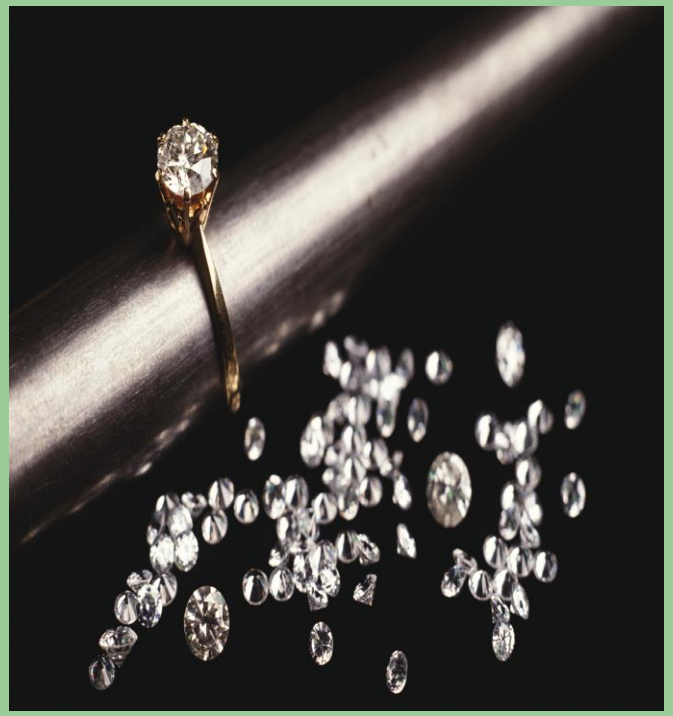
Describing Data:
Displaying and Exploring Data



Chapter 4

GOALS

1. Develop and interpret a *dot plot*.
2. Compute and understand *quartiles, deciles, and percentiles*.
3. Construct and interpret *box plots*.
4. Compute and understand the *coefficient of skewness*.
5. Draw and interpret a *scatter diagram*.
6. Construct and interpret a *contingency table*.

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 are most useful for smaller data sets, whereas histograms tend to be most useful for large data sets.

