



Describing Data: Displaying and Exploring Data



Chapter 4

Learning Objectives

- LO4-1 Construct and interpret a dot plot
- LO4-2 Construct and describe a stem-and-leaf display
- LO4-3 Identify and compute measures of position
- LO4-4 Construct and analyze a box plot
- LO4-5 Compute and interpret the coefficient of skewness
- LO4-6 Create and interpret a scatter diagram
- LO4-7 Develop and explain a contingency table

Dot Plots Example

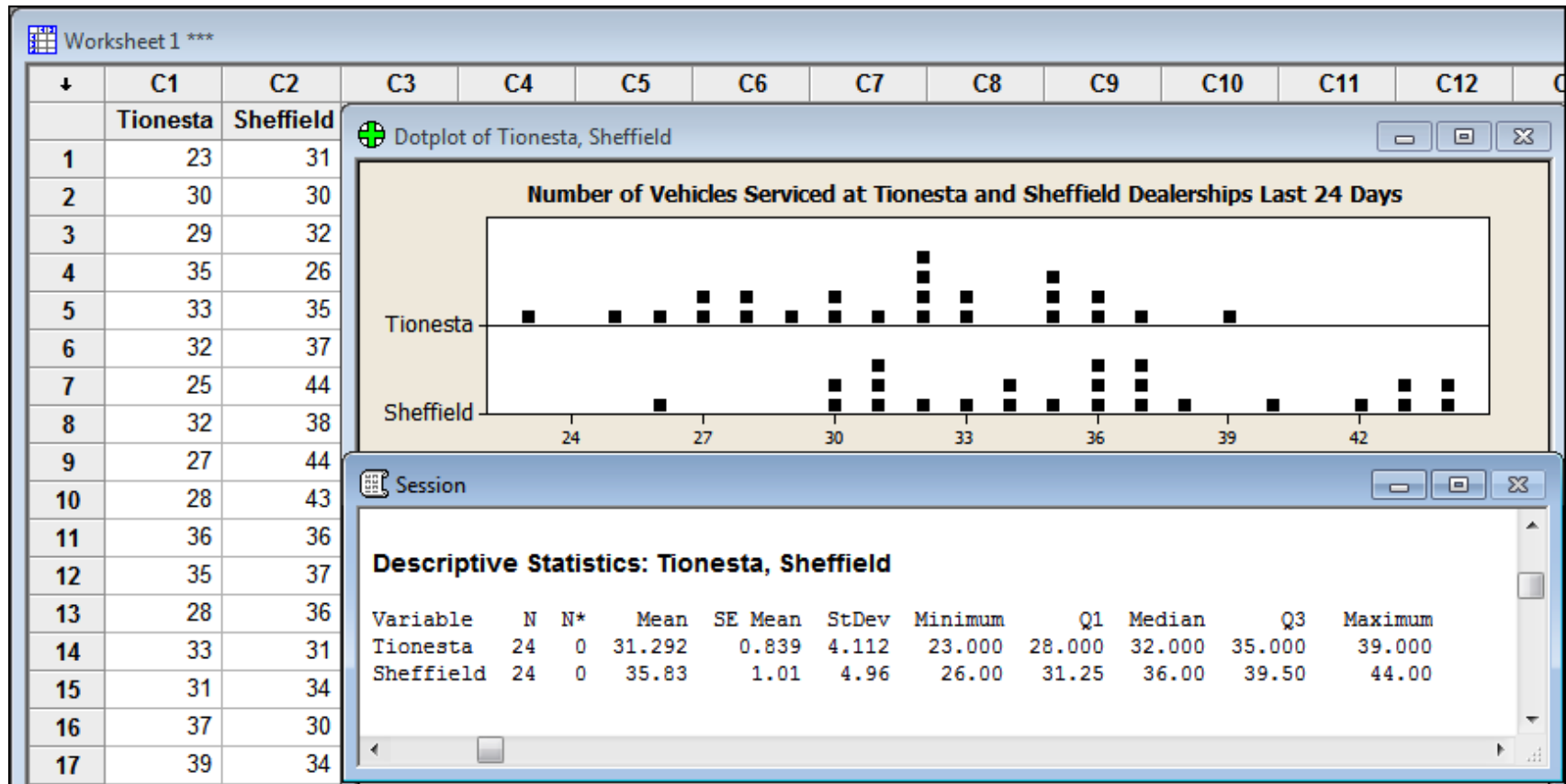
- ▶ Use dot plots to compare the two data sets like these of the number of vehicles serviced last month for two different dealerships

