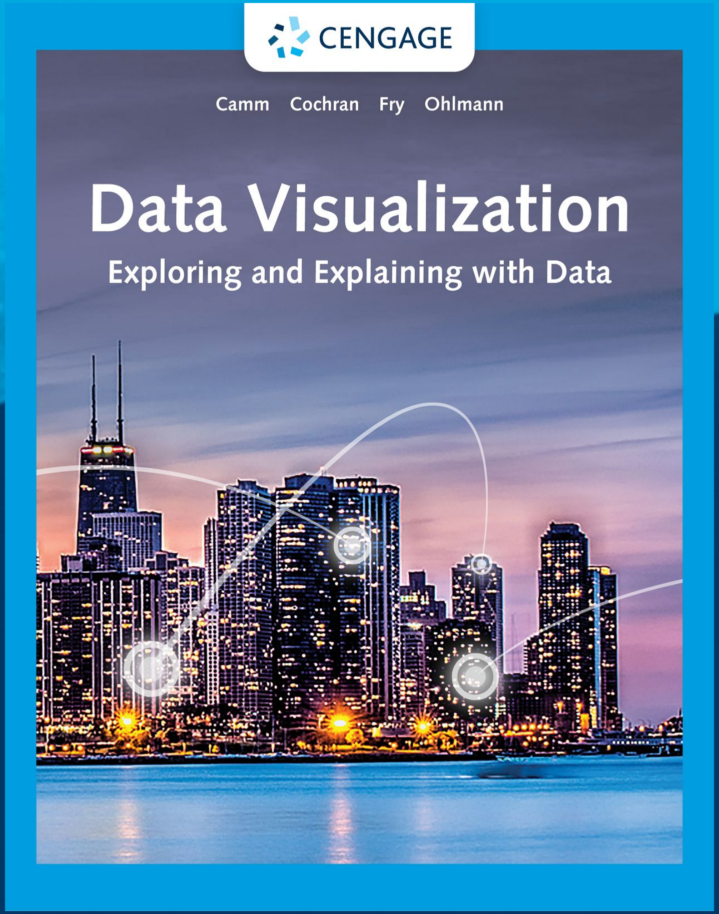


Data Visualization, 1e

Chapter 5: Visualizing Variability

Chapter Objectives (1 of 2)

After completing this chapter, you will be able to:

LO 5.1 Create and interpret charts used to visualize the frequency distribution of a categorical variable

LO 5.2 Create and interpret histograms and frequency polygons, charts used to visualize the distribution of a quantitative variable

LO 5.3 Create and interpret visualizations comparing the distributions of two or more variables

LO 5.4 Create and interpret strip charts, recognize situations to use them, and employ techniques to improve their clarity

Chapter Objectives (2 of 2)

LO 5.5 Describe basic statistical measures of central location, variability, and distribution shape

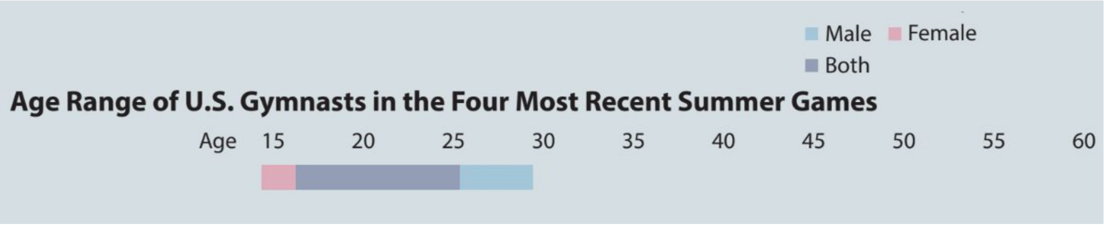
LO 5.6 Create and interpret a box and whisker chart

LO 5.7 Create and interpret visualizations that depict the uncertainty resulting from sampling error

LO 5.8 Create and interpret charts that depict the uncertainty in predictions from simple regression models and time series models

Data Visualization Makeover

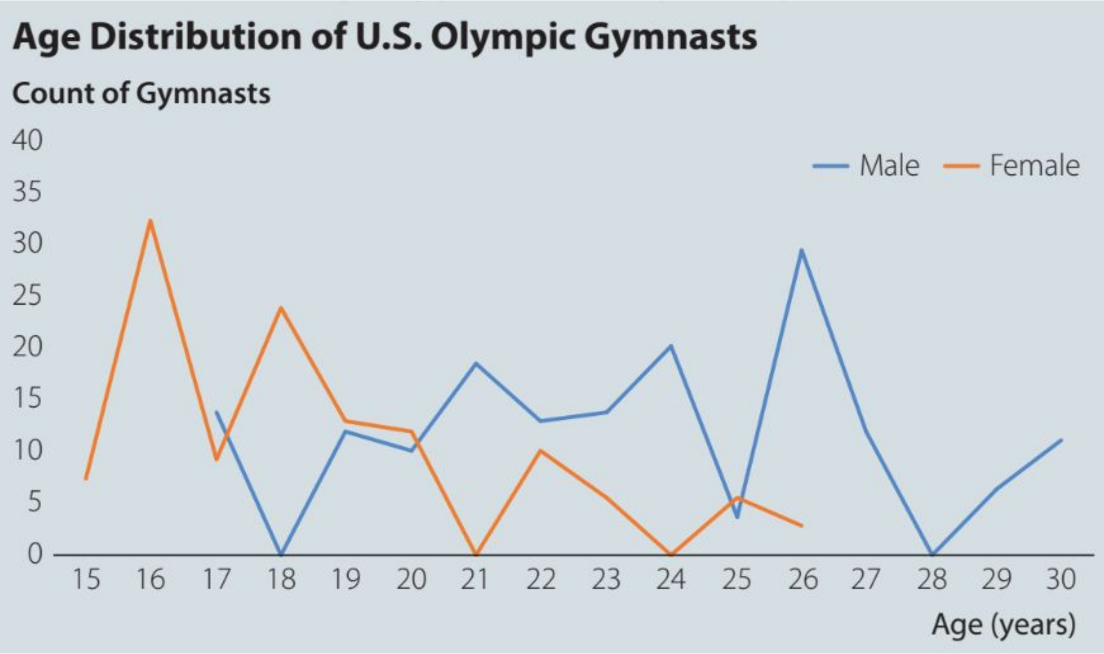
Overlapping range bar chart for ages of U.S. Olympic gymnasts



The third color in the range bar chart above is not intuitive and increases cognitive load.

The frequency polygons to the right reduce cognitive load and show the age distribution of male and female gymnasts.

Frequency polygon chart for U.S. Olympic gymnasts



Camm, Cochran, Fry & Ohlmann, Data Visualization - Exploring and Explaining with Data, 1st Edition.

5.1 Creating Distributions from Data: Definitions

A **random variable** is a quantity with values not known with certainty.

Variation is the difference in a variable measured over observations.

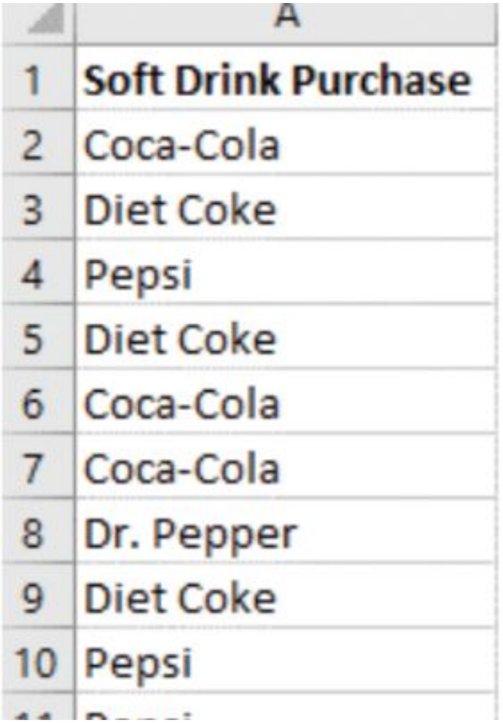
A **frequency distribution** describes the values of a variable and how often they appear in the data.

- For a **categorical variable**, data consist of labels or names for which arithmetical manipulation is impossible.
- For a **quantitative variable**, data consist of numerical values for which arithmetical manipulation is possible.

A **sample** is a subset of the **population** that makes data collection feasible.

5.1 Visualize a Frequency Distribution: Column Chart

Create a column chart for the categorical data in the file *Pop*



	A
1	**Soft Drink Purchase**
2	Coca-Cola
3	Diet Coke
4	Pepsi
5	Diet Coke
6	Coca-Cola
7	Coca-Cola
8	Dr. Pepper
9	Diet Coke
10	Pepsi
11	Pepsi

Step 1. Select cells A1:A51

Step 2. Click the **Insert** tab on the Ribbon

Step 3. Click the **Recommended Charts** button in the **Charts** group

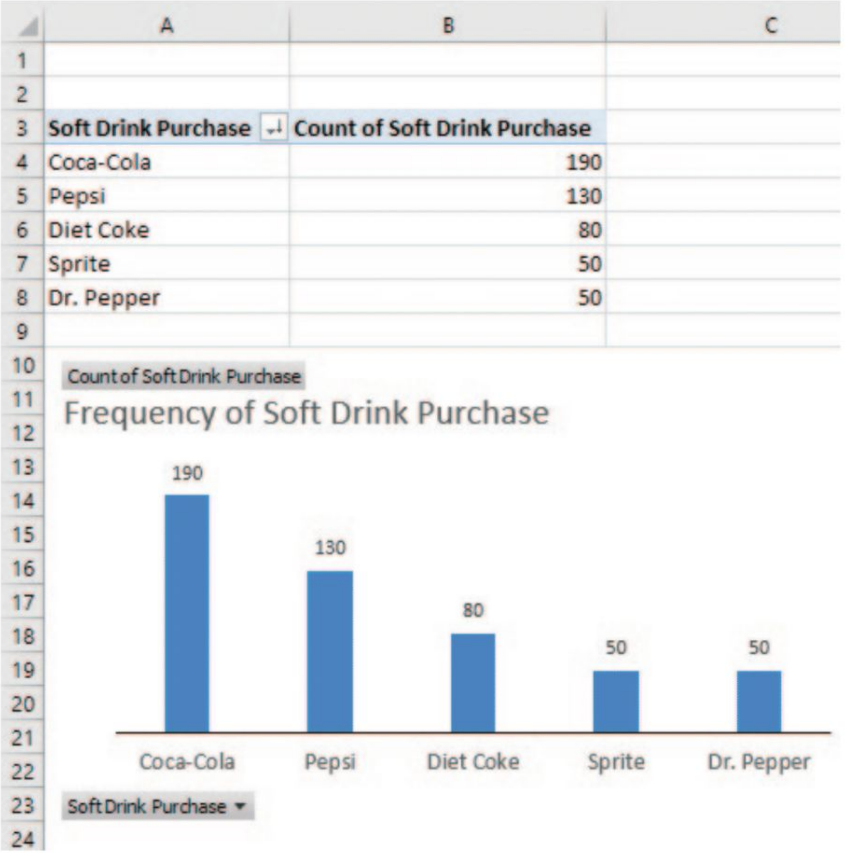
Step 4. When the **Insert Chart** dialog box appears:

- Select **Clustered Column**
- Click **OK**

Step 5. Click any of the columns in the chart that appears. While the columns are selected, right-click a column, then select **Sort** and **Sort Largest to Smallest**

5.1 Visualize a Frequency Distribution: PivotTable

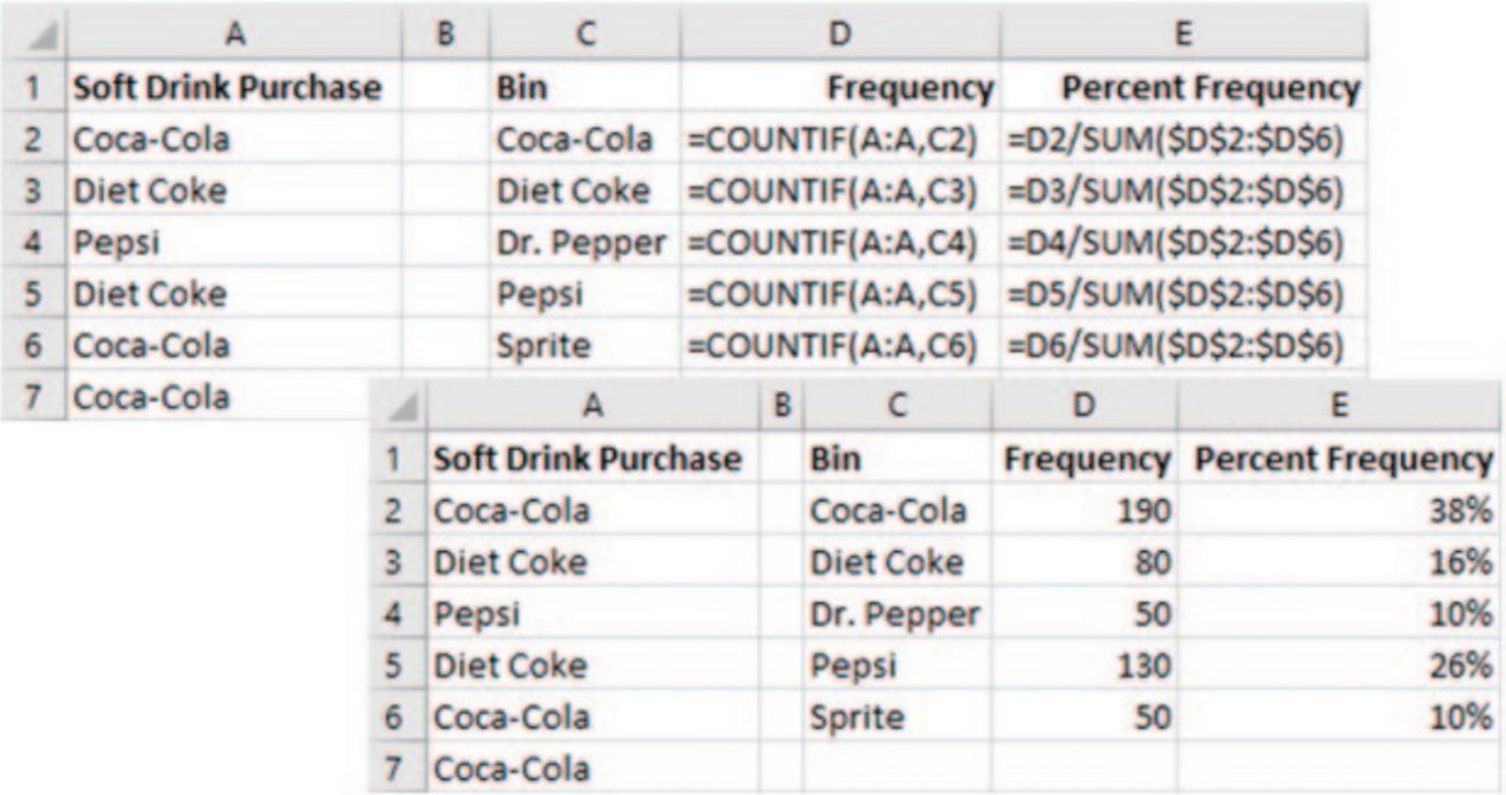
PivotTable and PivotChart to create a frequency distribution of soft drink purchase data