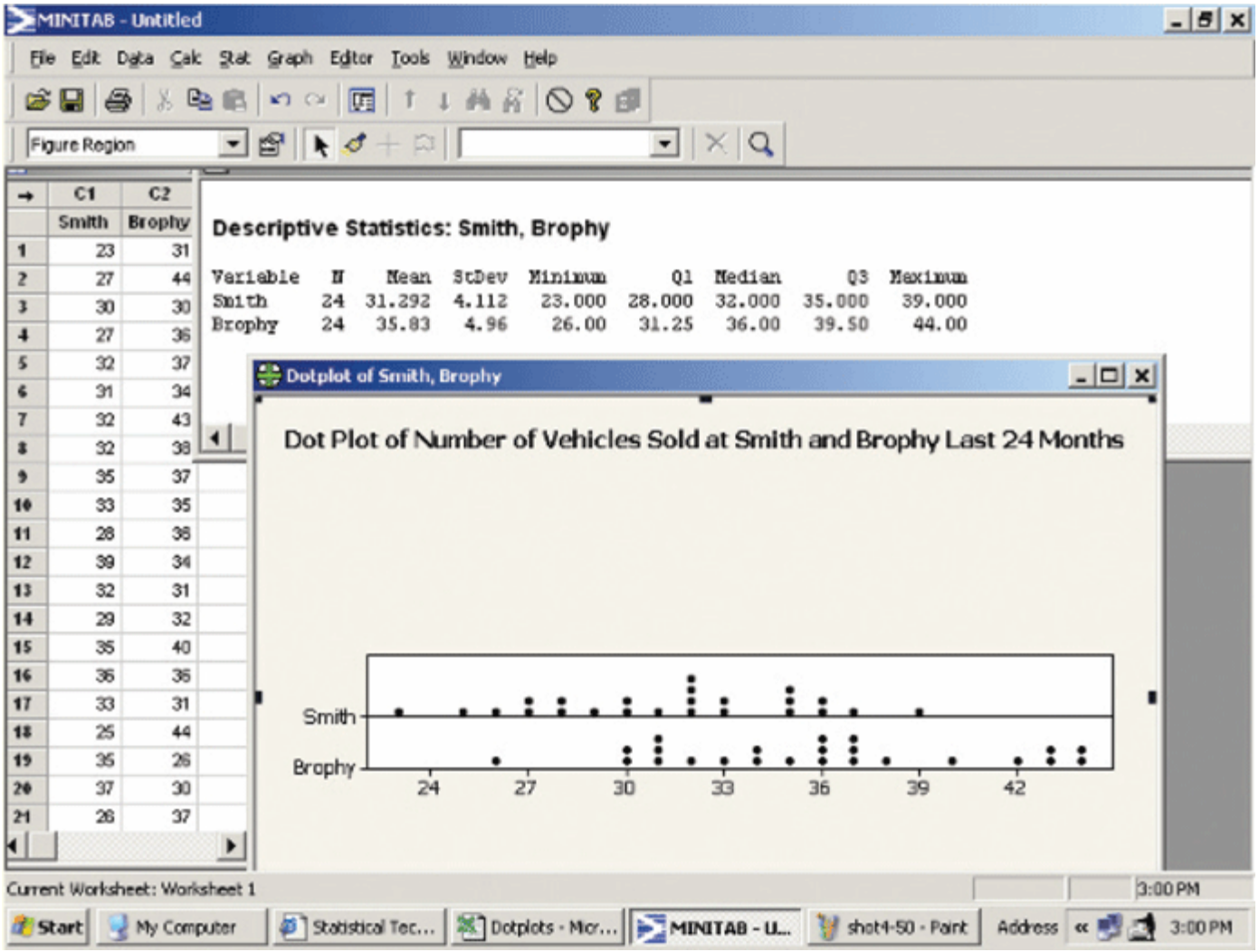
Dot Plots - Examples

Reported below are the number of vehicles sold in the last 24 months at Smith Ford Mercury Jeep, Inc., in Kane, Pennsylvania, and Brophy Honda Volkswagen in Greenville, Ohio. Construct dot plots and report summary statistics for the two small-town Auto USA lots.

Smith Ford Mercury Jeep, Inc.									
23	27	30	27	32	31	32	32	35	33
28	39	32	29	35	36	33	25	35	37
26	28	36	30						

Brophy Honda Volkswagen									
31	44	30	36	37	34	43	38	37	35
36	34	31	32	40	36	31	44	26	30
37	43	42	33						

Dot Plot – Minitab Example



Other Measures of Dispersion: Quartiles, Deciles and Percentiles

- 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

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

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

Percentile Computation

- Let L_p refer to the location of a desired percentile. 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

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

- The number of observations is *n.* To locate the median, its position is at **(*n* + 1)/2**. 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. Salomon Smith Barney is an investment company with offices located throughout the United States.

$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]**

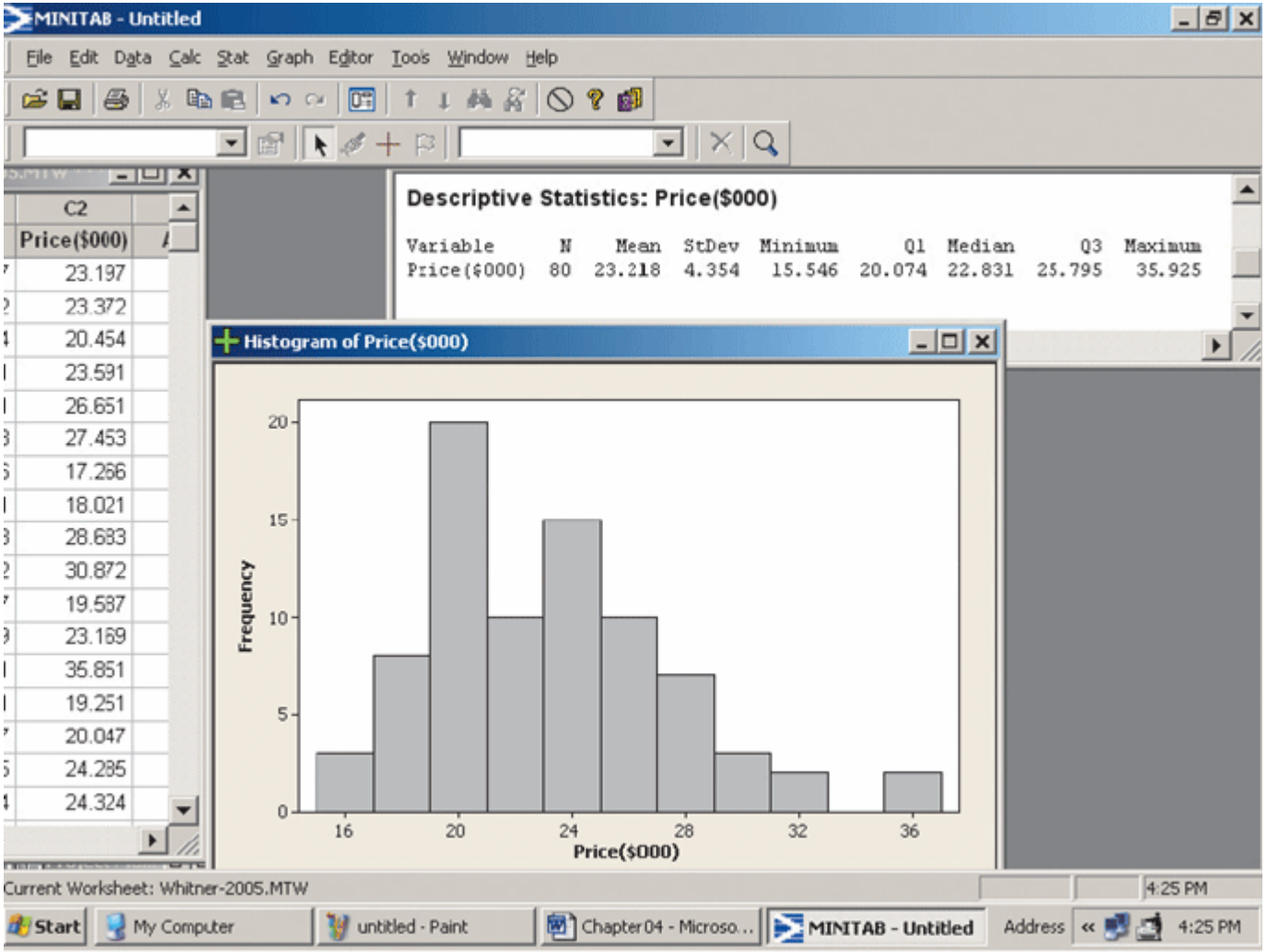
$$L_{25} = (15+1)\frac{25}{100} = 4 \qquad L_{75} = (15+1)\frac{75}{100} = 12$$