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
26	38	37	30	42	33

Dot Plots Example

- Minitab provides dot plots and summary statistics



Stem-and-Leaf Displays

- ▶ An alternative to a frequency distribution and histogram
- ▶ The advantages of the stem-and-leaf display
 - ▶ The identity of the each observation is not lost
 - ▶ The digits themselves give a picture of the distribution
 - ▶ The cumulative frequencies are also shown

STEM-AND-LEAF DISPLAY A statistical technique to present a set of data. Each numerical value is divided into two parts. The leading digit becomes the stem and the trailing digit the leaf. The stems are located along the vertical axis, and the leaf values are stacked against each other along the horizontal axis.

Stem-and-Leaf Display Example

TABLE 4-1 Number of People Attending Each of the 45 Performances at the Theater of the Republic

96	93	88	117	127	95	113	96	108	94	148	156
139	142	94	107	125	155	155	103	112	127	117	120
112	135	132	111	125	104	106	139	134	119	97	89
118	136	125	143	120	103	113	124	138			

Stem	Leaf
8	8 9
9	3 4 4 5 6 6 7
10	3 3 4 6 7 8
11	1 2 2 3 3 7 7 8 9
12	0 0 4 5 5 5 7 7
13	2 4 5 6 8 9 9
14	2 3 8
15	5 5 6

Measures of Position

- ▶ Measures of location also describes the shape of the distribution and can be expressed as percentiles

LOCATION OF A PERCENTILE

$$L_p = (n + 1) \frac{P}{100}$$

[4-1]

- ▶ Quartiles divide a set of observations into four equal parts
 - ▶ The interquartile range is the difference between the third quartile and the first quartile
- ▶ Deciles divide a set of observations into 10 equal parts
- ▶ Percentiles divide a set of observations into 100 equal parts

Measures of Position Example

- ▶ Morgan Stanley is an investment company with offices located throughout the United States. Listed below are the commissions earned last month by a sample of 15 brokers.

\$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			

- ▶ First, sort the data from smallest to largest

\$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	

Measures of Position Example

- ▶ Next, find the median
- ▶ $L_{50} = (15+1)*50/100 = 8$
- ▶ So the median is \$2,038, the value at position 8

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

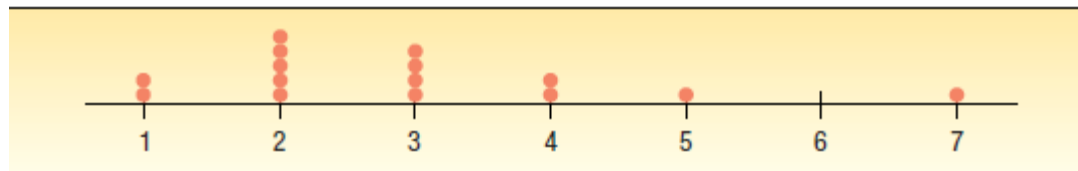
Therefore, the first and third quartiles are located at the 4th and 12th positions, respectively: $L_{25} = \$1,721$; $L_{75} = \$2,205$

L 25 (Q1) → means 25% of data is less than 1st quartile
L 75 (Q3) – L 25(Q1) → Interquartile range covers the middle 50 % of the data
L50 (Q2) → means 50% of data is less than 2nd quartile ;
the median is the same as the 50th percentile

\$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	

Exercise (1 / 4)

- ▶ Exercise 1: Describe the differences between a histogram and a dot plot. When might a dot plot be better than a histogram?
- ▶ Exercise 2: Describe the differences between a histogram and a stem-and-leaf display.
- ▶ Exercise 3: Consider the following chart.
 - ▶ a. What is this chart called?
 - ▶ b. How many observations are in the study?
 - ▶ c. What are the maximum and the minimum values?
 - ▶ d. Around what values do the observations tend to cluster?



