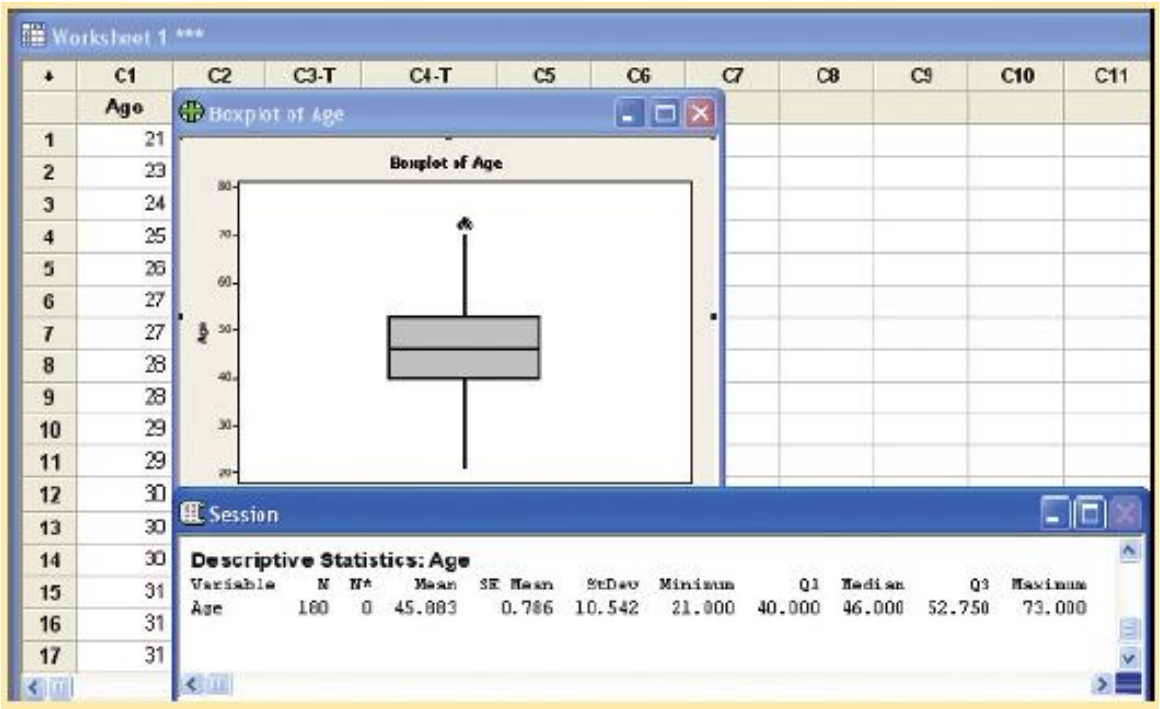
Step 1: Create an appropriate scale along the horizontal axis.

Step 2: Draw a box that starts at *Q1 (15 minutes) and ends at Q3 (22* minutes). Inside the box we place a vertical line to represent the median (18 minutes).

Step 3: Extend horizontal lines from the box out to the minimum value (13 minutes) and the maximum value (30 minutes).



Box Plot – Using Minitab



Refer to the Applewood Auto Group data. Develop a box plot for the variable age of the buyer. What can we conclude about the distribution of the age of the buyer?

The Minitab statistical software system was used to develop the following chart and summary statistics. What can we conclude about the distribution of the age of the buyers?