Soft Drink Purchase	Count of Soft Drink Purchase
Coca-Cola	190
Pepsi	130
Diet Coke	80
Sprite	50
Dr. Pepper	50

5.1 Build a Relative Frequency Distribution

Create a frequency distribution for categorical data using the COUNTIF function



	A	B	C	D	E
1	**Soft Drink Purchase**		**Bin**	**Frequency**	**Percent Frequency**
2	Coca-Cola		Coca-Cola	=COUNTIF(A:A,C2)	=D2/SUM($D$2:$D$6)
3	Diet Coke		Diet Coke	=COUNTIF(A:A,C3)	=D3/SUM($D$2:$D$6)
4	Pepsi		Dr. Pepper	=COUNTIF(A:A,C4)	=D4/SUM($D$2:$D$6)
5	Diet Coke		Pepsi	=COUNTIF(A:A,C5)	=D5/SUM($D$2:$D$6)
6	Coca-Cola		Sprite	=COUNTIF(A:A,C6)	=D6/SUM($D$2:$D$6)
7	Coca-Cola				

	A	B	C	D	E
1	**Soft Drink Purchase**		**Bin**	**Frequency**	**Percent Frequency**
2	Coca-Cola		Coca-Cola	190	38%
3	Diet Coke		Diet Coke	80	16%
4	Pepsi		Dr. Pepper	50	10%
5	Diet Coke		Pepsi	130	26%
6	Coca-Cola		Sprite	50	10%
7	Coca-Cola				

5.1 Relative Frequency and Percent Frequency

Definitions

The **relative frequency** of a bin equals the proportion of items belonging to a class:

$$\text{Relative Freq. of a bin} = \frac{\text{Freq. of a bin}}{\text{n}}$$

The **percent frequency** of a bin is the relative frequency multiplied by 100.

$$\text{Percent Freq. of a bin} = 100 \cdot \text{Relative Freq. of a bin}$$

A **probability distribution** characterizes the variability of a random variable.

- A percent frequency distribution estimates a probability distribution.

Obtain a percent frequency distribution

Step 1. Select any cell in the *Count of Soft Drink Purchase* column of the PivotTable (any cell in range B3:B8)

Step 2. From **PivotTable Fields** task pane:

In the **Values** area, select the triangle to the left of **Count of Soft Drink Purchase**

Select **Value Field Settings** from the list of options.

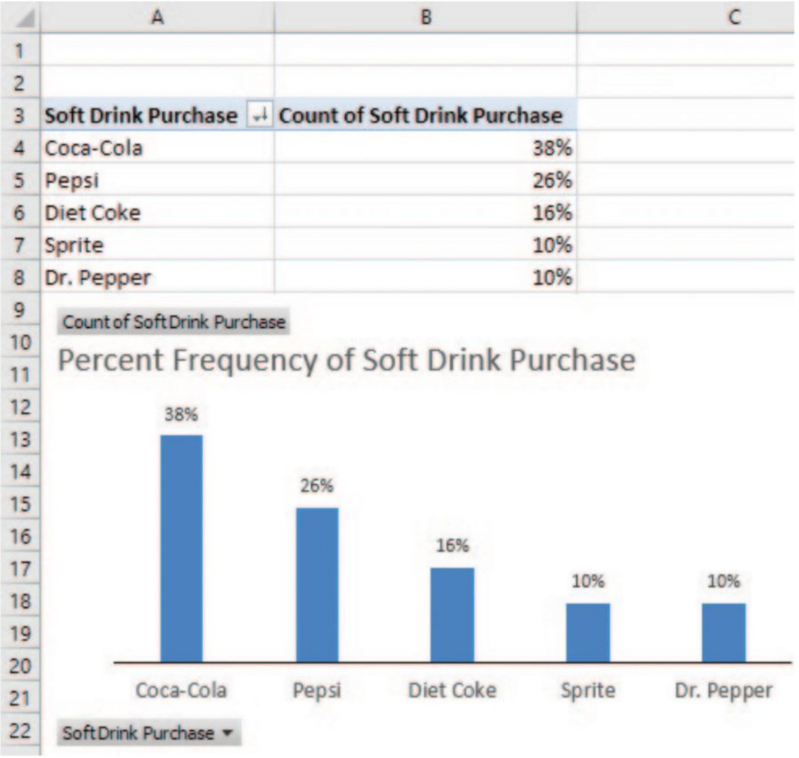
Step 3. From the **Value Field Settings** dialog box:

Click the **Show Values As** tab and in the box below **Show values as**, select **% of Grand Total**

CENGAGE

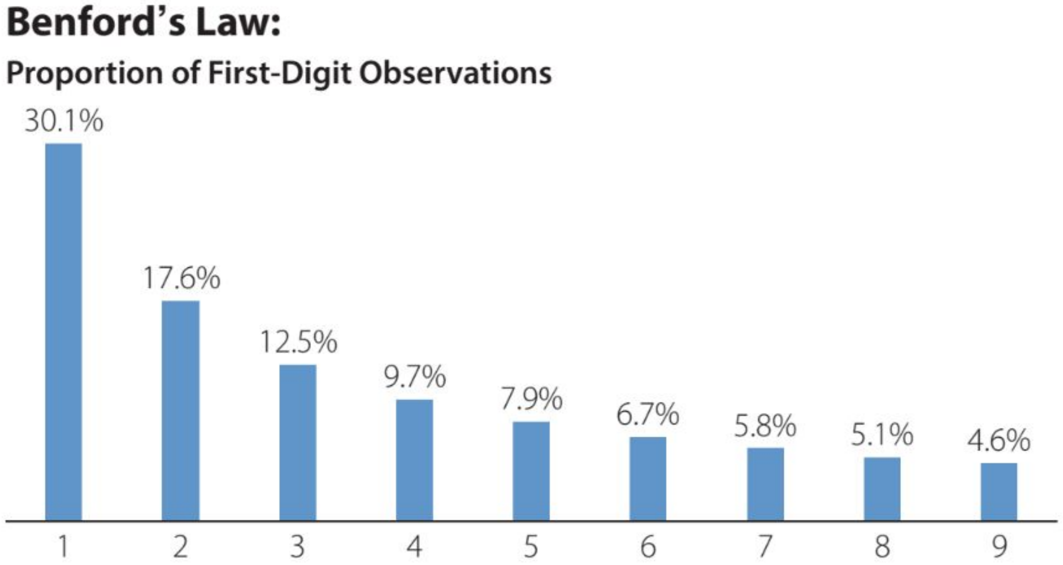
5.1 Application of Benford’s Law

Percent frequency distribution of soft drink purchase data



Soft Drink Purchase	Count of Soft Drink Purchase
Coca-Cola	38%
Pepsi	26%
Diet Coke	16%
Sprite	10%
Dr. Pepper	10%

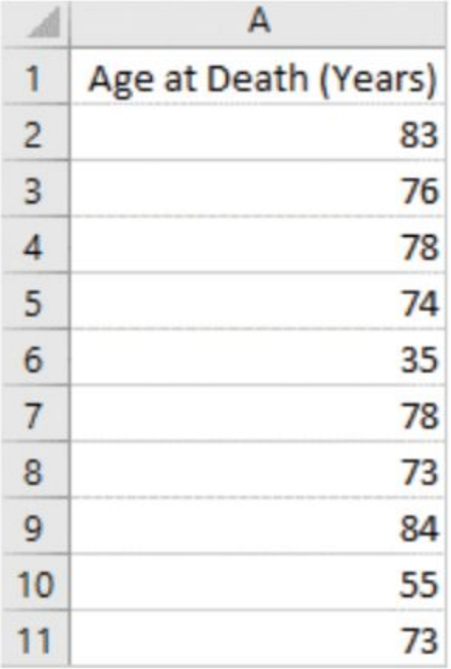
Relative frequency distribution of data obeying Benford’s Law



Benford’s Law states that in many data sets, the proportion of observations in which the first digit is 1, 2, 3, 4, 5, 6, 7, 8, or 9, respectively, follows the distribution shown to the right.

5.1 Distributions for Quantitative Data: The Histogram

Create a histogram for the data in file *Death*



	A
1	Age at Death (Years)
2	83
3	76
4	78
5	74
6	35
7	78
8	73
9	84
10	55
11	73

Three features need to be defined:

1. The number of nonoverlapping bins
2. The width (numerical range) for each bin
3. The range spanned by the set of bins

Follow these steps to create a histogram for the distribution of the age at death:

Step 1. Select cells A1:A701
Step 2. Click the Insert tab on the Ribbon
Step 3. Click the Insert **Statistic Chart** button in the **Charts** group. When the list of statistic charts appears, select **Histogram**

5.1 Build a Histogram: Definitions and Application

Definitions

Number of bins: between 5 and 20, depending on the number of observations.

Width of the bins: same for all bins.

$$\text{Approximate Bin Width} = \frac{\text{largest value} - \text{smallest value}}{\text{number of bins}}$$

Range spanned by bins: the first bin should begin at a value such that it includes the smallest value in the data.

Excel functions used:

FREQUENCY(data_array, bins_array)

CONCAT(text1, text2,…)

Application to the *Death* data

Number of bins = 16

Width of bins = (109 – 0) / 16 = 6.81 ≈ 7

Range spanned by bins = (0, 7], (7, 14], …, (105, 112]

Type the following in the *Death* Excel file:

Bin Lower Limit label in cell C1

Bin lower limit values in range C2:C17

Bin Upper Limit label in cell D1

Bin upper limit values in range D2:D17

=FREQUENCY(A2:A701,D2:D17) in cell E2

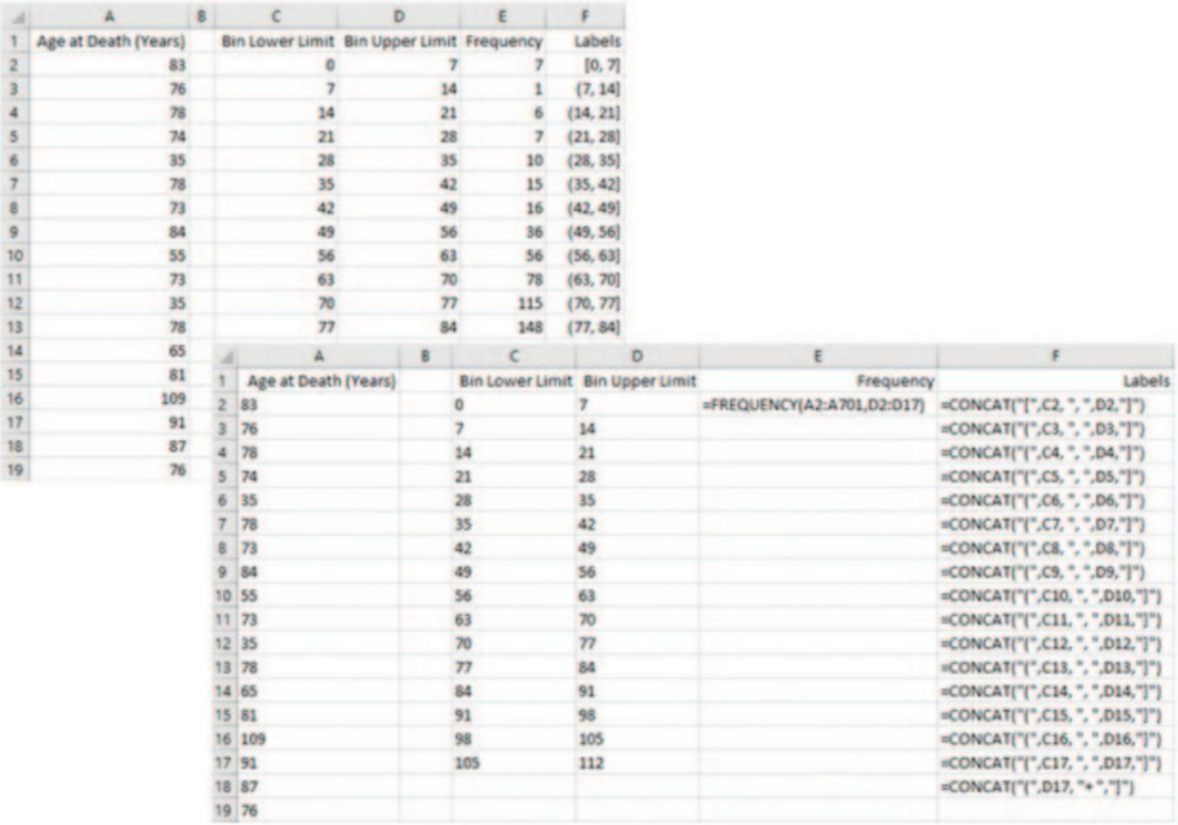
=CONCAT(“(“,C2,”,”,D2,”]”) in cell F2

Copy cell F2 to range F3:F17.