Exercise (2/4)

- ▶ Exercise 5: The first row of a stem-and-leaf chart appears as follows:
- ▶ 62 / 1 3 3 7 9. Assume whole number values.
 - ▶ a. What is the “possible range” of the values in this row?
 - ▶ b. How many data values are in this row?
 - ▶ c. List the actual values in this row of data.
- ▶ Exercise 7: The following stem-and-leaf chart from the Minitab software shows the number of units produced per day in a factory.
 - ▶ a. How many days were studied?
 - ▶ b. How many observations are in the first class?
 - ▶ c. What are the minimum value and the maximum value?
 - ▶ d. List the actual values in the fourth row.
 - ▶ e. List the actual values in the second row.
 - ▶ f. How many values are less than 70?
 - ▶ g. How many values are 80 or more?
 - ▶ h. What is the median?
 - ▶ i. How many values are between 60 and 89, inclusive?

1	3	8
1	4	
2	5	6
9	6	0133559
(7)	7	0236778
9	8	59
7	9	00156
2	10	36

Exercise (3/4)

- Exercise 9: A survey of the number of cell phone calls made by a sample of Verizon subscribers last week revealed the following information. Develop a stem-and-leaf chart. How many calls did a typical subscriber make? What were the maximum and the minimum number of calls made?

52	43	30	38	30	42	12	46	39
37	34	46	32	18	41	5		

Exercise (4/4)

Exercise 11: Determine the median and the values corresponding to the first and third quartiles in the following data.

46	47	49	49	51	53	54	54	55	55	59
----	----	----	----	----	----	----	----	----	----	----

Exercise 13: The Thomas Supply Company Inc. is a distributor of gas-powered generators. as with any business, the length of time customers take to pay their invoices is important. Listed below, arranged from smallest to largest, is the time, in days, for a sample of The Thomas Supply Company Inc. invoices.

- Determine the first and third quartiles.
- Determine the second decile and the eighth decile.
- Determine the 67th percentile.

13	13	13	20	26	27	31	34	34	34	35	35	36	37	38
41	41	41	45	47	47	47	50	51	53	54	56	62	67	82

Box Plots

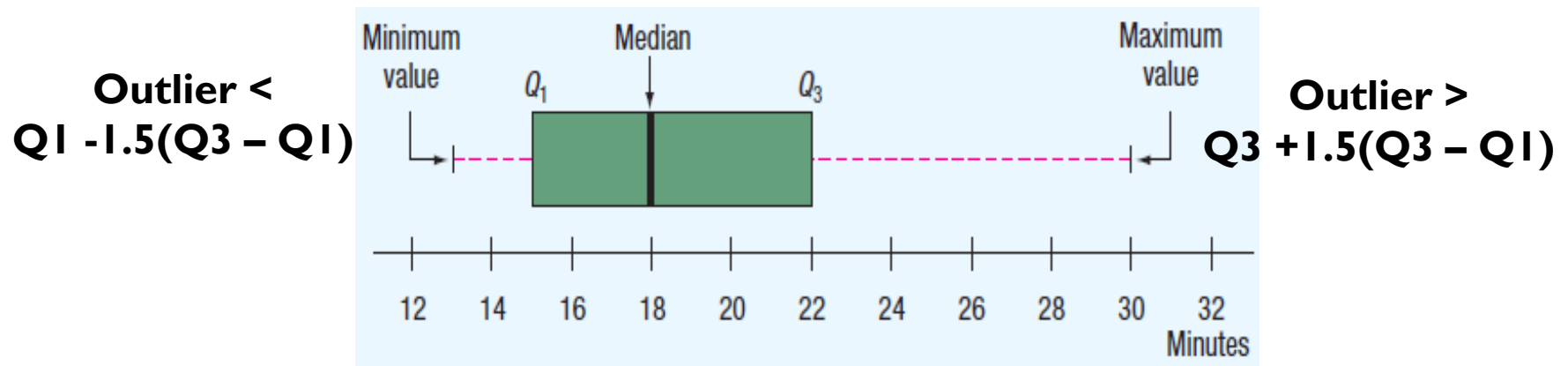
- ▶ A box plot is a graphical display using quartiles
- ▶ A box plot is based on five statistics:
 - ▶ Minimum value
 - ▶ 1st quartile
 - ▶ Median
 - ▶ 3rd quartile
 - ▶ Maximum value
- ▶ The interquartile range is $Q3 - Q1$
- ▶ Outliers are values that are inconsistent with the rest of the data and are identified with asterisks in box plots