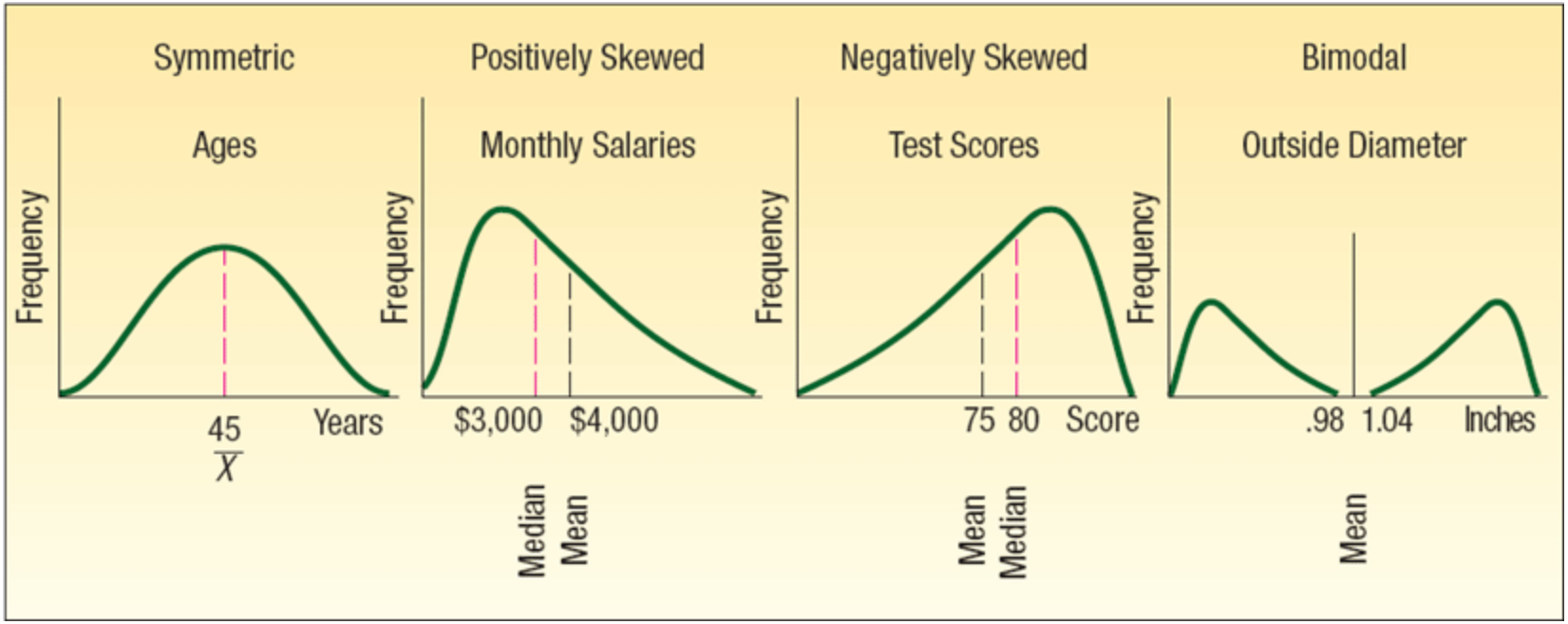
- The median age of purchaser was 46 yrs.
- 25 percent were more than 52.75 years of age.
- 50 percent of the purchasers were between the ages of 40 and 52.75 years.
- The distribution of age is symmetric.

LO 4-4 Compute and understand the *coefficient of skewness*.

Skewness

- In Chapter 3, measures of central location (the mean, median, and mode) for a set of observations and measures of data dispersion (e.g., range and the standard deviation) were introduced.
- Another characteristic of a set of data is the shape.
- There are four shapes commonly observed:
  - Symmetric
  - Positively skewed
  - Negatively skewed
  - Bimodal

Commonly Observed Shapes



Skewness – Formulas for Computing

The coefficient of skewness can range from –3 up to 3.

- A value near –3, indicates considerable negative skewness.
- A value such as 1.63 indicates moderate positive skewness.
- A value of 0, which will occur when the mean and median are equal, indicates the distribution is symmetrical and that there is no skewness present.

PEARSON'S COEFFICIENT OF SKEWNESS

$$sk = \frac{3(\bar{X} - \text{Median})}{s} \qquad [4\text{–}2]$$

SOFTWARE COEFFICIENT OF SKEWNESS

$$sk = \frac{n}{(n-1)(n-2)}\left[\sum\left(\frac{X - \bar{X}}{s}\right)^3\right] \qquad [4\text{–}3]$$

Skewness – An Example

Following are the earnings per share for a sample of 15 software companies for the year 2010. The earnings per share are arranged from smallest to largest.

