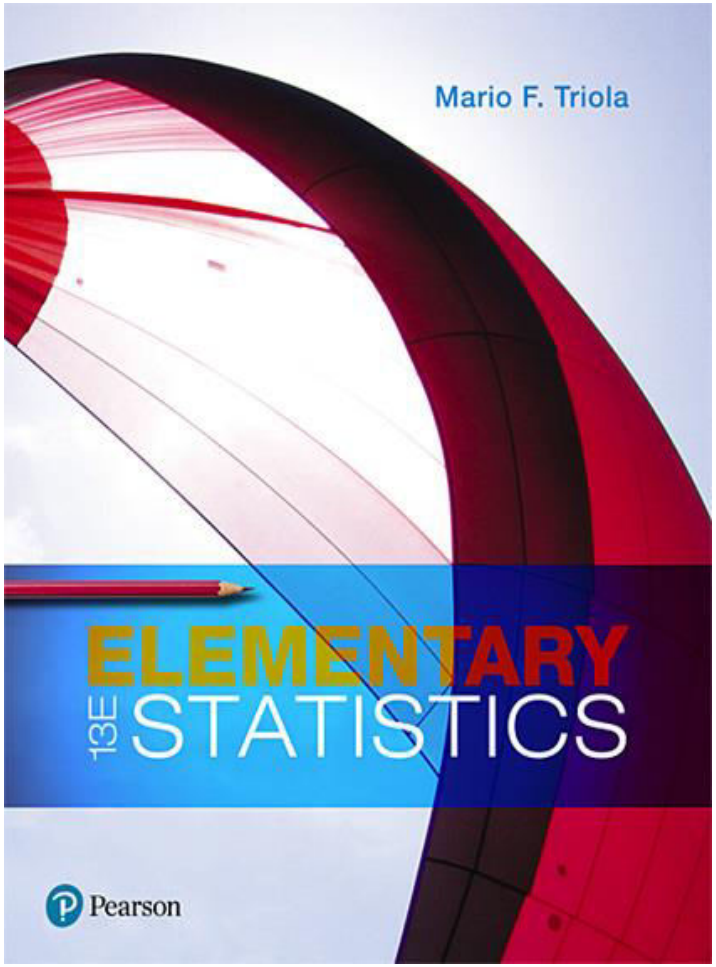
Elementary Statistics

Thirteenth Edition



Chapter 2

Exploring Data with Tables and Graphs

Exploring Data with Tables and Graphs

2-1 Frequency Distributions for Organizing and Summarizing Data

2-2 Histograms

2-3 Graphs that Enlighten and Graphs that Deceive

2-4 Scatterplots, Correlation, and Regression

Key Concept

Introduce the analysis of **paired** sample data.

Discuss **correlation** and the role of a graph called a **scatterplot,** and provide an introduction to the use of the **linear correlation coefficient.**

Provide a very brief discussion of **linear regression,** which involves the equation and graph of the straight line that best fits the sample paired data.