Box Plot Example

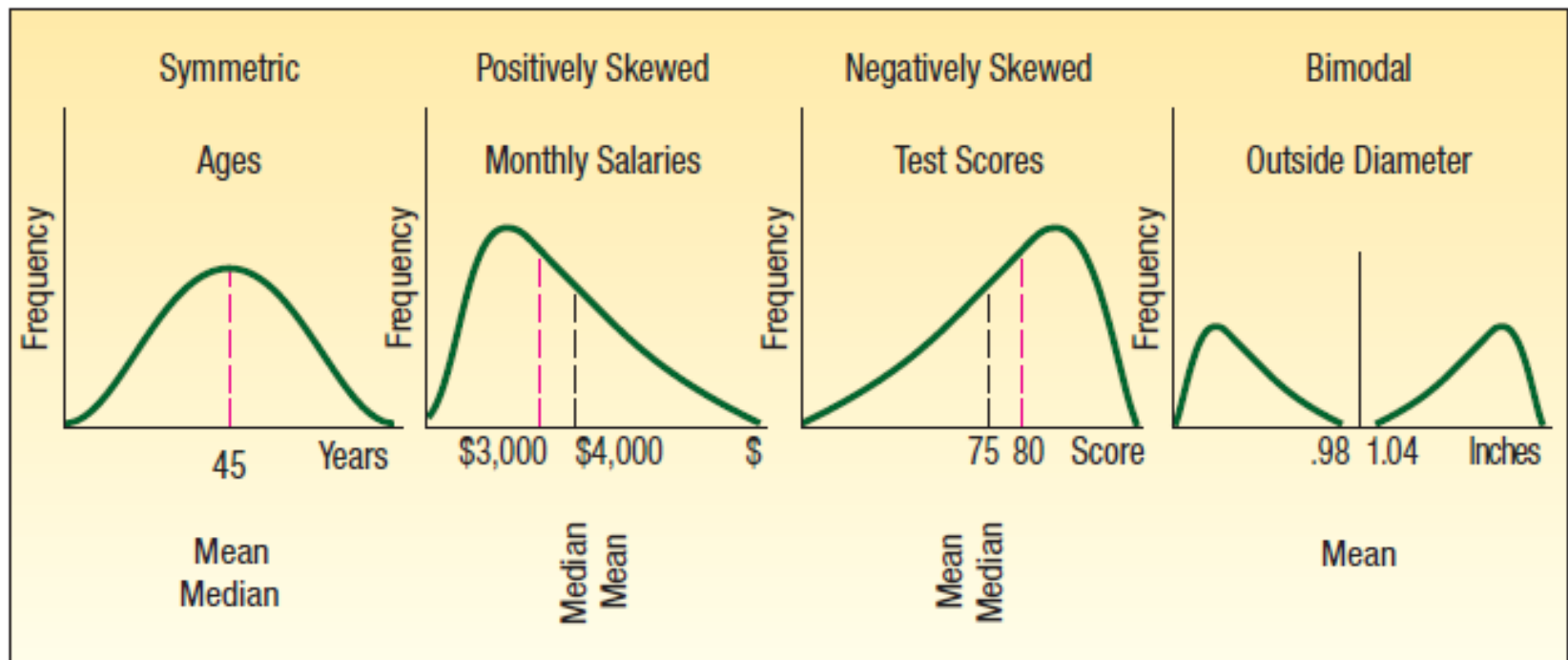
- ▶ Alexander's Pizza offers free delivery of its pizza within 15 miles. How long does a typical delivery take? Within what range will most deliveries be completed?
- ▶ Using a sample of 20 deliveries, Alexander determined the following:
 - ▶ Minimum value = 13 minutes
 - ▶ Q1 = 15 minutes
 - ▶ Median = 18 minutes
 - ▶ Q3 = 22 minutes
 - ▶ Maximum value = 30 minutes
- ▶ Develop a box plot for delivery times