$0.09	$0.13	$0.41	$0.51	$ 1.12	$ 1.20	$ 1.49	$3.18
3.50	6.36	7.83	8.92	10.13	12.99	16.40	

- Compute the mean, median, and standard deviation. Find the coefficient of skewness using Pearson's estimate.
- What is your conclusion regarding the shape of the distribution?

Skewness – An Example Using Pearson's Coefficient

Step 1: Compute the Mean

$$\overline{X} = \frac{\sum X}{n} = \frac{\$74.26}{15} = \$4.95$$

Step 2 : Compute the Standard Deviation

$$s = \sqrt{\frac{\Sigma\left(X - \overline{X}\right)^2}{n-1}} = \sqrt{\frac{(\$0.09 - \$4.95)^2 + \ldots + (\$16.40 - \$4.95)^2)}{15-1}} = \$5.22$$

Step 3 : Find the Median

The middle value in the set of data, arranged from smallest to largest is 3.18

Step 4 : Compute the Skewness

$$sk = \frac{3(\overline{X} - Median)}{s} = \frac{3(\$4.95 - \$3.18)}{\$5.22} = 1.017$$

Skewness – An Example Using the Software Method

Step 1: Compute the Mean

$$\overline{X} = \frac{\sum X}{n} = \frac{\$74.26}{15} = \$4.95$$

Step 2: Compute the Standard Deviation

$$s = \sqrt{\frac{\Sigma\left(X - \overline{X}\right)^2}{n-1}} = \sqrt{\frac{(\$0.09 - \$4.95)^2 + \ldots + (\$16.40 - \$4.95)^2)}{15-1}} = \$5.22$$

Step 3: Compute $\sum\left(\frac{x - \overline{X}}{s}\right)^3$

Skewness – An Example Using the Software Method

Step 3: Compute $\sum\left(\frac{x-\bar{X})}{s}\right)^3$

Earnings per Share	$\frac{(X-\bar{X})}{s}$	$\left(\frac{X-\bar{X}}{s}\right)^3$
0.09	−0.9310	−0.8070
0.13	−0.9234	−0.7873
0.41	−0.8697	−0.6579
0.51	−0.8506	−0.6154
1.12	−0.7337	−0.3950
1.20	−0.7184	−0.3708
1.49	−0.6628	−0.2912
3.18	−0.3391	−0.0390
3.50	−0.2778	−0.0214
6.36	0.2701	0.0197
7.83	0.5517	0.1679
8.92	0.7605	0.4399
10.13	0.9923	0.9772
12.99	1.5402	3.6539
16.40	2.1935	10.5537
		11.8274