5.1 Example of a Manual Histogram

DATA *file*

DeathFrequency



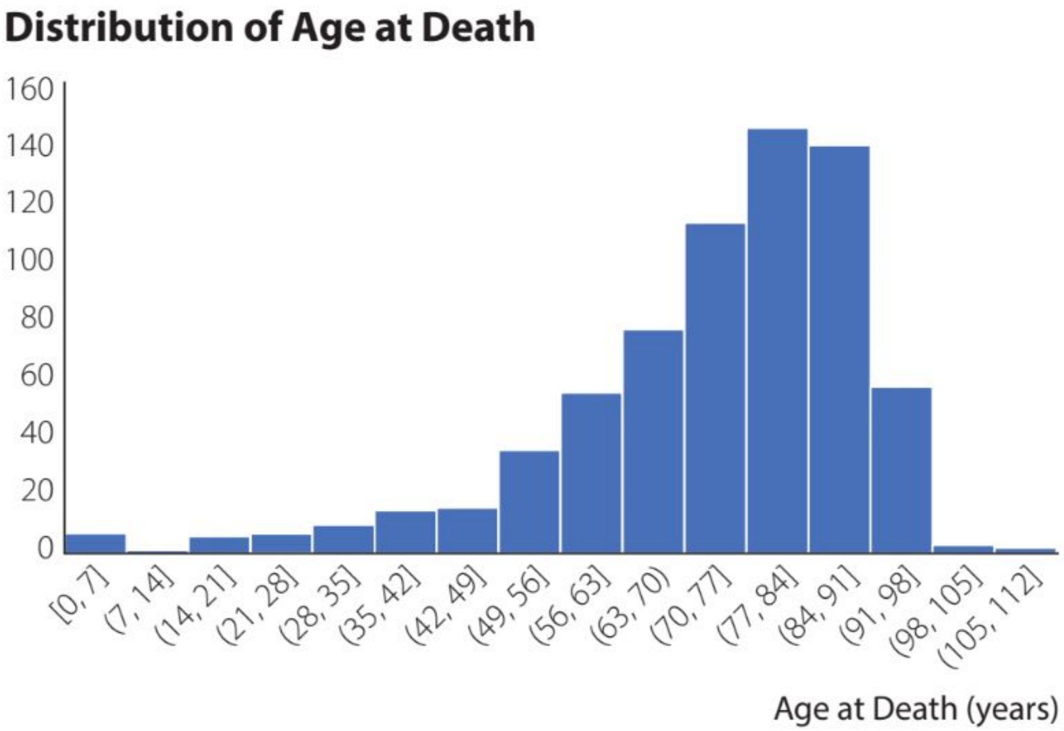
	A	B	C	D	E	F
1	Age at Death (Years)		Bin Lower Limit	Bin Upper Limit	Frequency	Labels
2	83		0	7	7	[0, 7]
3	76		7	14	1	(7, 14]
4	78		14	21	6	(14, 21]
5	74		21	28	7	(21, 28]
6	35		28	35	10	(28, 35]
7	78		35	42	15	(35, 42]
8	73		42	49	16	(42, 49]
9	84		49	56	36	(49, 56]
10	55		56	63	56	(56, 63]
11	73		63	70	78	(63, 70]
12	35		70	77	115	(70, 77]
13	78		77	84	148	(77, 84]
14	65					
15	81					
16	109					
17	91					
18	87					
19	76					

	A	B	C	D	E	F
1	Age at Death (Years)		Bin Lower Limit	Bin Upper Limit	Frequency	Labels
2	83		0	7	=FREQUENCY(A2:A701,D2:D17)	=CONCAT("[",C2, ", ",D2,"]")
3	76		7	14		=CONCAT("(",C3, ", ",D3,"]")
4	78		14	21		=CONCAT("(",C4, ", ",D4,"]")
5	74		21	28		=CONCAT("(",C5, ", ",D5,"]")
6	35		28	35		=CONCAT("(",C6, ", ",D6,"]")
7	78		35	42		=CONCAT("(",C7, ", ",D7,"]")
8	73		42	49		=CONCAT("(",C8, ", ",D8,"]")
9	84		49	56		=CONCAT("(",C9, ", ",D9,"]")
10	55		56	63		=CONCAT("(",C10, ", ",D10,"]")
11	73		63	70		=CONCAT("(",C11, ", ",D11,"]")
12	35		70	77		=CONCAT("(",C12, ", ",D12,"]")
13	78		77	84		=CONCAT("(",C13, ", ",D13,"]")
14	65		84	91		=CONCAT("(",C14, ", ",D14,"]")
15	81		91	98		=CONCAT("(",C15, ", ",D15,"]")
16	109		98	105		=CONCAT("(",C16, ", ",D16,"]")
17	91		105	112		=CONCAT("(",C17, ", ",D17,"]")
18	87					=CONCAT("(",D17, "+ ","]")
19	76					

*See notes for step-by-step instructions on how to build a manual histogram.

CENGAGE

5.1 Histogram for the *Death* Data



A **histogram** is a column chart with no spaces between the columns.

- The columns' height represents the frequency of the corresponding bin.
- An absence of space between the columns reflects the continuous nature of the variable of interest.

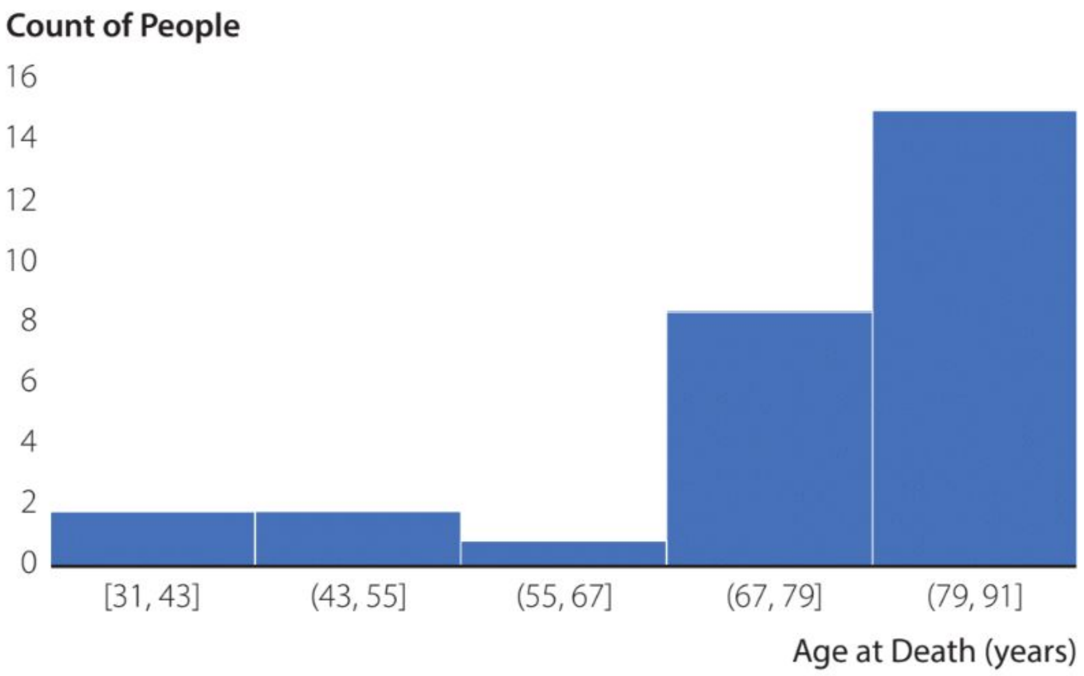
The tallest column corresponds to the bin (77, 84]

- The round parenthesis to the left means greater than (exclusive of 77).
- The square parenthesis to the right means less than or equal to (inclusive of 84).

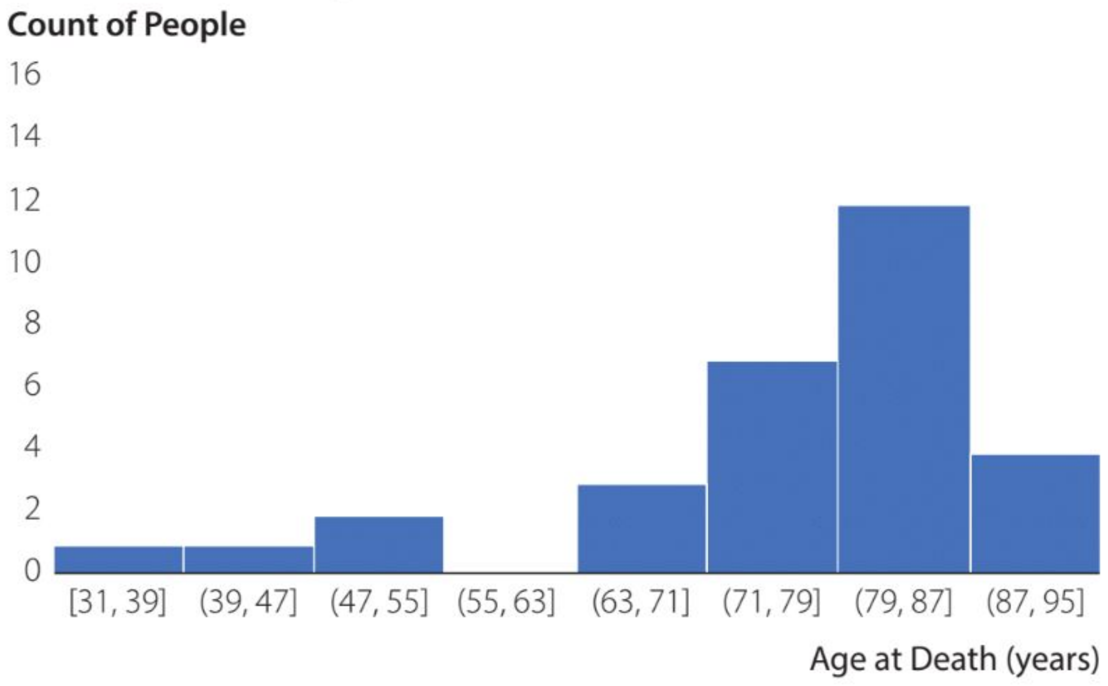
5.1 Effect of the Number of Bins on a Histogram

Varying the number of bins in the histogram for the age at death distribution

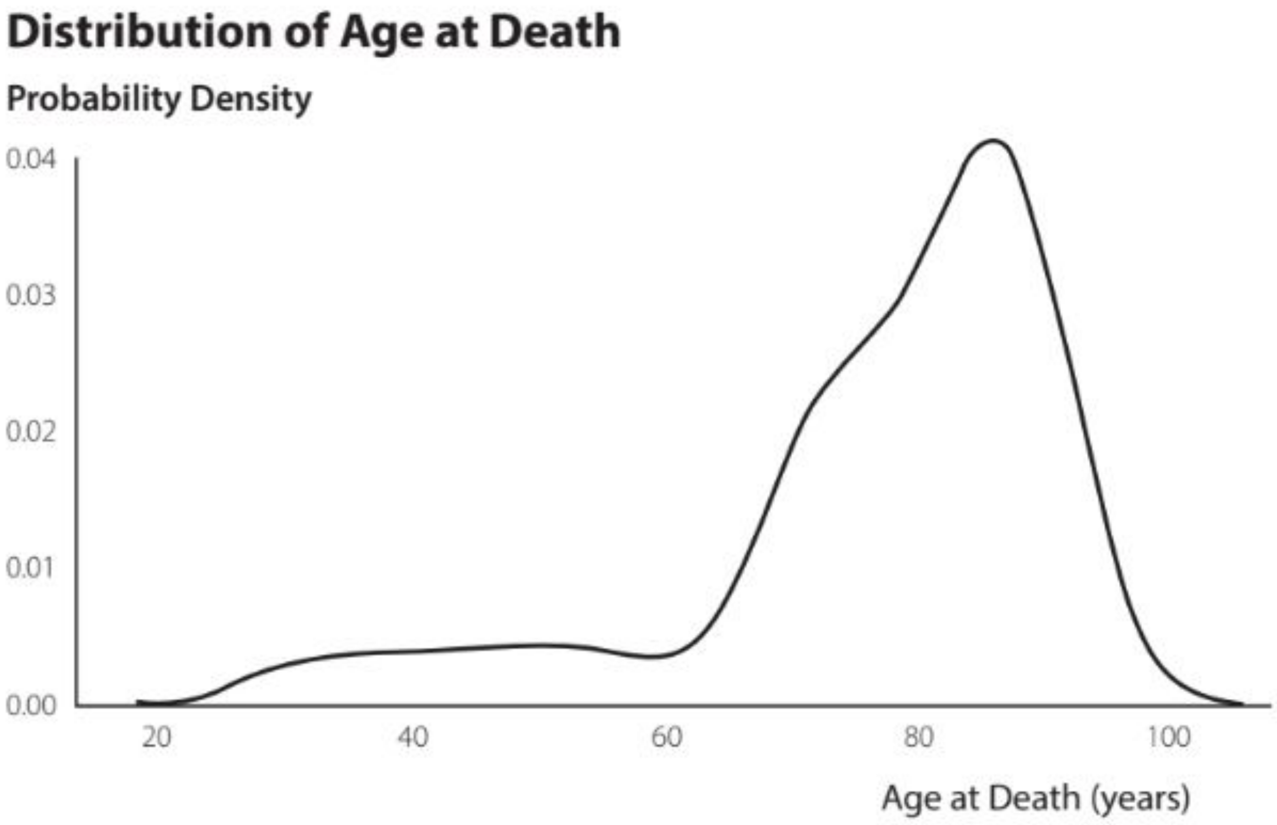
Distribution of Age at Death: 5 Bins



Distribution of Age at Death: 8 Bins



5.1 Additional Charts: the Kernel Density Chart

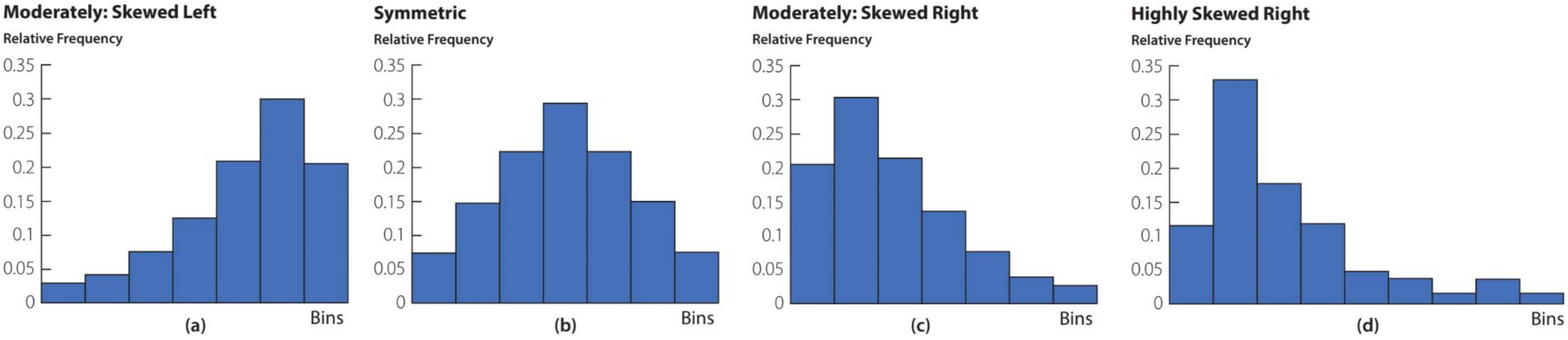


A continuous alternative to a histogram.