Box Plot Example Continued

- ▶ Begin by drawing a number line using an appropriate scale
- ▶ Next, draw a box that begins at Q_1 (15 minutes) and ends at Q_3 (22 minutes)
- ▶ Draw a vertical line at the median (18 minutes)
- ▶ Extend a horizontal line out from Q_3 to the maximum value (30 min) and out from Q_1 to the minimum value (13 min)



Common Shapes of Data



Skewness

- ▶ The coefficient of skewness is a measure of the symmetry of a distribution
- ▶ Two formulas for coefficient of skewness

PEARSON'S COEFFICIENT OF SKEWNESS $sk = \frac{3(\bar{x} - \text{Median})}{s}$ [4-2]

- ▶ The coefficient of skewness can range from -3 to +3
- ▶ A value near -3 indicates considerable negative skewness
- ▶ A value of 1.63 indicates moderate positive skewness
- ▶ A value of 0 means the distribution is symmetrical

Skewness Example

- ▶ Following are the earnings per share for a sample of 15 software companies for the year 2016. 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	

- ▶ Begin by finding the mean, median, and standard deviation. Find the coefficient of skewness.
- ▶ What do you conclude about the shape of the distribution?

Skewness Example

Step 1: Compute the Mean

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

Step 2: Compute the Standard Deviation

$$s = \sqrt{\frac{\sum (X - \bar{X})^2}{n - 1}} = \sqrt{\frac{(\$0.09 - \$4.95)^2 + \dots + (\$16.40 - \$4.95)^2}{15 - 1}} = \$5.22$$

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

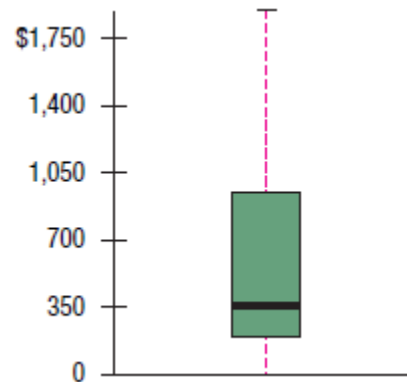
$$sk = \frac{3(\bar{X} - \text{Median})}{s} = \frac{3(\$4.95 - \$3.18)}{\$5.22} = 1.017$$

► What do you conclude about the shape of the distribution?

Exercise (1 / 3)

Exercise 15. The box plot below shows the amount spent for books and supplies per year by students at four-year public colleges.

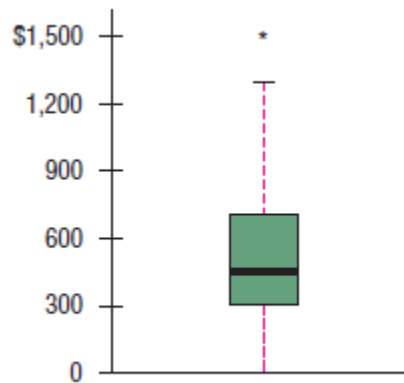
- a. Estimate the median amount spent.
- b. Estimate the first and third quartiles for the amount spent.
- c. Estimate the interquartile range for the amount spent.
- d. Beyond what point is a value considered an outlier?



Exercise (2/3)

Exercise 16: The box plot shows the undergraduate in-state charge per credit hour at four-year public colleges.

- Estimate the median.
- Estimate the first and third quartiles.
- Determine the interquartile range.
- Beyond what point is a value considered an outlier?
- Identify any outliers and estimate their value.
- Is the distribution symmetrical or positively or negatively skewed?



Exercise (3/3)

- a. Determine the mean, median, and the standard deviation.
- b. Determine the coefficient of skewness using Pearson's method.

Exercise 19: The following values are the starting salaries, in \$000, for a sample of five accounting graduates who accepted positions in public accounting last year.