Skewness – An Example Using the Software Method

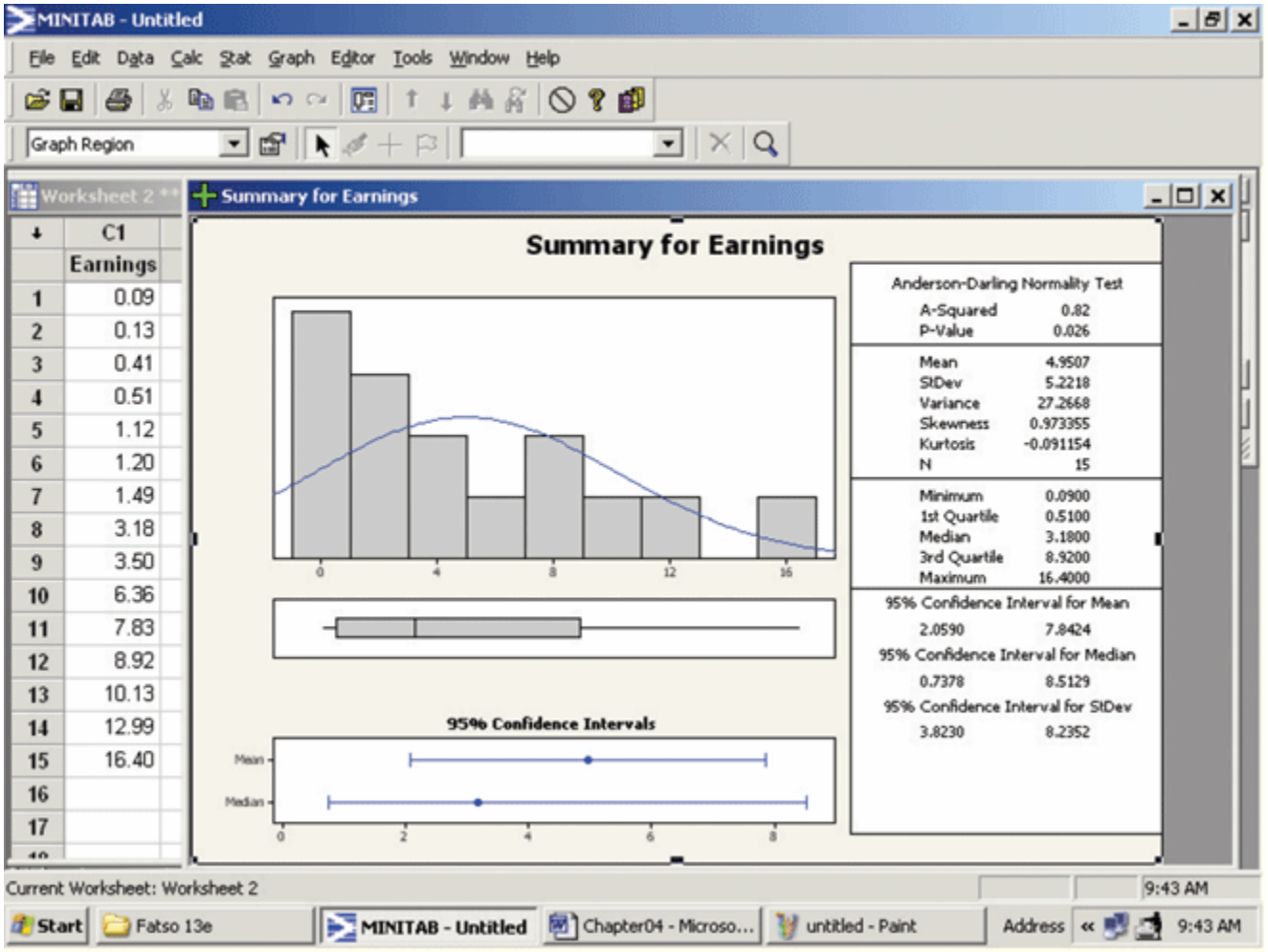
Step 4: Compute sk

$$= \frac{n}{(n-1)(n-2)} \sum \left(\frac{x - \overline{X}}{s} \right)^3$$

$$= \frac{15}{(15-1)(15-2)} (11.8274)$$

$$= 0.975$$

Skewness – A Minitab Example



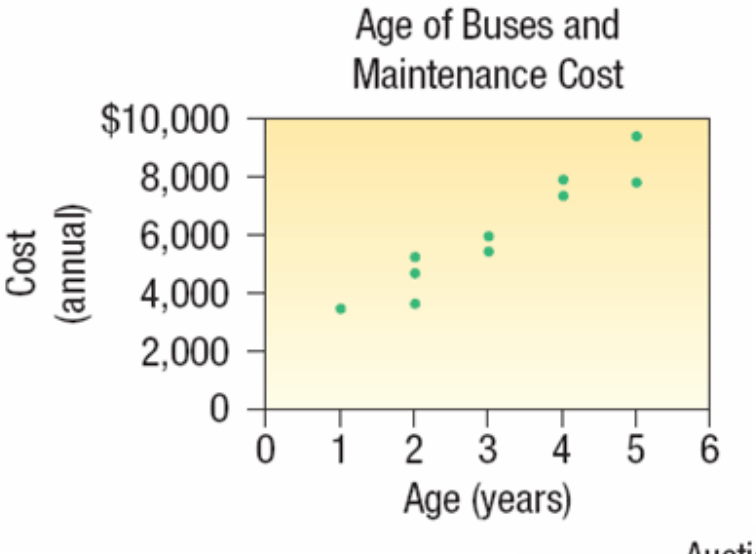
LO 4-5 Create and interpret a scatterplot.