Therefore, the first and third quartiles are the 4th and 12th observations in the array, respectively

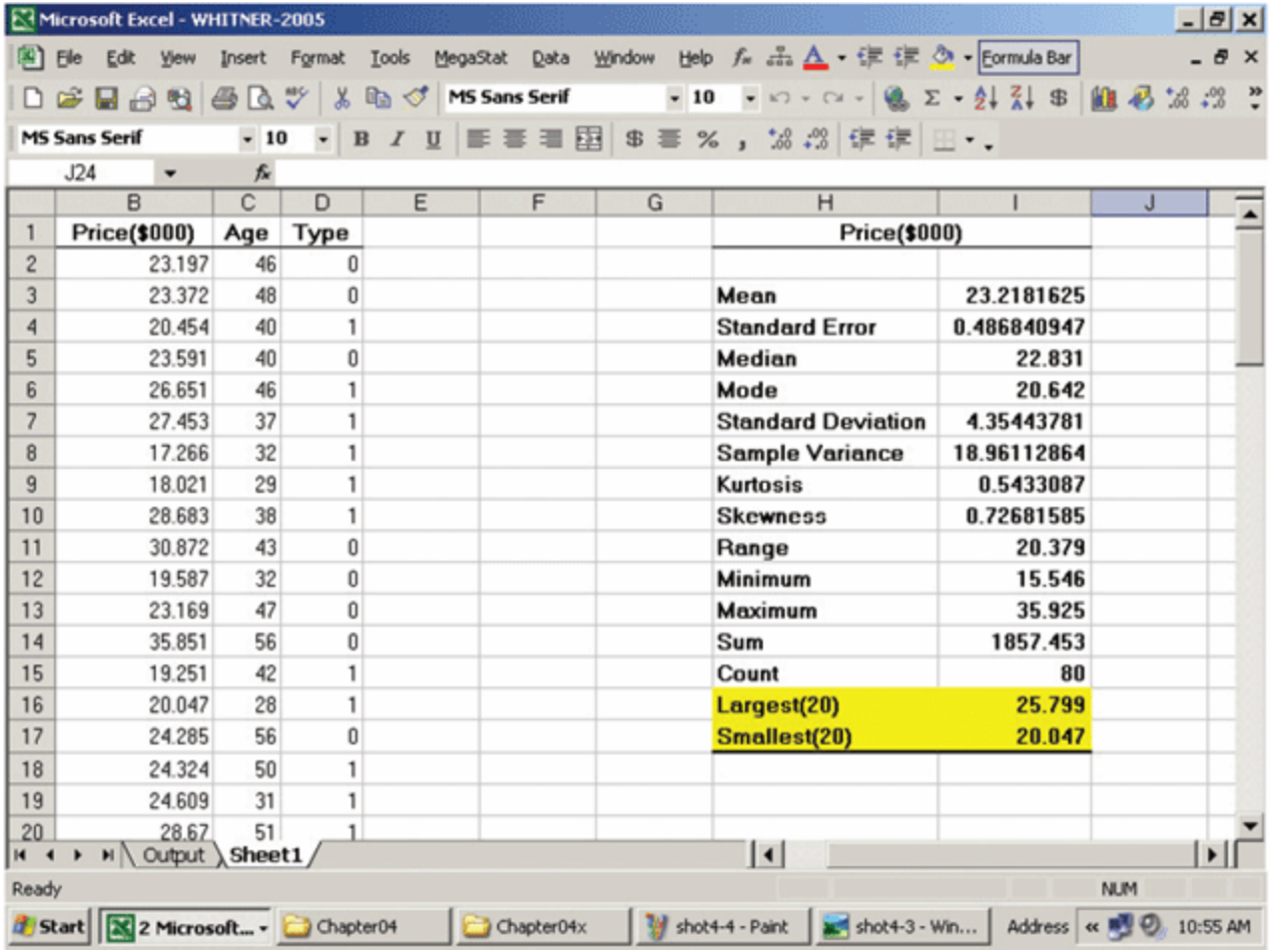
$L_{25} = \$1{,}721$

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

Percentiles – Example (Minitab)



Percentiles – Example (Excel)