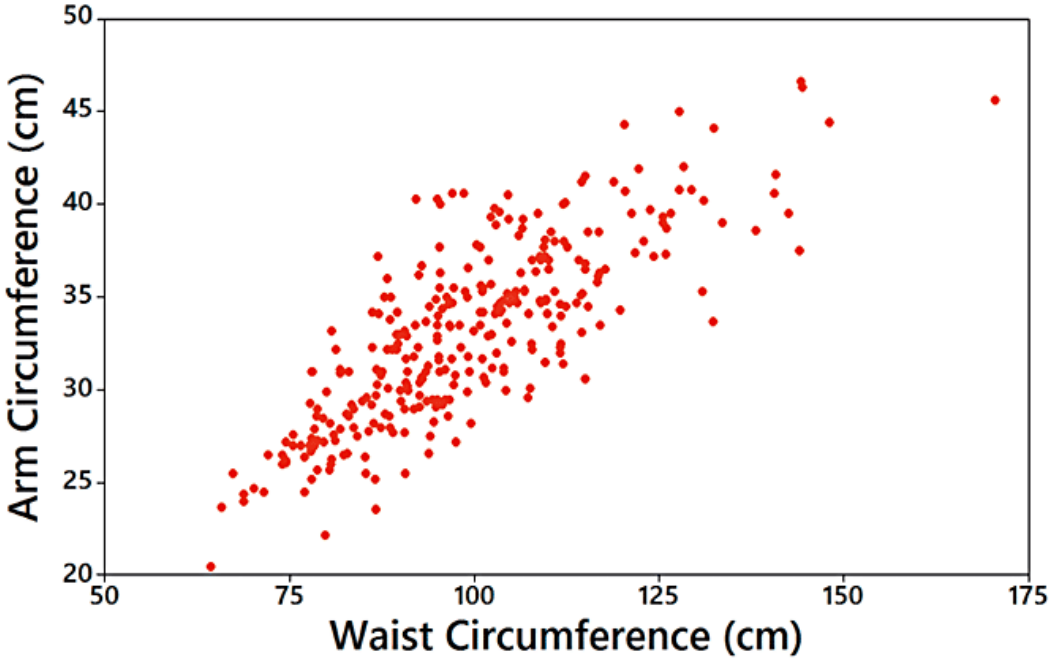
Scatterplot and Correlation (1 of 2)

- **Correlation**
  - A **correlation** exists between two variables when the values of one variable are somehow associated with the values of the other variable.
- **Linear Correlation**
  - A **linear correlation** exists between two variables when there is a correlation and the plotted points of paired data result in a pattern that can be approximated by a straight line.

Scatterplot and Correlation (2 of 2)

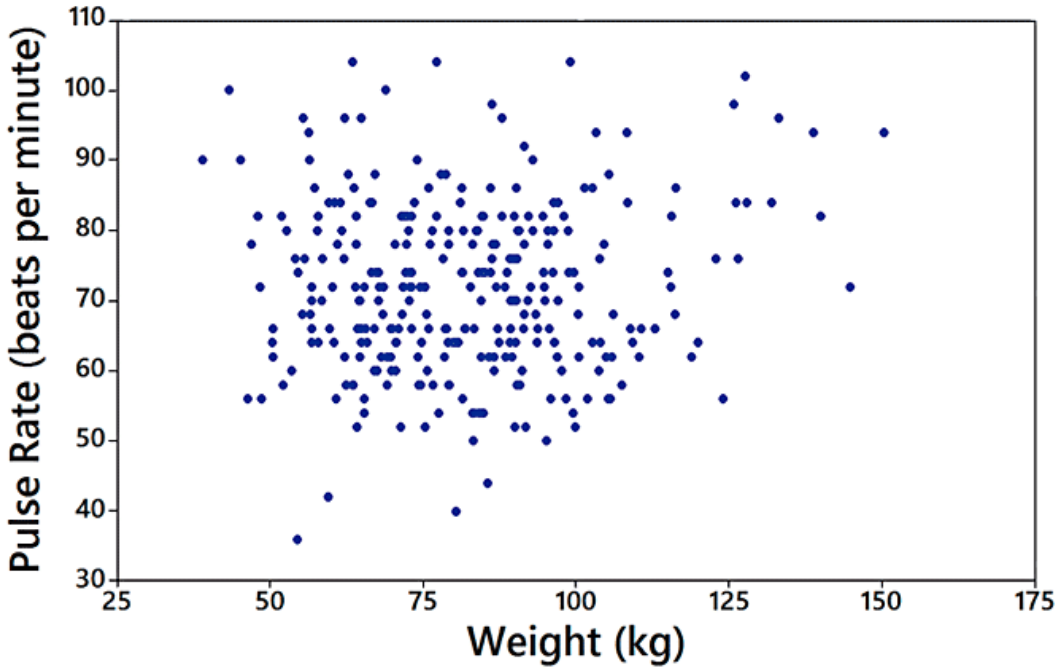
- **Scatterplot** (or **Scatter Diagram**)
  - A **scatterplot** (or **scatter diagram**) is a plot of paired (x, y) quantitative data with a horizontal x-axis and a vertical y-axis. The horizontal axis is used for the first variable (x), and the vertical axis is used for the second variable (y).