It employs a smoothing technique known as kernel density estimation.

Not available in Excel.

5.1 Skewness in a Quantitative Distribution



Skewness represents the lack of symmetry in a quantitative distribution.

a. In a skewed left histogram, the left tail extends farther than the right one (example: exam scores)

b. In a symmetric histogram, the two tails mirror each other (example: SAT scores)

c. In a skewed right histogram, the right tail extends farther than the left one (example: housing prices)

d. In a highly skewed histogram, one of the tails extends much farther than the other one (example: data on wealth and salaries are usually highly skewed right)

5.1 Comparing Distributions: the Frequency Polygon

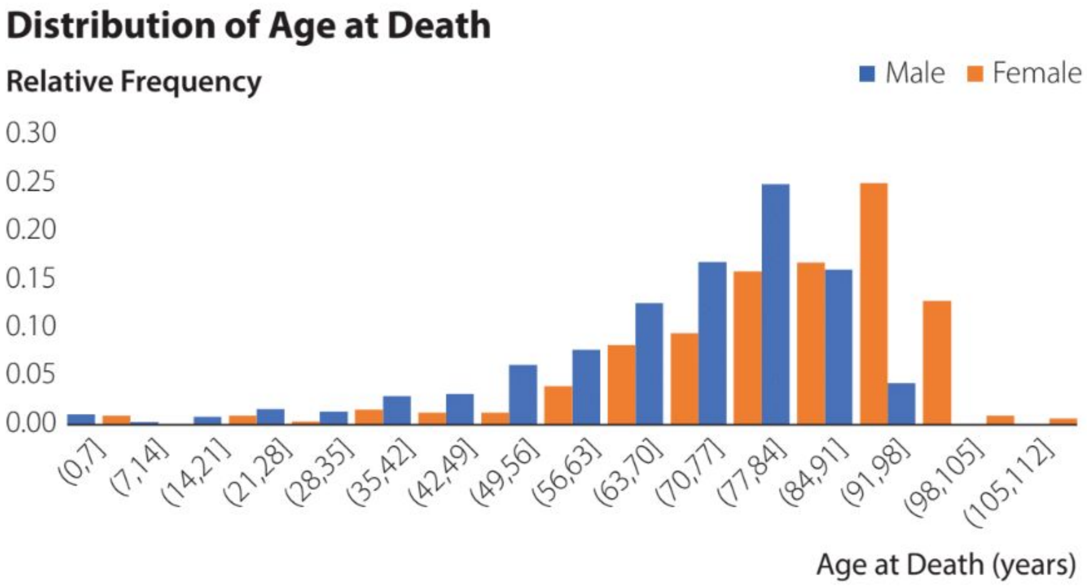
A **frequency polygon** is a visualization tool useful for comparing distributions.

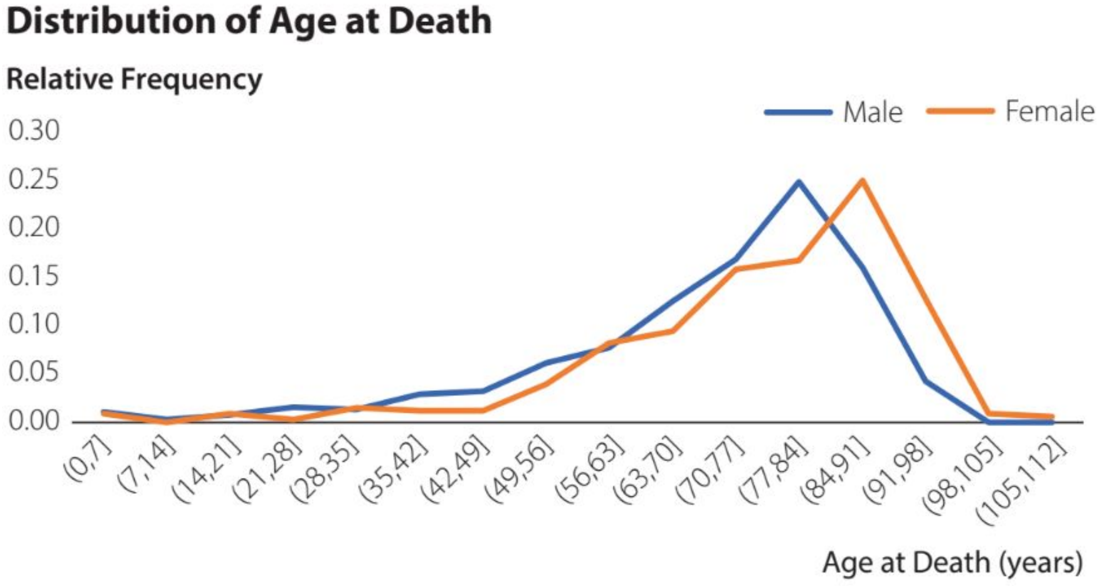
- Like a histogram, a frequency polygon plots the count of observations in a set of bins.
- Unlike a histogram, a frequency polygon uses lines instead of columns to connect the counts of different bins.
- In Excel, use the **Line Chart** option to create a frequency polygon.

*See notes for step-by-step instructions on how to create frequency distributions to build histograms and frequency polygons for the variables *Male* and *Female* in the file *DeathTwo*.

5.1 Comparing Displays for Quantitative Distributions

Comparing a clustered column display to a frequency polygon display

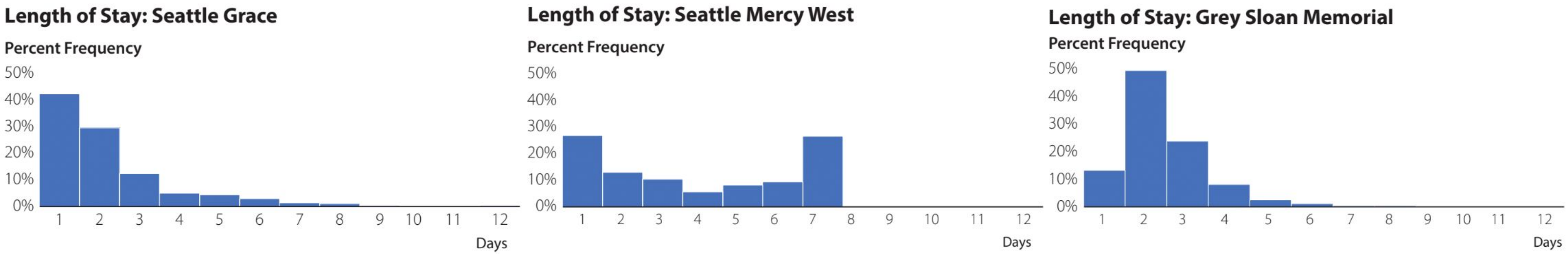




*See notes for step-by-step instructions on how to build the clustered column and frequency polygon displays.

5.1 Comparing Quantitative Distributions: Trellis Display

A Trellis display of the length of stay distributions for three hospitals

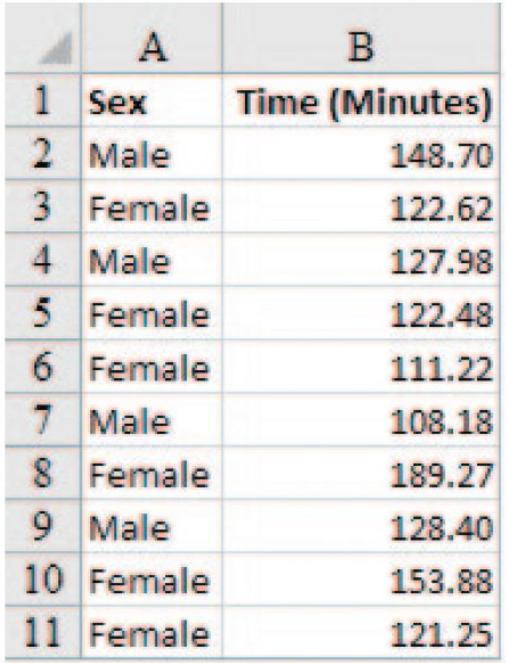


A **trellis display** is a vertical or horizontal arrangement of individual charts of the same type, size, scale, and formatting that differ only by the data they display.

A trellis display can be useful when comparing three or more distributions that would otherwise appear cluttered if plotted using several frequency polygons on the same chart.

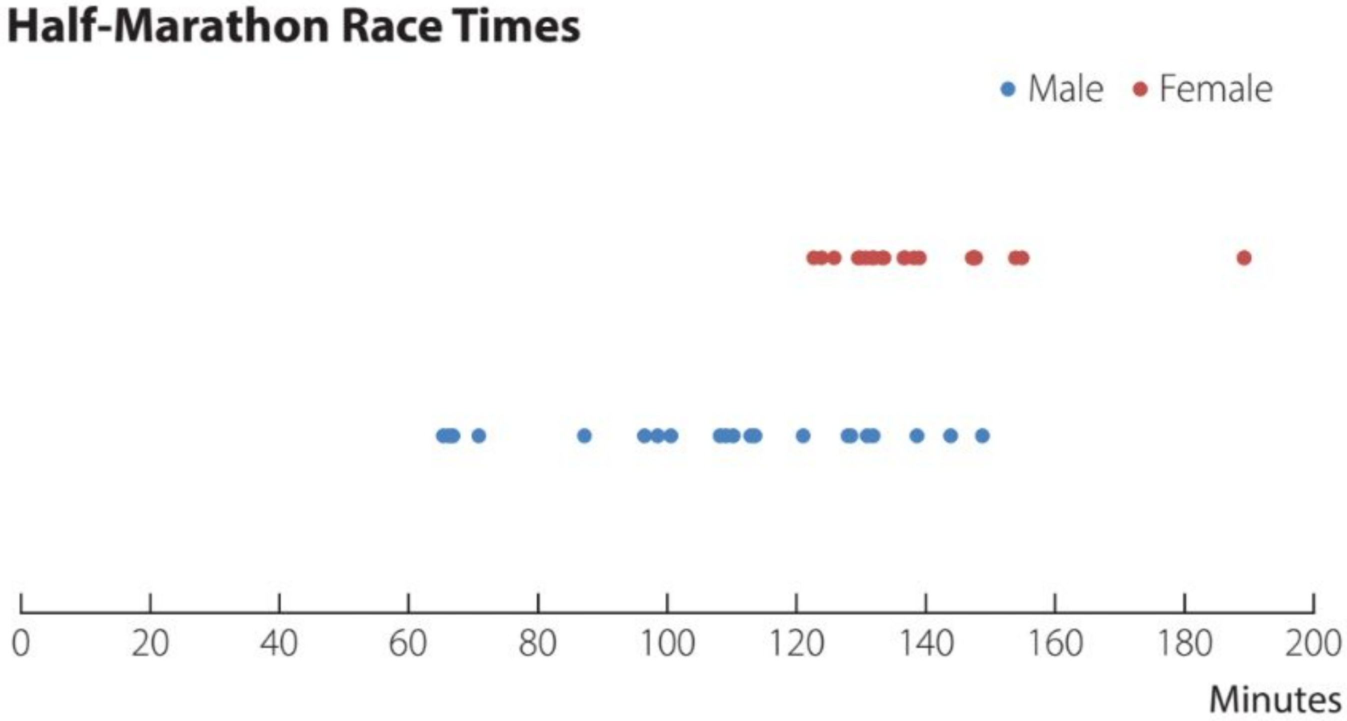
5.1 Displaying Individual Values: the Strip Chart

Portion of data in the file *HalfMarathon*



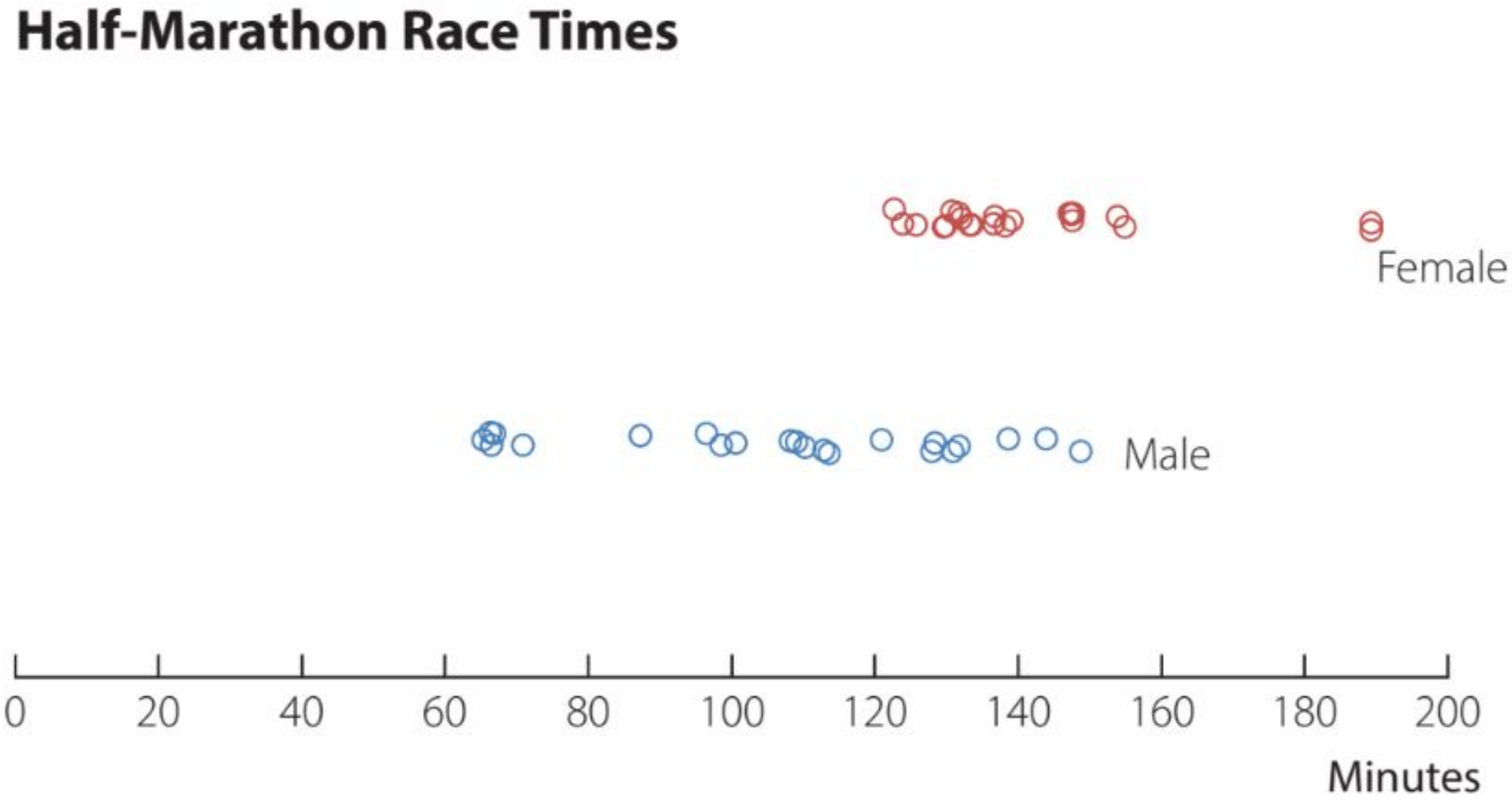
	A	B
1	Sex	Time (Minutes)
2	Male	148.70
3	Female	122.62
4	Male	127.98
5	Female	122.48
6	Female	111.22
7	Male	108.18
8	Female	189.27
9	Male	128.40
10	Female	153.88
11	Female	121.25

Strip chart of female and male half-marathon race times



*See step-by-step instructions on how to build a strip chart in the notes.