	B	C	D	E	F	G	H	I	J
1	Price($000)	Age	Type				Price($000)		
2	23.197	46	0						
3	23.372	48	0				Mean	23.2181625	
4	20.454	40	1				Standard Error	0.486840947	
5	23.591	40	0				Median	22.831	
6	26.651	46	1				Mode	20.642	
7	27.453	37	1				Standard Deviation	4.35443781	
8	17.266	32	1				Sample Variance	18.96112864	
9	18.021	29	1				Kurtosis	0.5433087	
10	28.683	38	1				Skewness	0.72681585	
11	30.872	43	0				Range	20.379	
12	19.587	32	0				Minimum	15.546	
13	23.169	47	0				Maximum	35.925	
14	35.851	56	0				Sum	1857.453	
15	19.251	42	1				Count	80	
16	20.047	28	1				Largest(20)	25.799	
17	24.285	56	0				Smallest(20)	20.047	
18	24.324	50	1						
19	24.609	31	1						
20	28.67	51	1						

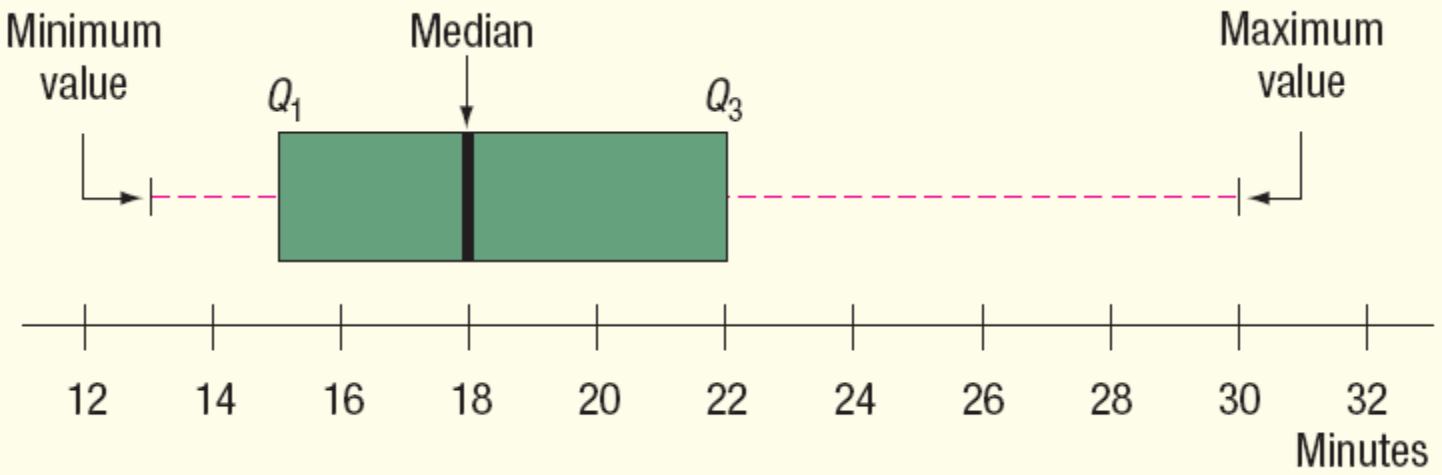
Boxplot - Example

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:

$$\text{Minimum value} = 13 \text{ minutes}$$
$$Q_1 = 15 \text{ minutes}$$
$$\text{Median} = 18 \text{ minutes}$$
$$Q_3 = 22 \text{ minutes}$$
$$\text{Maximum value} = 30 \text{ minutes}$$