Example: Waist and Arm Correlation (1 of 2)



- **Correlation:** The distinct pattern of the plotted points suggests that there is a correlation between waist circumferences and arm circumferences.

Example: Waist and Arm Correlation (2 of 2)



- **No Correlation:** The plotted points do not show a distinct pattern, so it appears that there is no correlation between weights and pulse rates.

Linear Correlation Coefficient *r*

- **Linear Correlation Coefficient *r***
  - The **linear correlation coefficient** is denoted by *r*, and it measures the strength of the linear association between two variables.

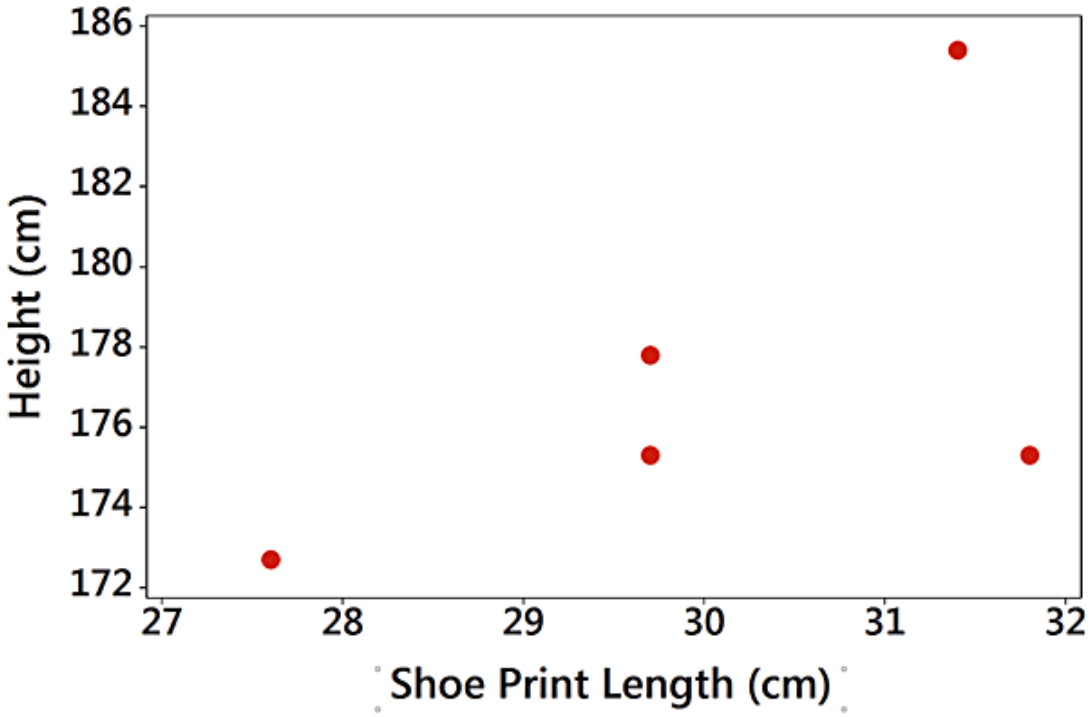
Using *r* for Determining Correlation

The computed value of the linear correlation coefficient, *r*, is always between −1 and 1.

- If *r* is close to −1 or close to 1, there appears to be a correlation.
- If *r* is close to 0, there does not appear to be a linear correlation.

Example: Correlation between Shoe Print Lengths and Heights? (1 of 2)

Shoe Print Length (cm)	29.7	29.7	31.4	31.8	27.6
Height (cm)	175.3	177.8	185.4	175.3	172.7



Example: Correlation between Shoe Print Lengths and Heights? (2 of 2)

It isn't very clear whether there is a linear correlation.

Statdisk

Sample size, n: 5
Degrees of freedom: 3

Correlation Results:
Correlation coeff, r: 0.5912691
Critical r: ±0.8783393
P-value (two-tailed): 0.29369

Regression Results:
Y= b0 + b1x:
Y Intercept, b0: 125.4073
Slope, b1: 1.727452

Total Variation: 95.02
Explained Variation: 33.21891
Unexplained Variation: 61.80109
Standard Error: 4.538762
Coeff of Det, R^2: 0.3495991

P-Value

- ***P*-Value**
  - If there really is no linear correlation between two variables, the ***P*-value** is the probability of getting paired sample data with a linear correlation coefficient *r* that is at least as extreme as the one obtained from the paired sample data.