5.1 A Remedy to Occlusion: the Jittered Strip Chart



*See step-by-step instructions on how to build a jittered strip chart in the notes.

5.2 Measures of Central Location: Definitions

Definitions

Mean

- Sum of the values divided by the sample size
- AVERAGE function in Excel

Median

- Sort the data in ascending order:
  - If the sample size is even, take the average of the two middle points
  - If the sample size is odd, take the middle value
- MEDIAN function in Excel

Mode

- Most frequent value(s) in the data set
- MODE.MULT function in Excel

Application to the *CincySales* data

Mean (*see notes for details on the equation)

$$\frac{456{,}400 + 298{,}000 + \cdots + 108{,}000}{12} = \$219{,}950$$

=AVERAGE(A2:A13) = $219,950

Median

$$\frac{\$208{,}000 + \$199{,}500}{2} = \$203{,}750$$

=MEDIAN(A2:A13) = $203,750

Mode

$138,000 and $254,000 both occur twice

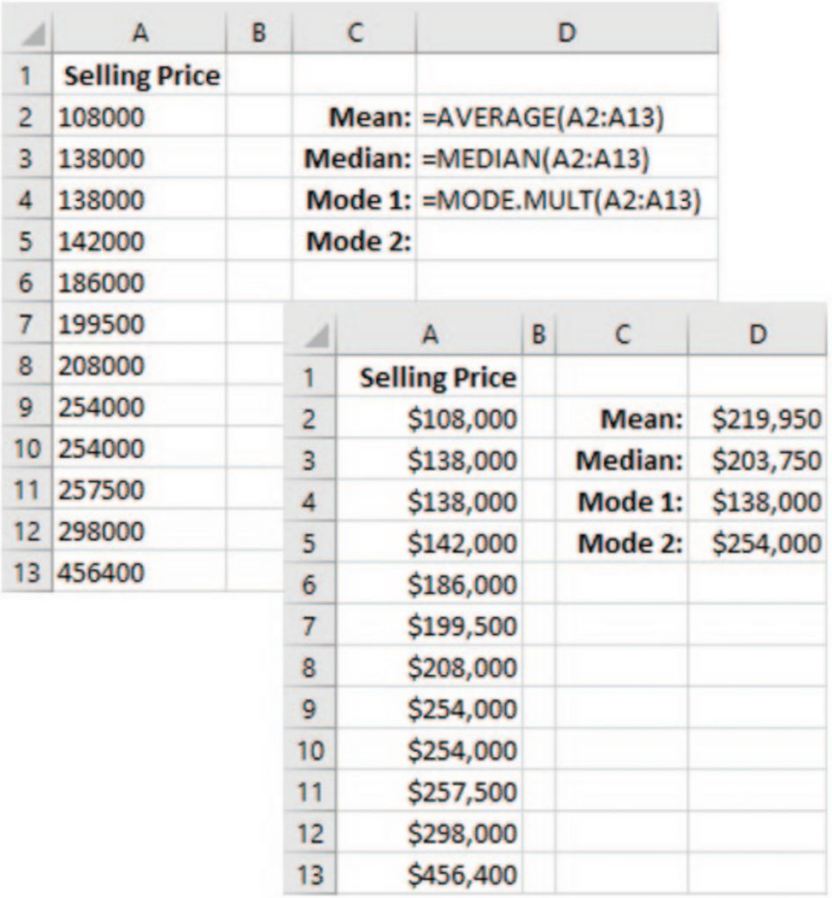
=MODE.MULT(A2:A13) = $138,000 and $254,000

CENGAGE

5.2 Measures of Central Location: Application

DATA file

CincySales



	A	B	C	D
1	Selling Price			
2	108000		Mean:	=AVERAGE(A2:A13)
3	138000		Median:	=MEDIAN(A2:A13)
4	138000		Mode 1:	=MODE.MULT(A2:A13)
5	142000		Mode 2:	
6	186000			
7	199500			
8	208000			
9	254000			
10	254000			
11	257500			
12	298000			
13	456400			

	A	B	C	D
1	Selling Price			
2	$108,000		Mean:	$219,950
3	$138,000		Median:	$203,750
4	$138,000		Mode 1:	$138,000
5	$142,000		Mode 2:	$254,000
6	$186,000			
7	$199,500			
8	$208,000			
9	$254,000			
10	$254,000			
11	$257,500			
12	$298,000			
13	$456,400			

5.2 Measures of Variability: Range and Standard Deviation

Definitions

Range

- Largest value minus smallest value in the set
- MAX minus MIN functions in Excel

Standard Deviation

- Based on average deviation from the mean
- STDEV.S function in Excel

Application to the *CincySales* data

Range

$$\$456{,}400 - \$108{,}000 = \$348{,}400$$

=MAX(A3:A12) – MIN(A3:A12) = $348,400

Standard Deviation (*see notes for details on the equation)

$$\sqrt{\frac{(456{,}400 - 219{,}950)^2 + (298{,}000 - 219{,}950)^2 + \cdots}{12 - 1}} =$$

=STDEV.S(A3:A12) = $95,100

5.2 Measures of Variability: Percentiles

Definitions

Percentile

- The *p*th percentile is a value that exceeds *p%* of the observations in the set
- PERCENTILE.EXT function in Excel

Quartile

Q1 = 1st quartile = 25th percentile

2nd quartile = 50th percentile = median

Q1 = 3rd quartile = 75th percentile

Interquartile Range (IQR)

- Q3 – Q1
- 75th percentile minus 25th percentile

Application to the *CincySales* data

Percentile

Location 25th percentile = $\frac{25}{100} \times (12+1) = 3.25$

25th percentile $= \$138{,}000 + (3.25 - 3) \times (\$142{,}000 - \$138{,}000) = \$139{,}000$

=PERCENTILE.EXT(A3:A12,0.25) = $139,000

Location 75th percentile = $\frac{75}{100} \times (12+1) = 9.75$

75th percentile $= \$254{,}000 + (9.75 - 9) \times (\$257{,}500 - \$254{,}000) = \$256{,}625$

=PERCENTILE.EXT(A3:A12,0.75) = $256,625

Interquartile Range (IQR)

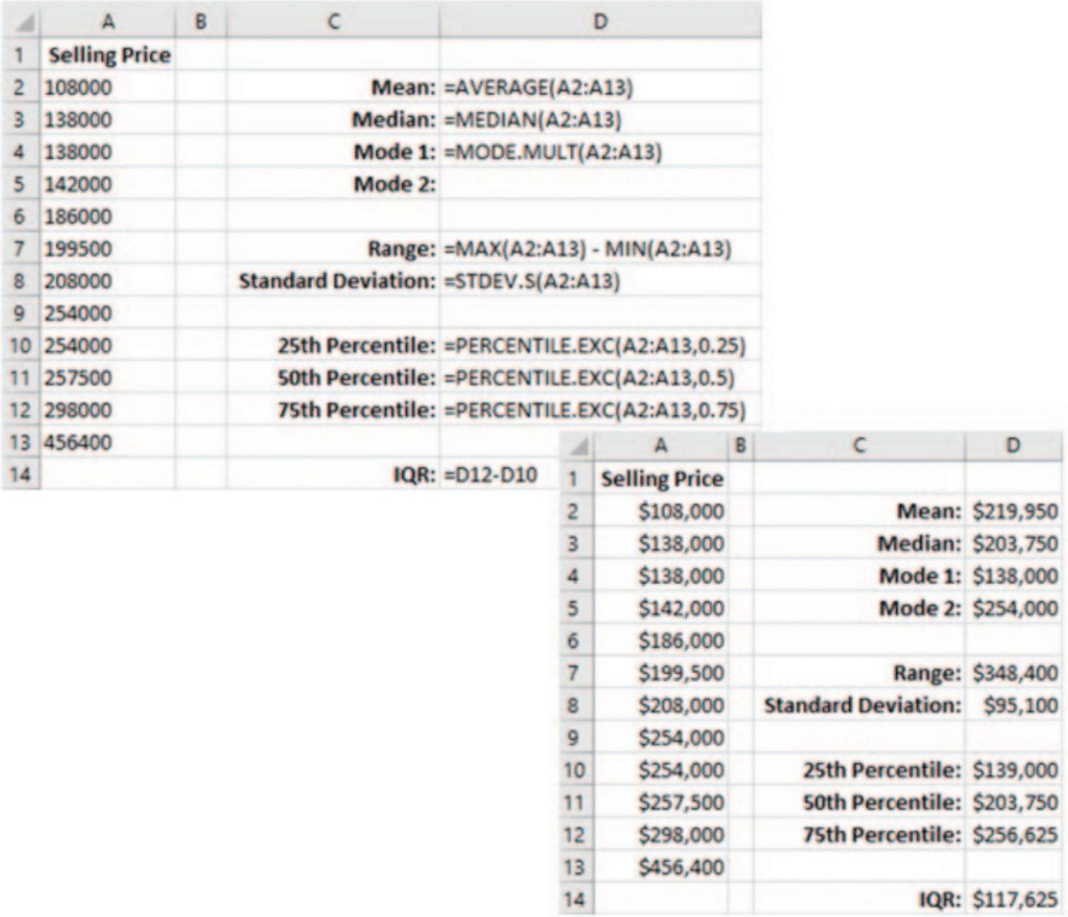
$Q3 - Q1 = \$256{,}625 - \$139{,}000 = \$117{,}625$

CENGAGE

5.2 Measures of Variability: Application

DATA file

CincySales

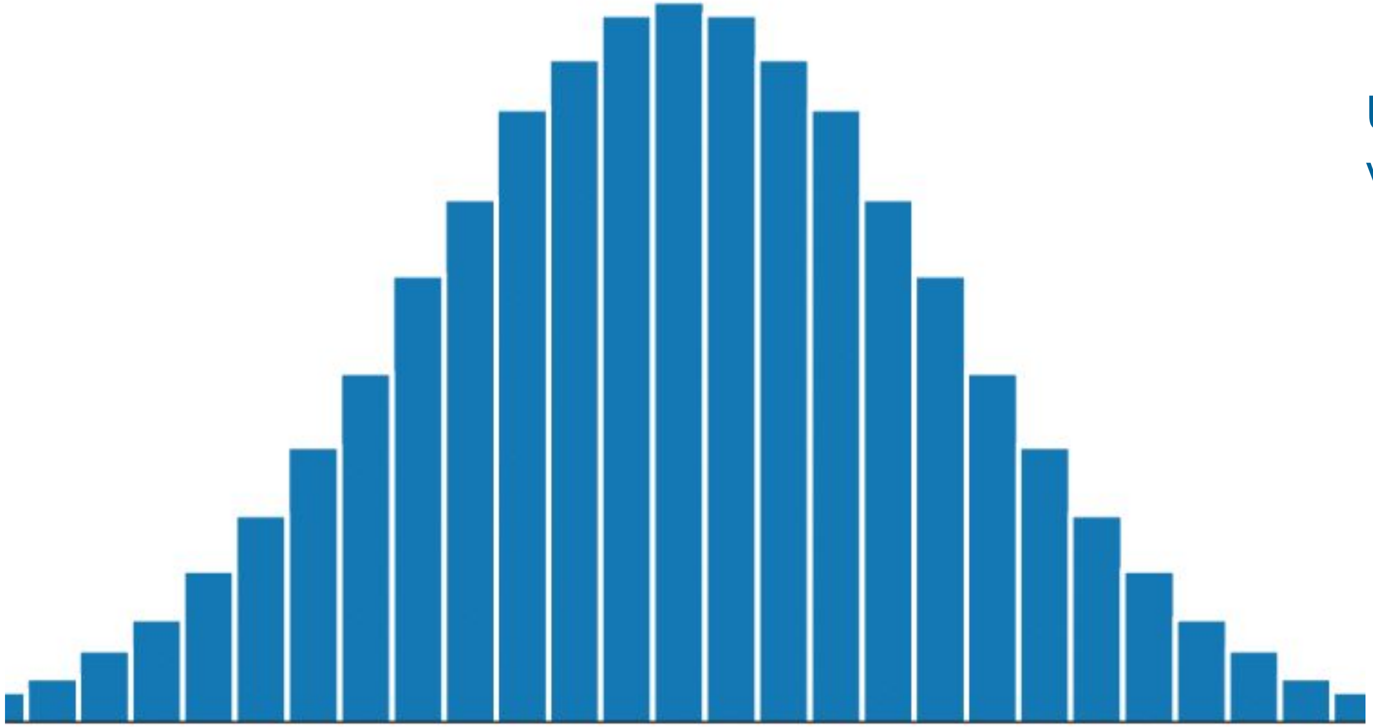


	A	B	C	D
1	Selling Price			
2	108000		Mean:	=AVERAGE(A2:A13)
3	138000		Median:	=MEDIAN(A2:A13)
4	138000		Mode 1:	=MODE.MULT(A2:A13)
5	142000		Mode 2:	
6	186000			
7	199500		Range:	=MAX(A2:A13) - MIN(A2:A13)
8	208000		Standard Deviation:	=STDEV.S(A2:A13)
9	254000			
10	254000		25th Percentile:	=PERCENTILE.EXC(A2:A13,0.25)
11	257500		50th Percentile:	=PERCENTILE.EXC(A2:A13,0.5)
12	298000		75th Percentile:	=PERCENTILE.EXC(A2:A13,0.75)
13	456400			
14			IQR:	=D12-D10