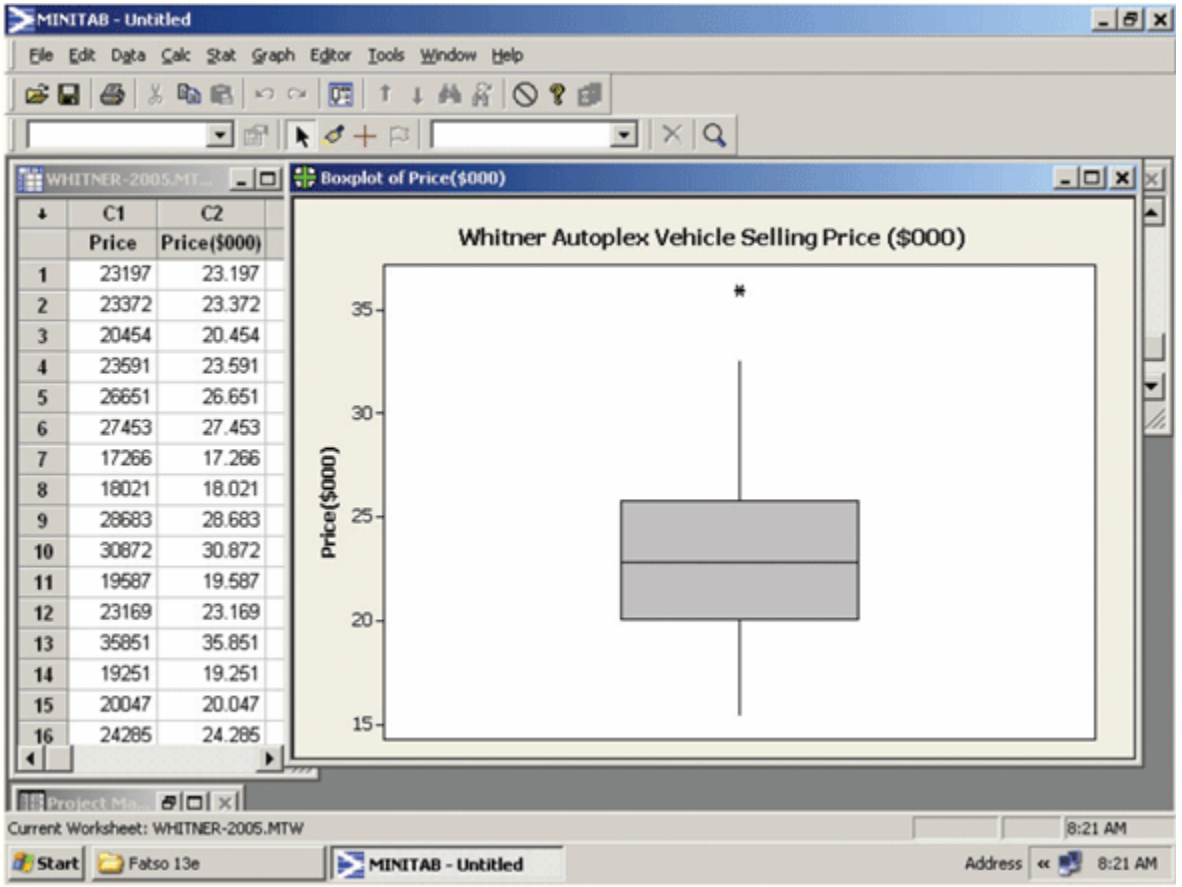
Develop a box plot for the delivery times. What conclusions can you make about the delivery times?

Boxplot Example



Boxplot – Using Minitab

Refer to the Whitner Autoplex data in Table 2–4. Develop a box plot of the data. What can we conclude about the distribution of the vehicle selling prices?



Skewness

- In Chapter 3, measures of central location for a set of observations (the mean, median, and mode) 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:
  1. symmetric,
  2. positively skewed,
  3. negatively skewed,
  4. bimodal.

Skewness - Formulas for Computing

PEARSON'S COEFFICIENT OF SKEWNESS

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

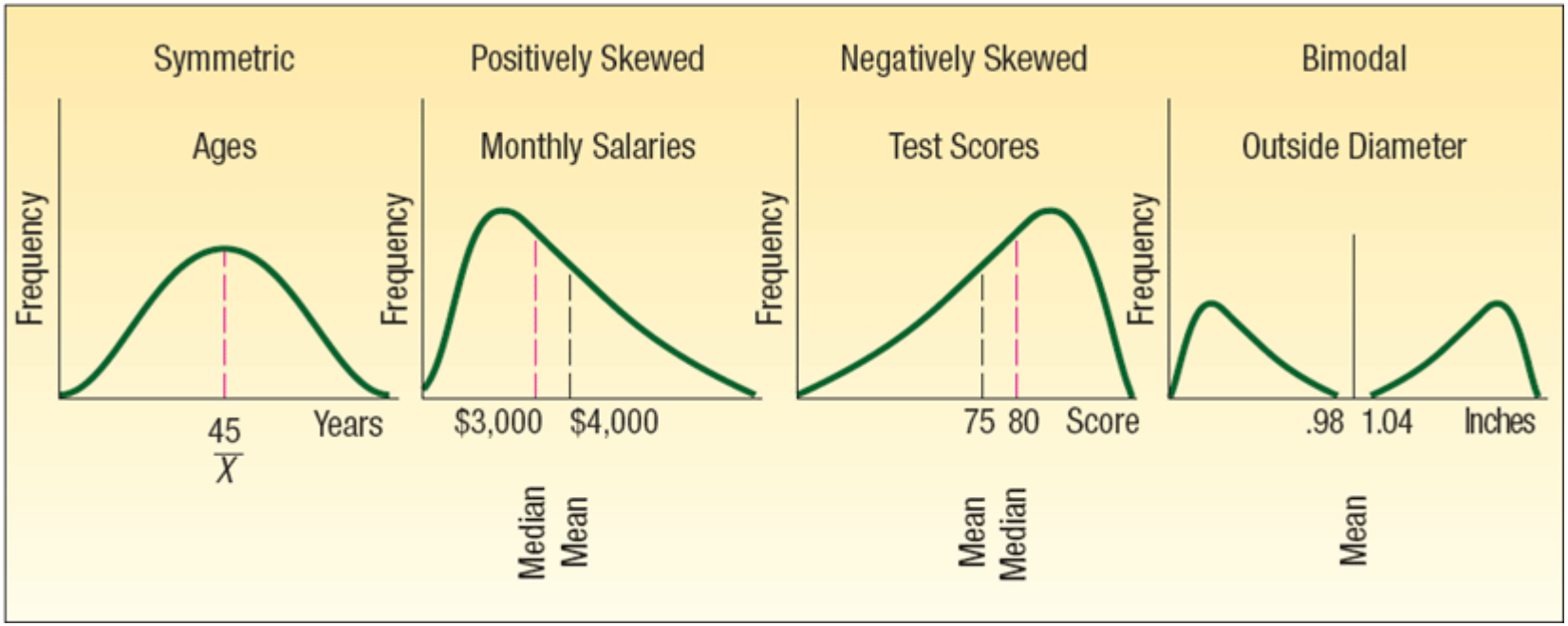
SOFTWARE COEFFICIENT OF SKEWNESS

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

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

- A value near -3, such as -2.57, 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.