Interpreting a *P*-Value from the Previous Example

The *P*-value of 0.294 is high. It shows there is a high chance of getting a linear correlation coefficient of $r = 0.591$ (or more extreme) by chance when there is no linear correlation between the two variables.

Statdisk

Sample size, n: 5
Degrees of freedom: 3

Correlation Results:
Correlation coeff, r: 0.5912691
Critical r: ±0.8783393
P-value (two-tailed): 0.29369

Regression Results:
Y= b0 + b1x:
Y Intercept, b0: 125.4073
Slope, b1: 1.727452

Total Variation: 95.02
Explained Variation: 33.21891
Unexplained Variation: 61.80109
Standard Error: 4.538762
Coeff of Det, R^2: 0.3495991

Interpreting a *P*-Value from the Example Where $n = 5$

Because the likelihood of getting $r = 0.591$ or a more extreme value is so high (29.4% chance), we conclude there is not sufficient evidence to conclude there is a linear correlation between shoe print lengths and heights.

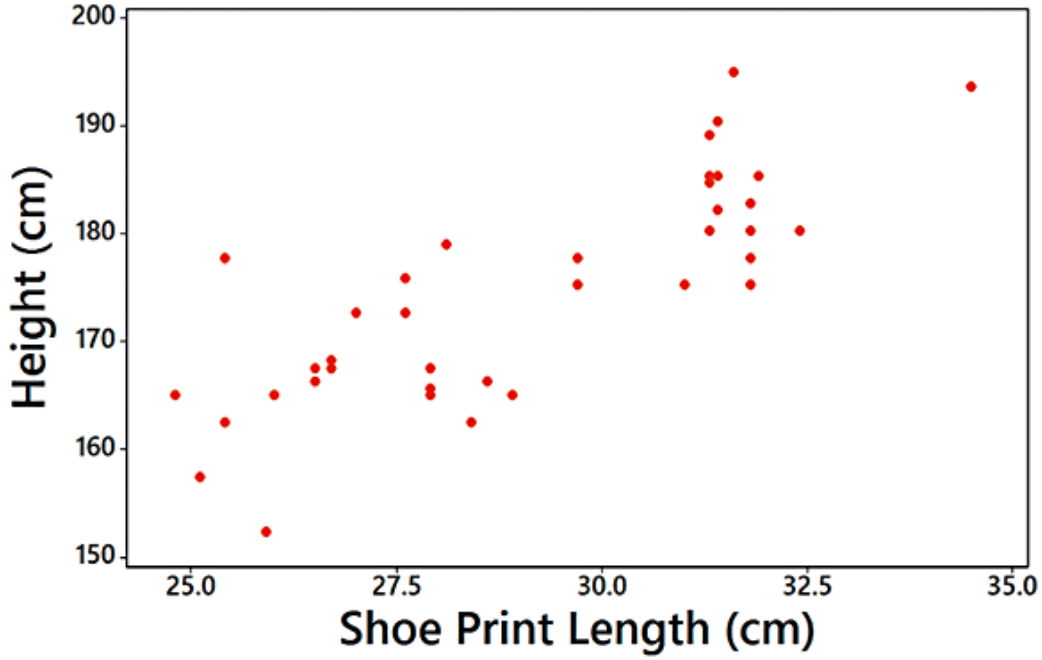
Interpreting a *P*-Value

Only a **small** *P*-value, such as 0.05 or less (or a 5% chance or less), suggests that the sample results are **not** likely to occur by chance when there is no linear correlation, so a small *P*-value supports a conclusion that there is a linear correlation between the two variables.