36.0	26.0	33.0	28.0	31.0
------	------	------	------	------

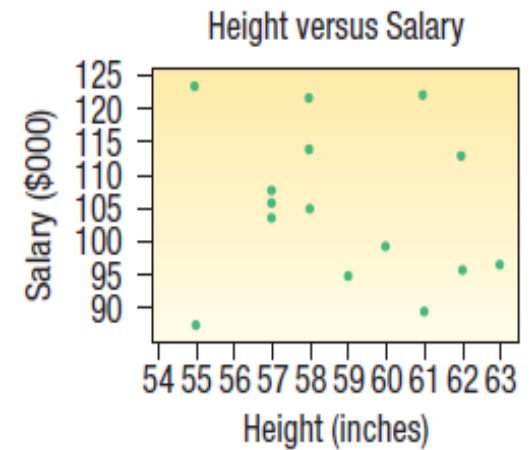
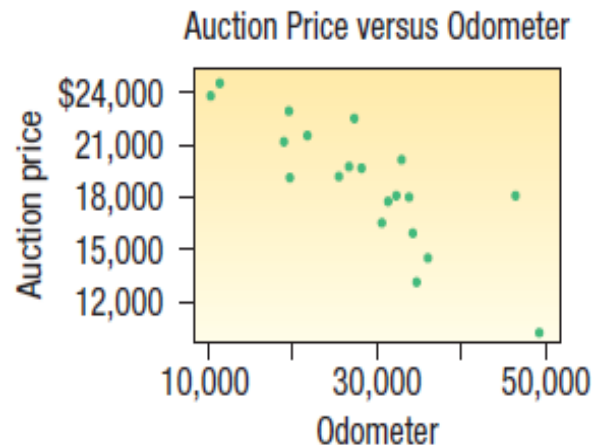
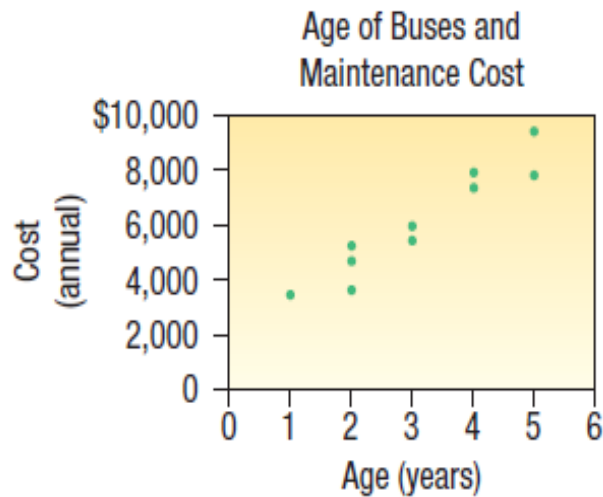
Exercise 20: Listed below are the salaries, in \$000, for a sample of 15 chief financial officers in the electronics industry.

\$516.0	\$548.0	\$566.0	\$534.0	\$586.0	\$529.0
546.0	523.0	538.0	523.0	551.0	552.0
486.0	558.0	574.0			

Describing the Relationship Between Two Variables

- ▶ A scatter diagram is a graphical tool to portray the relationship between two variables or bivariate data
- ▶ Both variables are measured with interval or ratio level scale
- ▶ If the scatter of points moves from the lower left to the upper right, the variables under consideration are directly or positively related
- ▶ If the scatter of points moves from the upper left to the lower right, the variables are inversely or negatively related

Scatter Diagrams



Contingency Tables

- ▶ A contingency table is used to classify nominal scale observations according to two characteristics

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

- ▶ It is a cross-tabulation that simultaneously summarizes two variables of interest
- ▶ Both variables need only be nominal or ordinal