Commonly Observed Shapes



Skewness – An Example

Following are the earnings per share for a sample of 15 software companies for the year 2005. 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 sample mean

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

Step 2: Compute the sample standard deviation

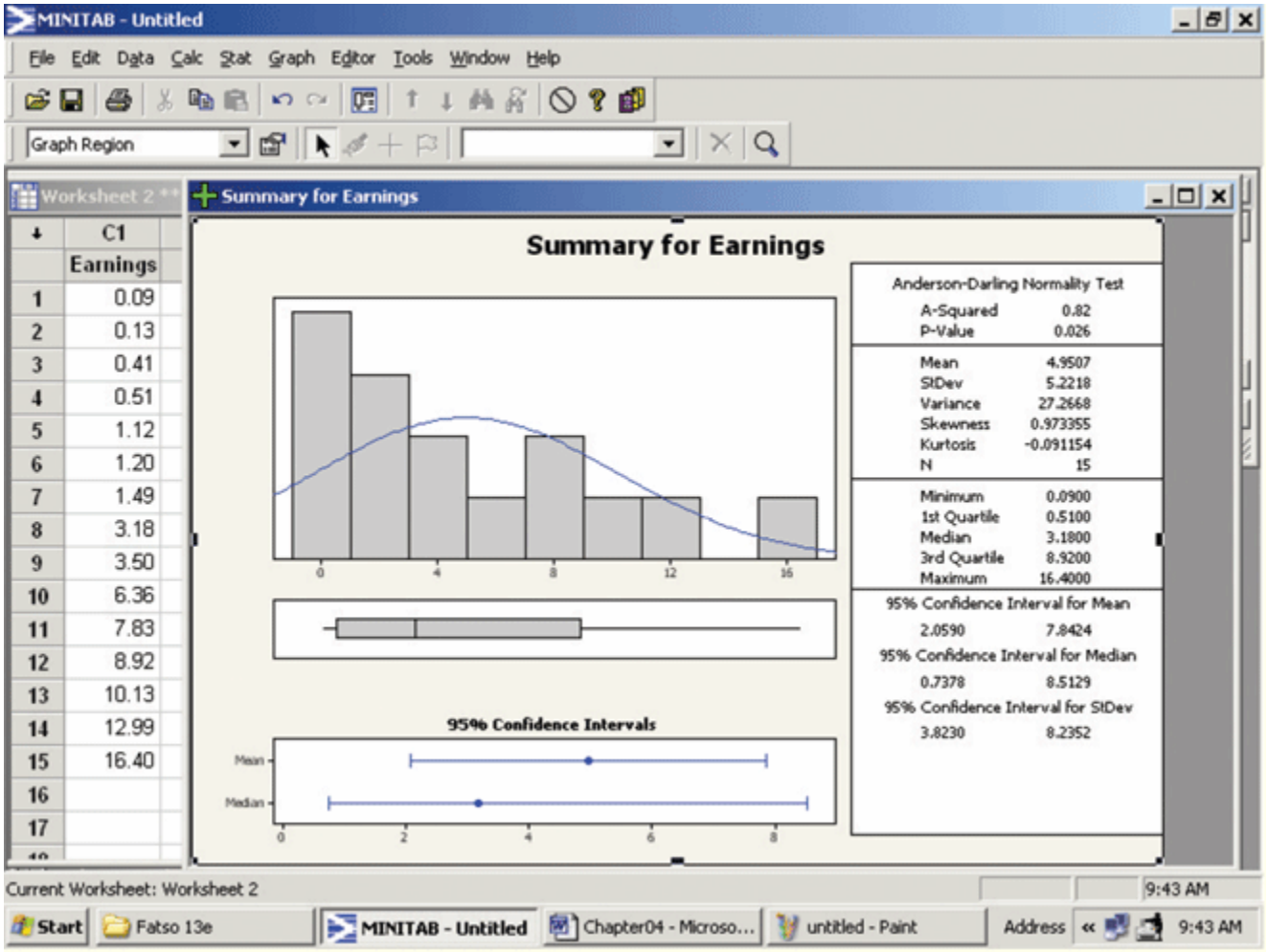
$$s = \sqrt{\frac{\Sigma\left(X - \overline{X}\right)^2}{n-1}} = \sqrt{\frac{(\$0.09 - \$4.95)^2 + \ldots + (\$16.40 - \$4.95)^2)}{15-1}} = \$5.22$$

Step 3: Determine the median - the middle value in a set of data, arranged from smallest to largest. In this case the middle value is \$3.18, so the median earnings per share is \$3.18.