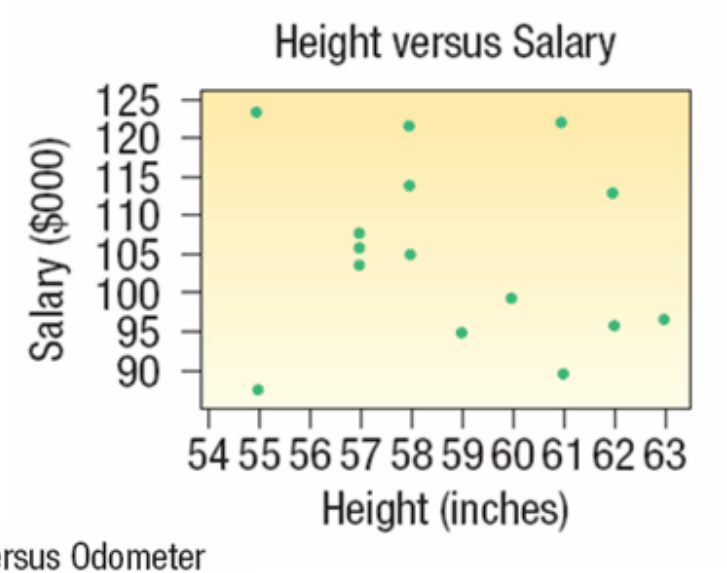
Describing a Relationship between Two Variables

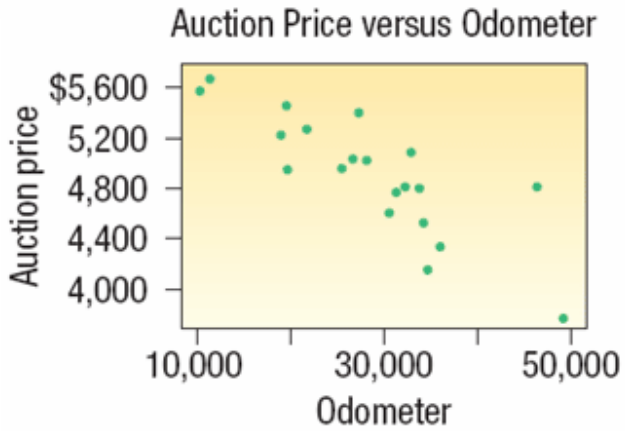
- When we study the relationship between two variables we refer to the data as **bivariate.**
- One graphical technique we use to show the relationship between variables is called a **scatter diagram.**
- To draw a scatter diagram, we need two variables. We scale one variable along the horizontal axis (*X*-axis) of a graph and the other variable along the vertical axis (*Y*-axis).

LO 4-5

Describing a Relationship between Two Variables – Scatter Diagram Examples







Describing a Relationship between Two Variables – Scatter Diagram Excel Example

In the Introduction to Chapter 2, we presented data from the Applewood Auto Group. We gathered information concerning several variables, including the profit earned from the sale of 180 vehicles sold last month. In addition to the amount of profit on each sale, one of the other variables is the age of the purchaser.