	A	B	C	D
1	Selling Price			
2	$108,000		Mean:	$219,950
3	$138,000		Median:	$203,750
4	$138,000		Mode 1:	$138,000
5	$142,000		Mode 2:	$254,000
6	$186,000			
7	$199,500		Range:	$348,400
8	$208,000		Standard Deviation:	$95,100
9	$254,000			
10	$254,000		25th Percentile:	$139,000
11	$257,500		50th Percentile:	$203,750
12	$298,000		75th Percentile:	$256,625
13	$456,400			
14			IQR:	$117,625

5.2 The Empirical Rule for Bell-Shaped Distributions

Bell-shaped distribution (symmetric, single-mode)



Using the standard deviation to describe variability:

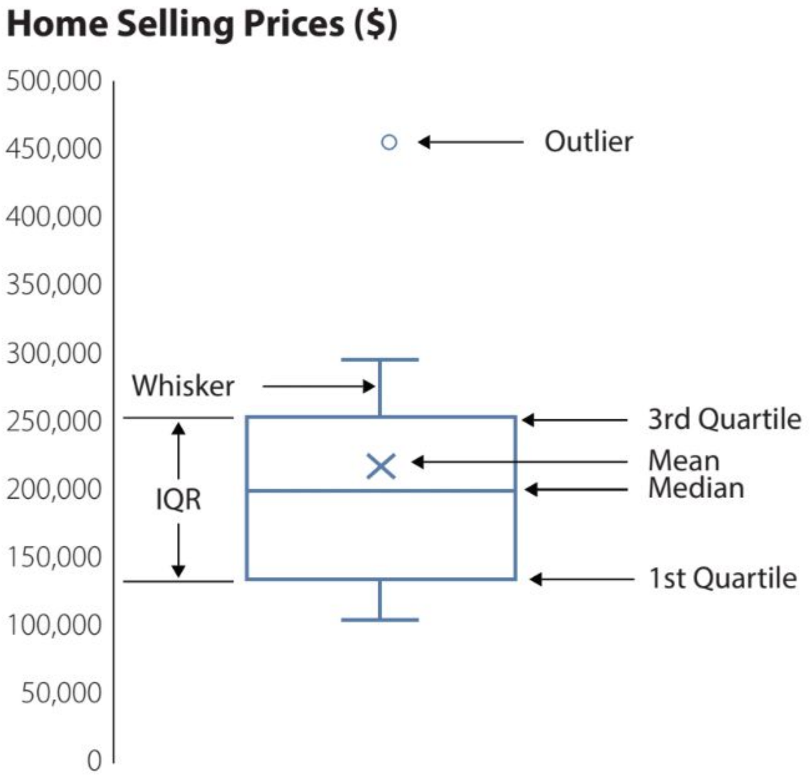
≈ 68% of data values lie within one standard deviation of the mean

≈ 95% of data values lie within two standard deviations of the mean

≈ 99.7% of data values lie within three standard deviations of the mean

5.2 Box and Whisker Charts: Definitions

Box and whisker chart for *CincySales* data



Outliers are values outside the range:

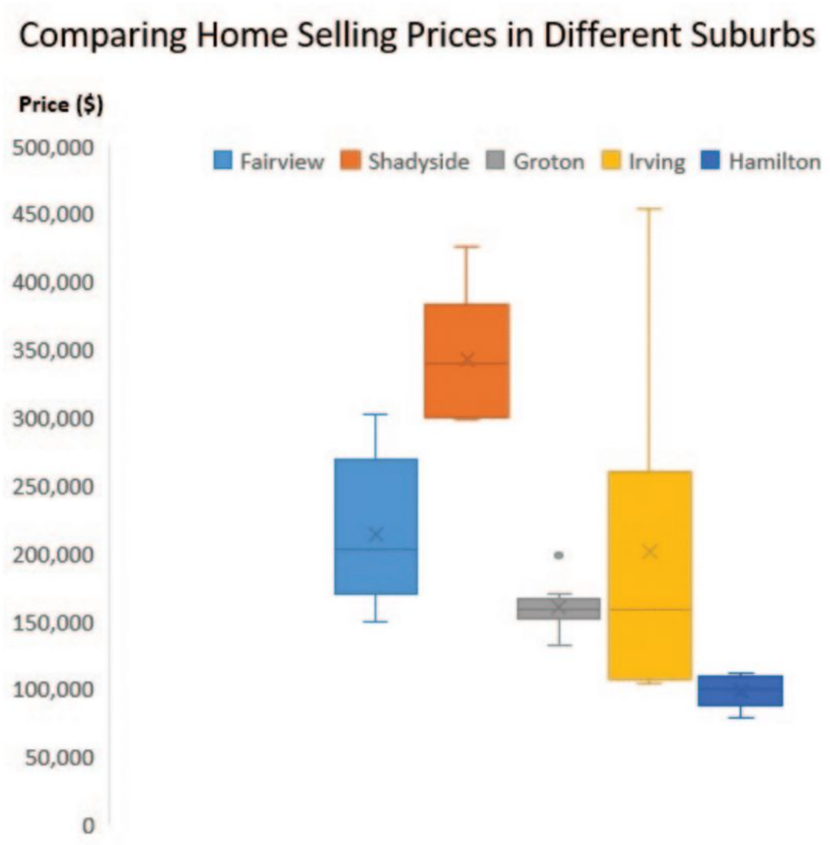
$$[Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR]$$

To build a box and whisker chart for *CincySales* data, follow these steps:

Step 1. Select cells *A1:A13*
Step 2. Click the **Insert** tab on the Ribbon
Step 3. Click the **Insert Statistic Chart** button in the **Charts** group

When the list of statistic charts appears, select **Box and Whisker**

5.2 Box and Whisker Chart to Compare Multiple Variables



We can use the *SalesComparison* file to demonstrate the construction of box and whisker charts for multiple variables:

Step 1. Select cells *B1:F11*

Step 2. Click the **Insert** tab on the Ribbon

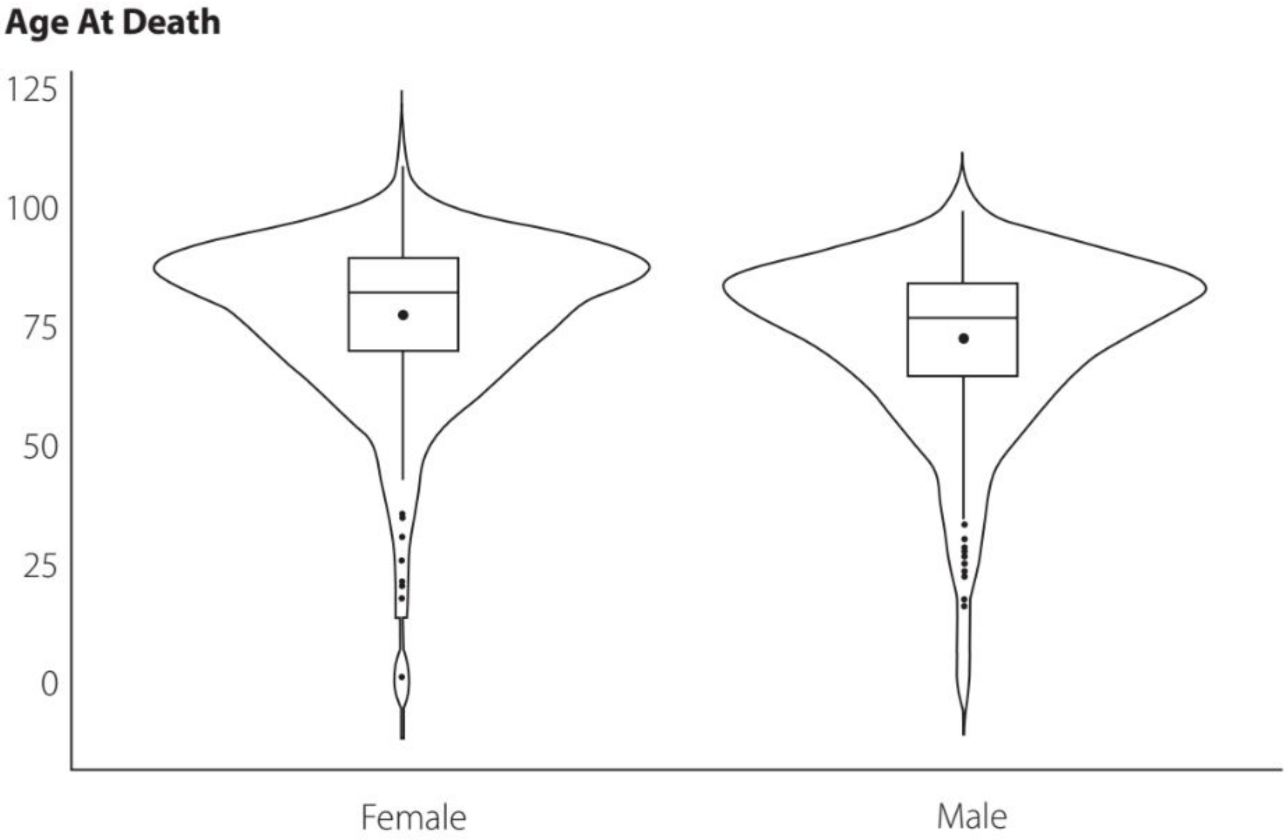
Step 3. Click the **Insert Statistic Chart** button in the **Charts** group

When the list of statistic charts appears, select **Box and Whisker**

We can now make several observations, among which:

- Shadyside has the highest home selling prices, while Hamilton has the lowest.
- Irving and Groton have the same median selling price, but Irving houses have a higher price variability.

5.2 Additional charts: the Violin Chart



A **Violin chart** is an advanced visualization that combines the statistical descriptors of a box and whisker chart with a rotated and mirrored kernel density chart.

Shown here is an example of a violin chart for the *DeathTwo* data.

The violin chart is not available in Excel.

5.3 Statistical Inference

Statistical inference is the process of collecting sample data to make estimates of or draw conclusions about one or more characteristics of a population.

Examples of statistical inference:

- Use the proportion of support for a candidate to the U.S. Senate out of a sample of registered voters in Texas to make an inference about the population proportion.
- Collect a sample of weekly grocery bills to estimate the average amount of money spent on groceries by a target population of potential customers of a grocery delivery service.

5.3 Confidence Interval

A **confidence interval** is a parameter estimate such as the mean or the proportion of a population of interest.

The confidence interval on a mean:

sample mean ± margin of error

The confidence interval on a proportion:

sample proportion ± margin of error

The **margin of error** represents the uncertainty on the parameter estimate at a given *confidence level*, such as 95% or 99%.

5.3 The Confidence Interval on a Mean: Definitions

The margin of error for a confidence interval on a mean depends on three factors:

1) The confidence level
2) The variability of sample values (the standard deviation)
3) The sample size

In Excel:

margin of error = CONFIDENCE.T(significance level, std dev, sample size)

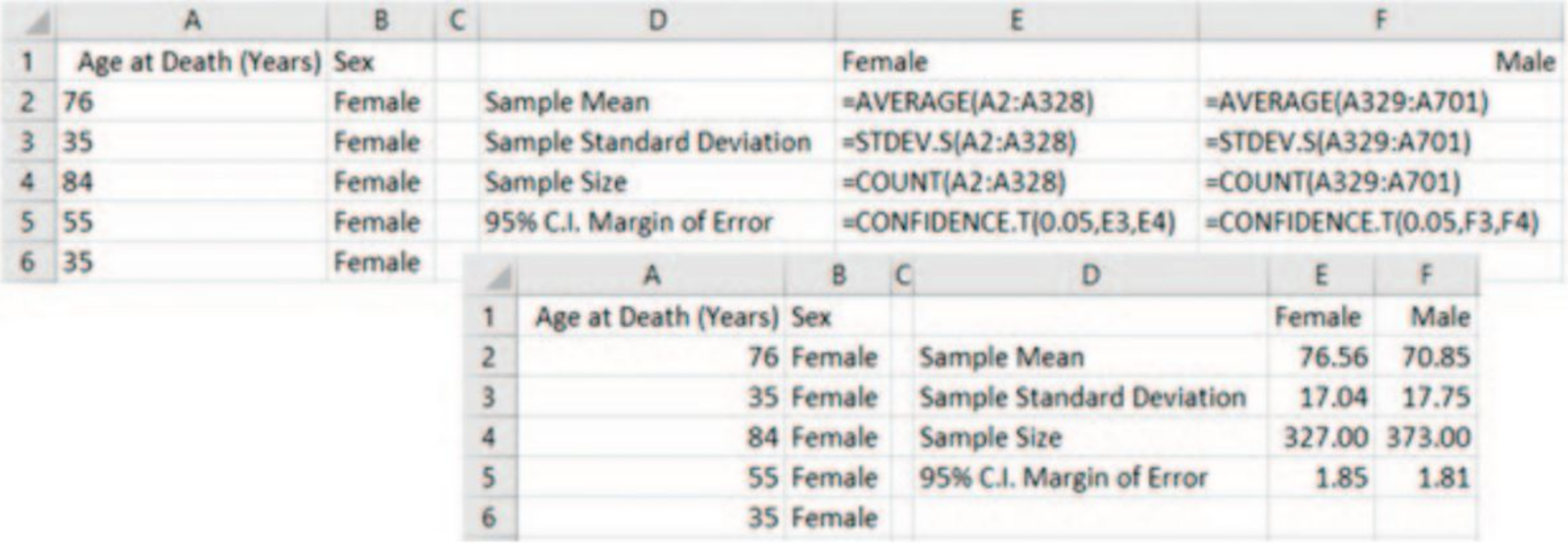
Where the significance level is the complement of the confidence level.