Step 4: Compute the Skewness

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

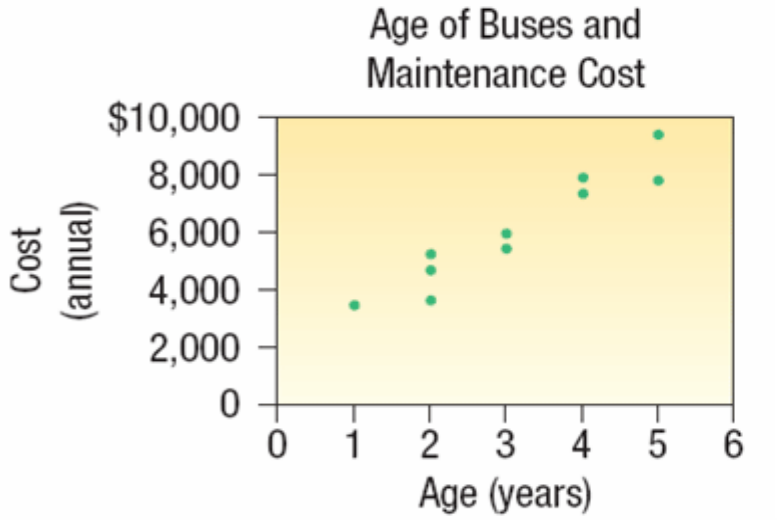
Skewness – A Minitab Example

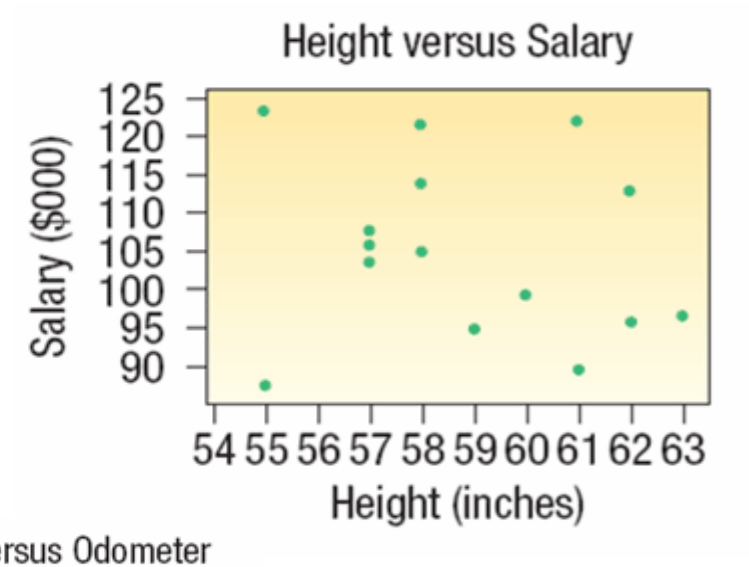


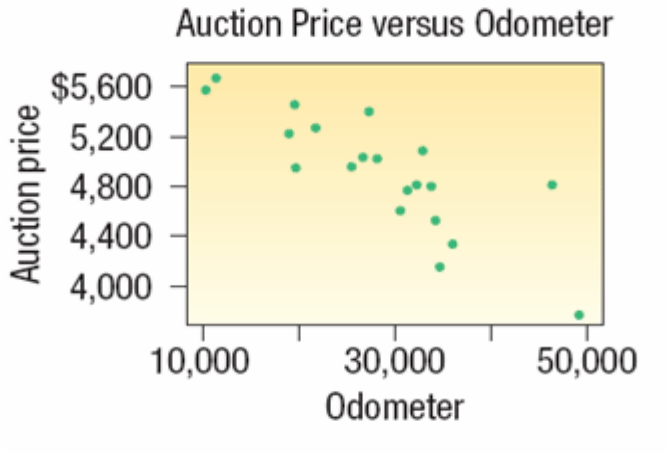
Describing Relationship between Two Variables

- 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).