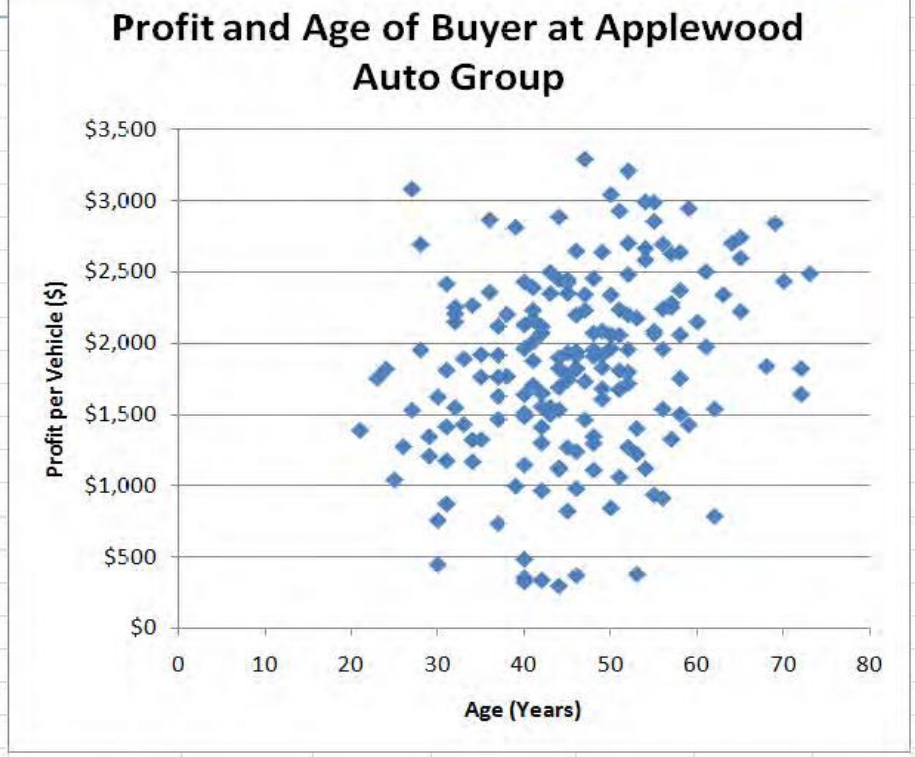
Is there a relationship between the profit earned on a vehicle sale and the age of the purchaser?

Would it be reasonable to conclude that the more expensive vehicles are purchased by older buyers?

Describing the Relationship between Two Variables – Scatter Diagram Excel Example

U27

	A	B
1	Age	Profit
2	21	$1,387
3	23	$1,754
4	24	$1,817
5	25	$1,040
6	26	$1,273
7	27	$1,529
8	27	$3,082
9	28	$1,951
10	28	$2,692
11	29	$1,206
12	29	$1,342
13	30	$443
14	30	$754
15	30	$1,621
16	31	$870
17	31	$1,174
18	31	$1,412
19	31	$1,809
20	31	$2,415
21	32	$1,546
22	32	$2,148
23	32	$2,207
24	32	$2,252
25	33	$1,428



LO 4-6 Develop and explain a *contingency table*.

Contingency Tables

- A scatter diagram requires that both of the variables be at least interval scale.
- What if we wish to study the relationship between two variables when one or both are nominal or ordinal scale? In this case, we tally the results in a **contingency table.**

CONTINGENCY TABLE A table used to classify observations according to two identifiable characteristics.

Contingency Tables

A contingency table is a cross-tabulation that simultaneously summarizes two variables of interest.

Examples:

1. Students at a university are classified by gender and class rank.
2. A product is classified as acceptable or unacceptable and by the shift (day, afternoon, or night) on which it is manufactured.
3. A voter in a school bond referendum is classified as to party affiliation (Democrat, Republican, other) and the number of children that voter has attending school in the district (0, 1, 2, etc.).

Contingency Tables – An Example

There are four dealerships in the Applewood Auto Group. Suppose we want to compare the profit earned on each vehicle sold by the particular dealership. To put it another way, is there a relationship between the amount of profit earned and the dealership? The table below is the cross-tabulation of the raw data of the two variables.

Contingency Table Showing the Relationship between Profit and Dealership

Above/Below Median Profit	Kane	Olean	Sheffield	Tionesta	Total
Above	25	20	19	26	90
Below	27	20	26	17	90
Total	52	40	45	43	180

From the contingency table, we observe the following:

1. From the Total column on the right, 90 of the 180 cars sold had a profit above the median and half below. From the definition of the median, this is expected.
2. For the Kane dealership, 25 out of the 52, or 48 percent, of the cars sold were sold for a profit more than the median.
3. The percent profits above the median for the other dealerships are 50 percent for Olean, 42 percent for Sheffield, and 60 percent for Tionesta.