- Example: If the confidence level is 95%, the significance level is 1 – 95% = 0.05

CENGAGE

5.3 The Confidence Interval on a Mean: Calculations

Calculations for the confidence interval on a mean using the *DeathAvgAgeChart* data



	A	B	C	D	E	F
1	Age at Death (Years)	Sex			Female	Male
2	76	Female		Sample Mean	=AVERAGE(A2:A328)	=AVERAGE(A329:A701)
3	35	Female		Sample Standard Deviation	=STDEV.S(A2:A328)	=STDEV.S(A329:A701)
4	84	Female		Sample Size	=COUNT(A2:A328)	=COUNT(A329:A701)
5	55	Female		95% C.I. Margin of Error	=CONFIDENCE.T(0.05,E3,E4)	=CONFIDENCE.T(0.05,F3,F4)
6	35	Female				

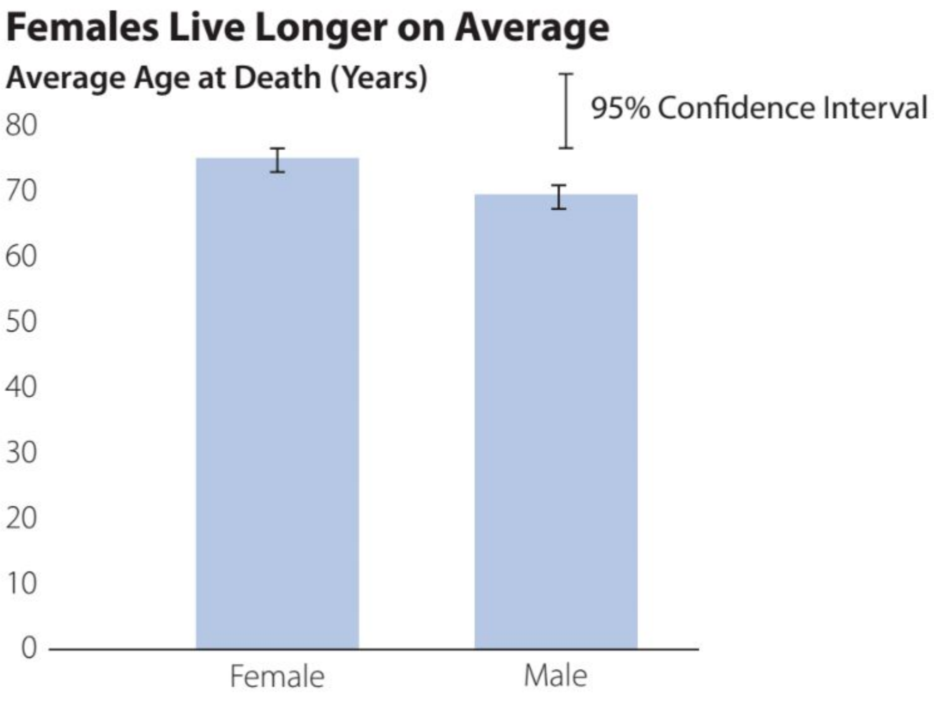
	A	B	C	D	E	F
1	Age at Death (Years)	Sex			Female	Male
2	76	Female		Sample Mean	76.56	70.85
3	35	Female		Sample Standard Deviation	17.04	17.75
4	84	Female		Sample Size	327.00	373.00
5	55	Female		95% C.I. Margin of Error	1.85	1.81
6	35	Female				

We are 95% confident that the overall population's mean age at death is (see notes for details):

- For females: $76.56 \pm 1.85 = [74.71, 78.41]$
- For males: $70.85 \pm 1.81 = [69.04, 72.66]$

5.3 Column Charts with Confidence Interval Bars

A column chart with error bars for mean age at death:



To visualize a confidence interval on a mean for the *DeathAvgAgeChart* data, follow these steps:

Step 1. Click on the column chart

Step 2. Click the **Chart Elements** button and select **Error Bars**

Click the black triangle to the right of **Error Bars** and select **More Options ...**

Step 3. When the **Format Error Bars** task pane appears:

Click **Error Bar Options**

In the **Error Amount** area, select **Custom**

Click the **Specify Value** button next to **Custom**

In the **Custom Error Bars** dialog box, enter *=Data!$E$5:$F$5* in both the **Positive Error Value** box and **Negative Error Value** box

*See additional comments in the notes.

CENGAGE

5.3 The Confidence Interval on a Proportion: Definitions

The margin of error for a confidence interval on a proportion depends on three factors:

1) The confidence level
2) The sample proportion
3) The sample size

In Excel:

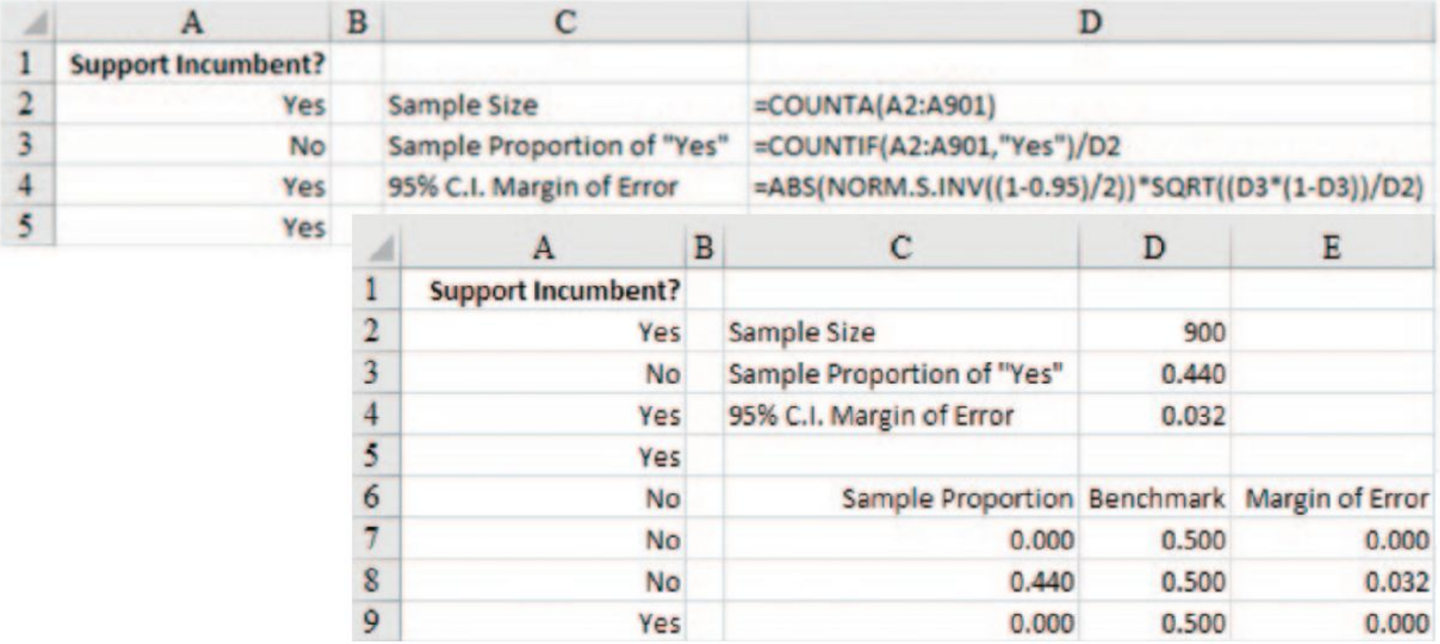
count of favorable outcomes = COUNTIF(data_array, condition)

sample size = COUNTA(data_array)

sample proportion = count of favorable outcomes / sample size

5.3 The Confidence Interval on a Proportion: Calculations

Calculations for the confidence interval on a proportion using the *Incumbent* data



	A	B	C	D
1	Support Incumbent?			
2	Yes		Sample Size	=COUNTA(A2:A901)
3	No		Sample Proportion of "Yes"	=COUNTIF(A2:A901,"Yes")/D2
4	Yes		95% C.I. Margin of Error	=ABS(NORM.S.INV((1-0.95)/2))*SQRT((D3*(1-D3))/D2)
5	Yes			

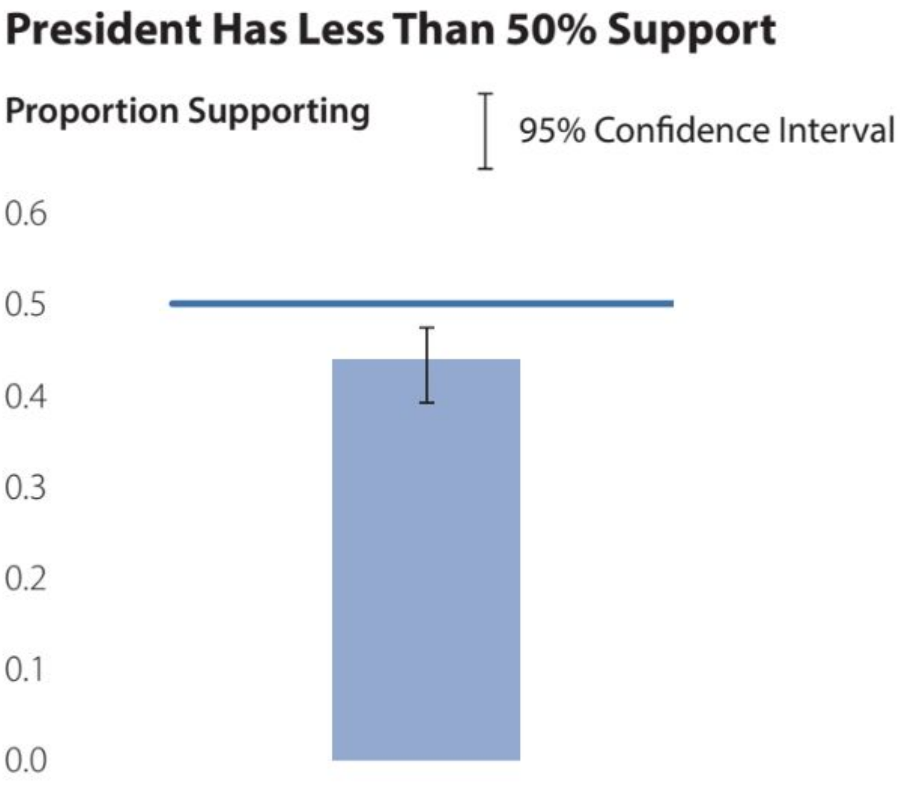
	A	B	C	D	E
1	Support Incumbent?				
2	Yes		Sample Size	900	
3	No		Sample Proportion of "Yes"	0.440	
4	Yes		95% C.I. Margin of Error	0.032	
5	Yes				
6	No		Sample Proportion	Benchmark	Margin of Error
7	No		0.000	0.500	0.000
8	No		0.440	0.500	0.032
9	Yes		0.000	0.500	0.000

We are 95% confident that the proportion of citizens who support the incumbent president is (see notes for details):

$$0.440 \pm 0.032 = [0.408, 0.472]$$

5.3 A Column Chart with Confidence Interval Bars

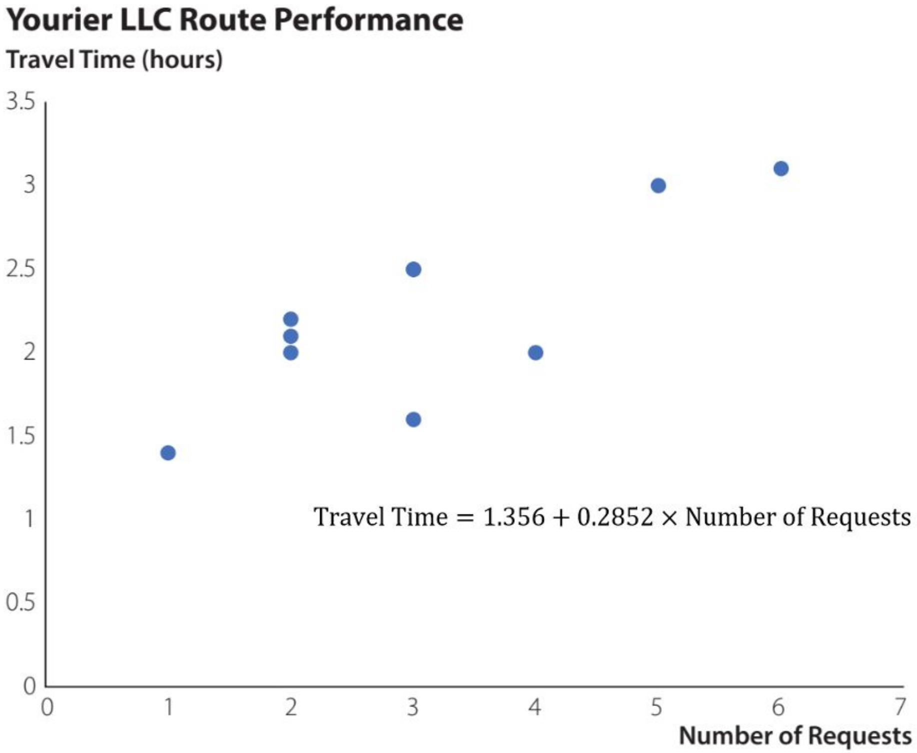
Combo chart with error bars on sample proportion for *Incumbent* data



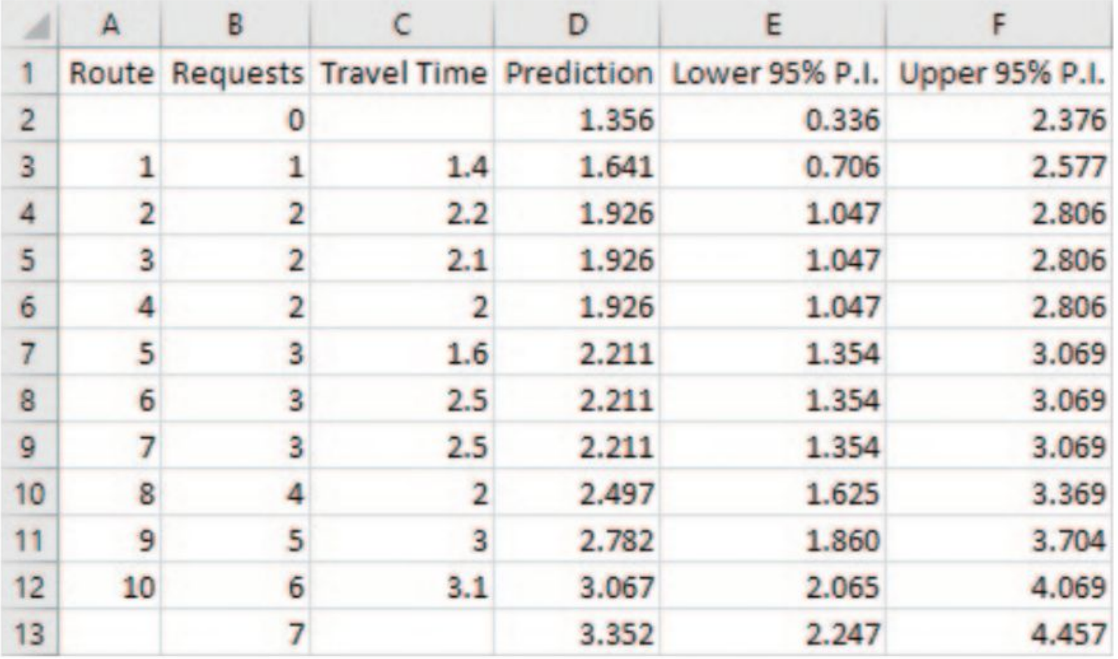
To visualize the confidence interval for a proportion on the *Incumbent* data with a 50% benchmark, see step-by-step instructions in the notes.