Describing Relationship between Two Variables – Scatter Diagram Examples





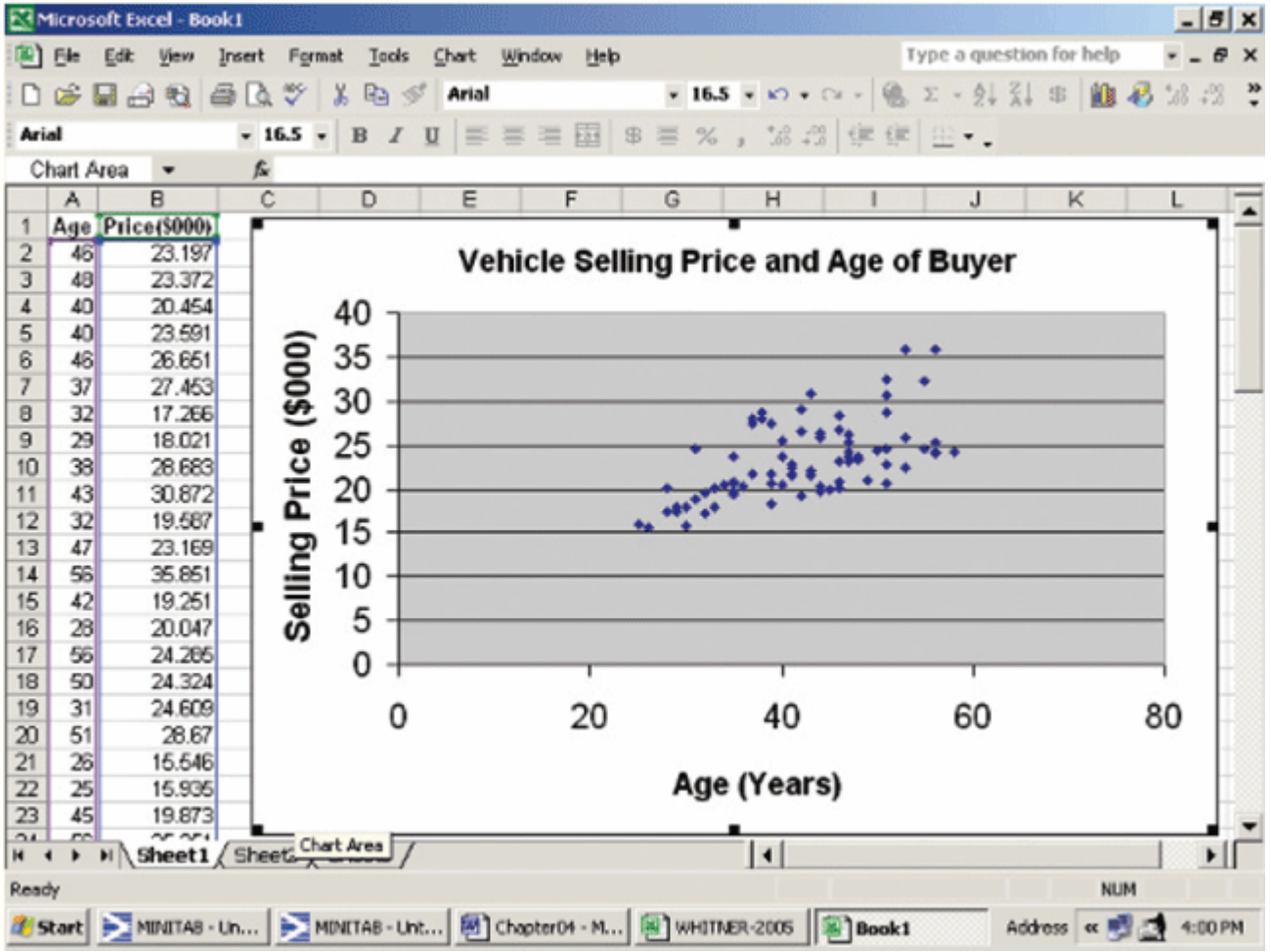


Describing Relationship between Two Variables – Scatter Diagram Excel Example

In the Introduction to Chapter 2 we presented data from AutoUSA. In this case the information concerned the prices of 80 vehicles sold last month at the Whitner Autoplex lot in Raytown, Missouri. The data shown include the selling price of the vehicle as well as the age of the purchaser.

Is there a relationship between the selling price of a vehicle and the age of the purchaser? Would it be reasonable to conclude that the more expensive vehicles are purchased by older buyers?

Describing Relationship between Two Variables – Scatter Diagram Excel Example



	A	B
1	Age	Price($000)
2	46	23.197
3	48	23.372
4	40	20.454
5	40	23.591
6	46	26.651
7	37	27.453
8	32	17.266
9	29	18.021
10	38	28.683
11	43	30.872
12	32	19.587
13	47	23.169
14	56	35.851
15	42	19.251
16	28	20.047
17	56	24.266
18	50	24.324
19	31	24.609
20	51	28.67
21	26	15.546
22	25	15.936
23	45	19.873

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 – An Example

A manufacturer of preassembled windows produced 50 windows yesterday. This morning the quality assurance inspector reviewed each window for all quality aspects. Each was classified as acceptable or unacceptable and by the shift on which it was produced. The two variables are shift and quality. The results are reported in the following table.

	Shift			
	Day	Afternoon	Night	Total
Defective	3	2	1	6
Acceptable	17	13	14	44
Total	20	15	15	50

Contingency Tables – An Example

	Shift			
	Day	Afternoon	Night	Total
Defective	3	2	1	6
Acceptable	17	13	14	44
Total	20	15	15	50

Usefulness of the Contingency Table:

By organizing the information into a contingency table we can compare the quality on the three shifts.

For example, on the day shift, 3 out of 20 windows or 15 percent are defective. On the afternoon shift, 2 of 15 or 13 percent are defective and on the night shift 1 out of 15 or 7 percent are defective.

Overall 12 percent of the windows are defective. Observe also that 40 percent of the windows are produced on the day shift, found by (20/50)(100).

End of Chapter 4