Contingency Table Example

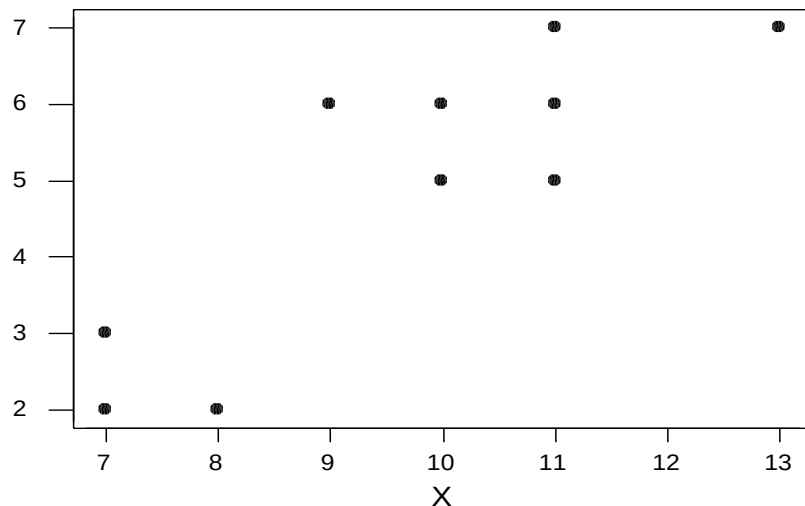
- ▶ Applewood Auto Group's profit comparison

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	<u>27</u>	<u>20</u>	<u>26</u>	<u>17</u>	<u>90</u>
Total	52	40	45	43	180

- ▶ 90 of the 180 cars sold had a profit above the median and half below. This meets the definition of median.
- ▶ The percentage of profits above the median are Kane 48%, Olean 50%, Sheffield 42% , and Tionesta 60%.

Exercise (1 / 4)

- **Exercise 23:** Develop a scatter diagram for the following sample data. How would you describe the relationship between the values?



X-Value	Y-Value	X-Value	Y-Value
10	6	11	6
8	2	10	5
9	6	7	2
11	5	7	3
13	7	11	7

Exercise (2/4)

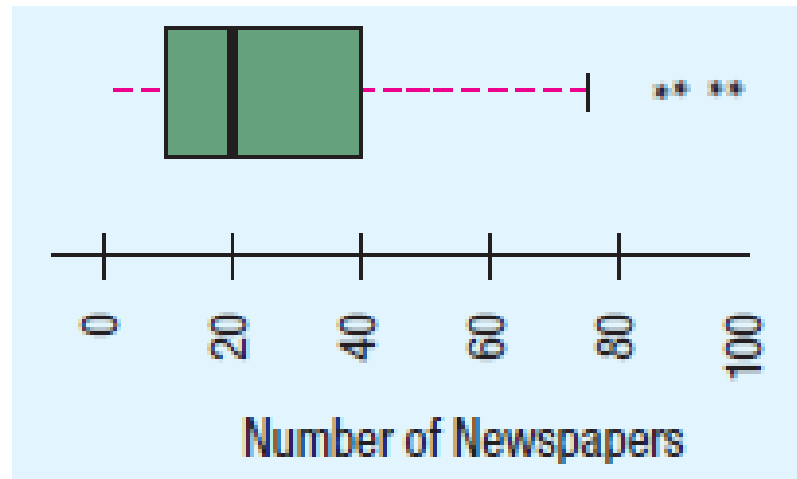
- **Exercise 26:** Ski Resorts of Vermont Inc. is considering a merger with Gulf Shores Beach Resorts Inc. of Alabama. The board of directors surveyed 50 stockholders concerning their position on the merger. The results are reported below.

Number of Shares Held	Opinion			Total
	Favor	Oppose	Undecided	
Under 200	8	6	2	16
200 up to 1,000	6	8	1	15
Over 1,000	6	12	1	19
Total	20	26	4	50

- What level of measurement is used in this table?
- What is this table called?
- What group seems most strongly opposed to the merger?

Exercise (3/4)

- **Exercise 34:** The following box plot shows the number of daily newspapers published in each state and the District of Columbia. Write a brief report summarizing the number published. Be sure to include information on the values of the first and third quartiles, the median, and whether there is any skewness. If there are any outliers, estimate their value.



Exercise (4/4)

Exercise 40: The manager of Information Services at Wilkin Investigations, a private investigation firm, is studying the relationship between the age (in months) of a combination printer, copy, and fax machine and its monthly maintenance cost. For a sample of 15 machines, the manager developed the following chart. What can the manager conclude about the relationship between the variables?