5.4 Prediction Intervals for Regression: Data

Scatter chart of travel time versus number of requests



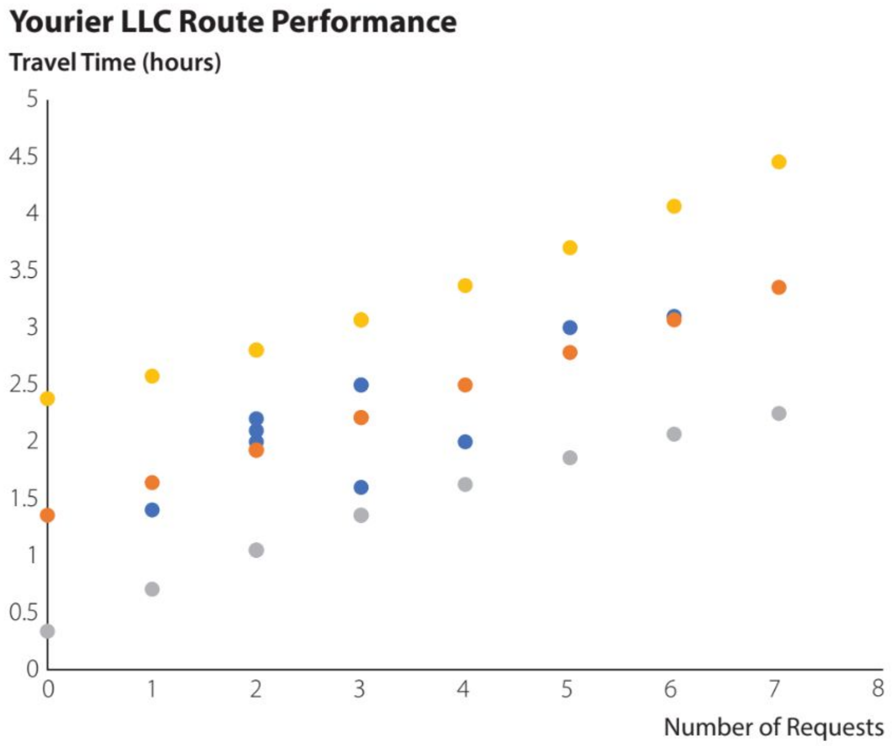
Simple regression predictions and 95% prediction interval limits* for the *Yourier* data



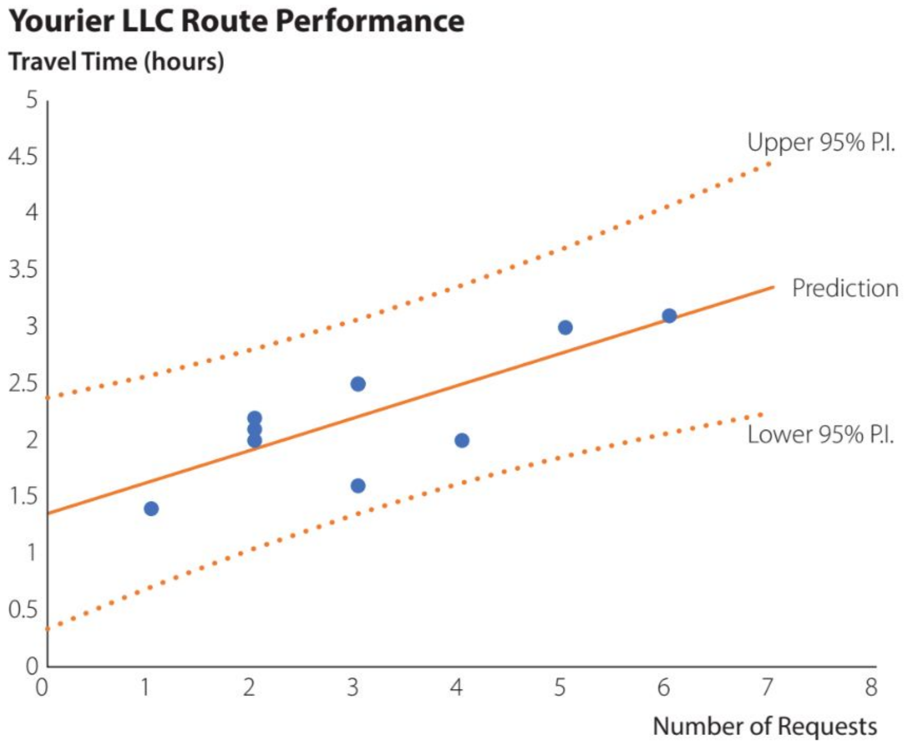
	A	B	C	D	E	F
1	Route	Requests	Travel Time	Prediction	Lower 95% P.I.	Upper 95% P.I.
2		0		1.356	0.336	2.376
3	1	1	1.4	1.641	0.706	2.577
4	2	2	2.2	1.926	1.047	2.806
5	3	2	2.1	1.926	1.047	2.806
6	4	2	2	1.926	1.047	2.806
7	5	3	1.6	2.211	1.354	3.069
8	6	3	2.5	2.211	1.354	3.069
9	7	3	2.5	2.211	1.354	3.069
10	8	4	2	2.497	1.625	3.369
11	9	5	3	2.782	1.860	3.704
12	10	6	3.1	3.067	2.065	4.069
13		7		3.352	2.247	4.457

5.4 Prediction Intervals for Regression: Visualization

Scatter chart of observations, predictions, and prediction interval limits



Combo chart displaying 95% prediction intervals on travel time



*See step-by-step instructions in the notes.

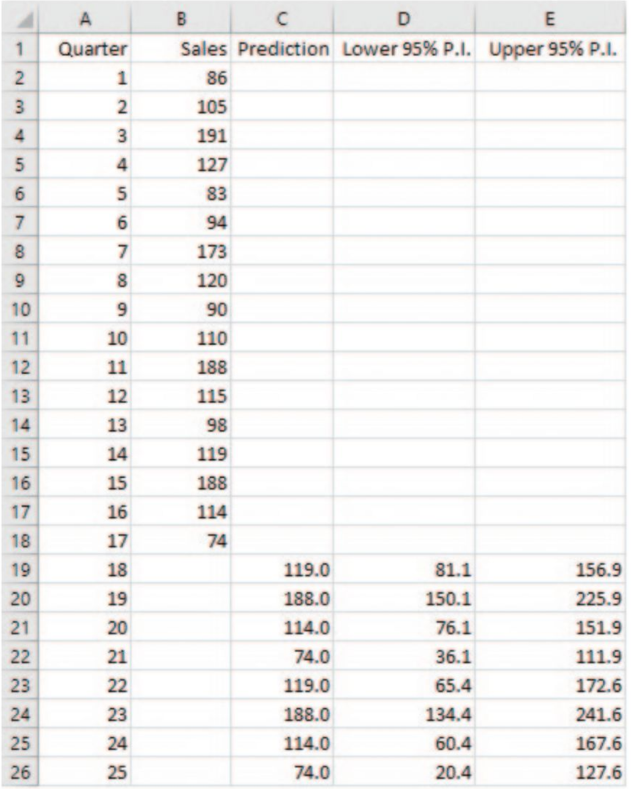
5.4 Time Series

Time series data is a sequence of observations on a variable measured at successive points in time.

A **time series chart** is a line chart with the time unit displayed on the horizontal axis and the values of the variable on the vertical axis.

5.4 Prediction Intervals for a Time Series Model

Time series data and predictions for quarterly root beer sales



	A	B	C	D	E
1	Quarter	Sales	Prediction	Lower 95% P.I.	Upper 95% P.I.
2	1	86			
3	2	105			
4	3	191			
5	4	127			
6	5	83			
7	6	94			
8	7	173			
9	8	120			
10	9	90			
11	10	110			
12	11	188			
13	12	115			
14	13	98			
15	14	119			
16	15	188			
17	16	114			
18	17	74			
19	18		119.0	81.1	156.9
20	19		188.0	150.1	225.9
21	20		114.0	76.1	151.9
22	21		74.0	36.1	111.9
23	22		119.0	65.4	172.6
24	23		188.0	134.4	241.6
25	24		114.0	60.4	167.6
26	25		74.0	20.4	127.6

Follow these steps to visualize the prediction information on the time series chart of quarterly root beer sales using the *BundabergChart* file:

Step 1. In cell C18, enter *=B18*

Step 2. Right-click the chart and select **Select Data...**

Step 3. Click the **Add** button in the **Select Data Source** dialog box

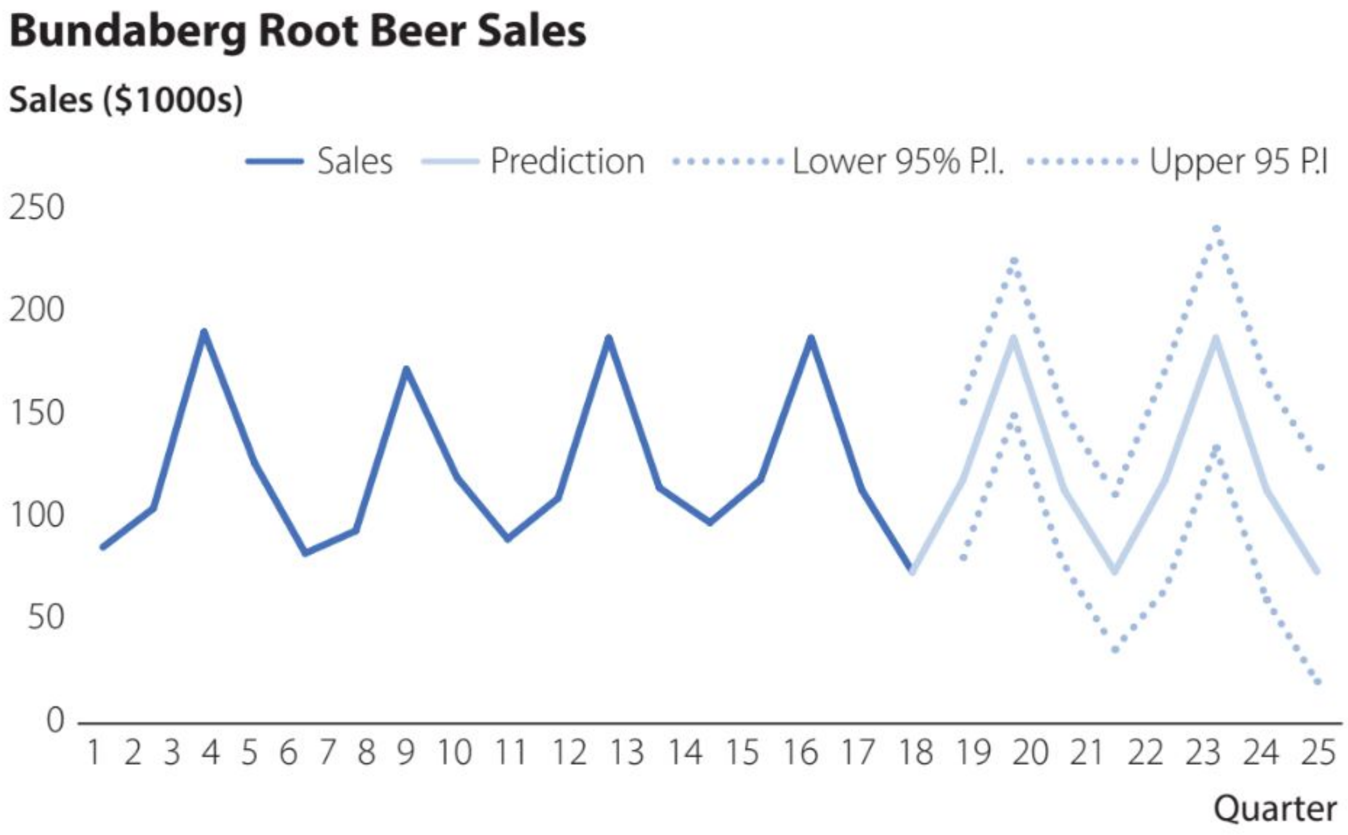
Step 4. In the **Edit Series** dialog box, enter *=Data!$C$1* in the box under **Series name:**, enter *=Data!$C$2:$C$26* in the box under **Series values:**, and click **OK**

Step 5. Repeat Steps 3 and 4, entering *=Data!$D$1* in the box under **Series name:**, and *=Data!$D$2:$D$26* in the box under **Series Y values:**

Step 6. Repeat Steps 3 and 4, entering *=Data!$E$1* in the box under **Series name:**, and *=Data!$E$2:$E$26* in the box under **Series Y values:**

5.4 A Line Chart for a Time Series Model

Time series chart illustrating predictions for future quarterly root beer sales



CENGAGE

Discussion Activity 1

- Consider the four examples of histograms with different levels of skewness shown in slide 16.
  - Using your common sense and your general knowledge, try to guess the expected skewness (either left, symmetric, or right) for the following nine examples.
    1. Length of hospital stay
    2. Birth weight
    3. Family income
    4. Daily stock market returns
    5. Age of death
    6. Female shoe size
    7. Scores in easy exam
    8. Housing prices
    9. Adult male height

CENGAGE

Discussion Activity 2

- Consider the frequency polygon chart in slide 18. The distribution of female and male age at death is skewed left.
    - How could we have used the summary statistics of location, such as mean and median described in slide 23, and the measure of variation such as the quartiles, described in slide 26, to predict the existence of skewness in the distributions of female and male age at death?
    - Besides the charts shown in slide 19, which other data visualization technique you could have used to display the distribution of the female and male age at death data
    - How many outliers either distribution has?

Check Your Knowledge

a. When points in a scatter chart are clustered along a line with a downward slope, what type of relationship is shown?

   a. Negative linear relationship
   b. Positive linear relationship
   c. Nonlinear relationship
   d. No relationship

b. A chart that display a small set of values in a manner that shows the individual values, is called a ________.

   a. strip chart
   b. scatter chart
   c. box plot
   d. prediction chart

CENGAGE

Summary

In this chapter, you should have learned how to:

- Visualize a frequency distribution for a categorical variable with a column chart.
- Visualize a frequency distribution for a quantitative variable in a variety of ways such as a histogram, a frequency polygon, and a jittered strip chart with hollow dots to avoid occlusion.
- Define statistical measures for central location such as mean, median, and mode.
- Define statistical measures for variability such as range, standard deviation, and interquartile range.
- Construct a box and whisker chart and use it to interpret statistical measures.
- Convey uncertainty that arises in:
  - statistical inference using error bars to portray the margin of error.
  - predictive analytics using prediction intervals generated by causal models and forecasts on times series models.