Example: Correlation between Shoe Print Lengths and Heights (n = 40)

Minitab

```
Pearson correlation of Shoe Print Length and Height = 0.813
P-Value = 0.000
```



Example: Correlation between Shoe Print Lengths and Heights

Minitab

```
Pearson correlation of Shoe Print Length and Height = 0.813
P-Value = 0.000
```

The scatterplot shows a distinct pattern. The value of the linear correlation coefficient is $r = 0.813$, and the P-value is 0.000. Because the P-value of 0.000 is **small**, we have sufficient evidence to conclude there is a linear correlation between shoe print lengths and heights.

Regression

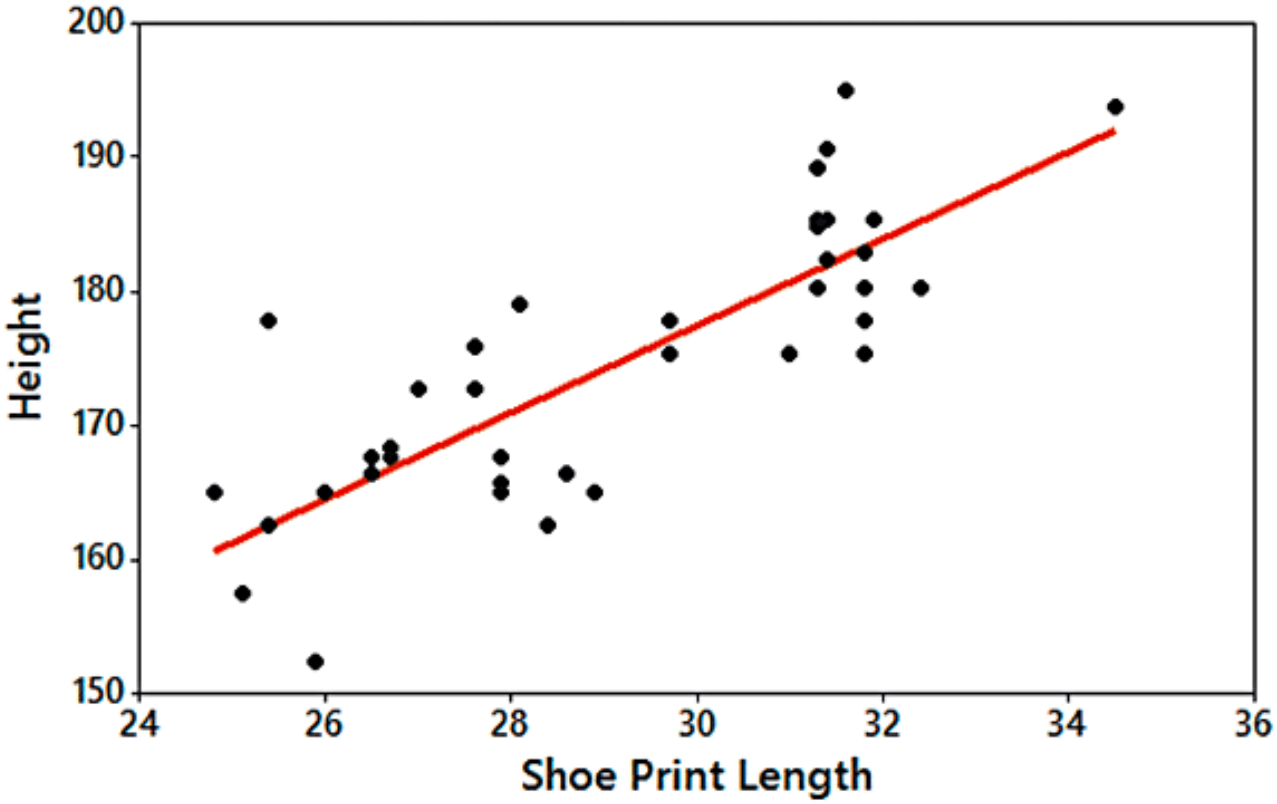
- **Regression**
  - Given a collection of paired sample data, the **regression line** (or **line of best fit,** or **least-squares line**) is the straight line that "best" fits the scatterplot of the data.

The **regression equation**

$$\hat{y} = b_0 + b_1 x$$

algebraically describes the regression line.

Example: Regression Line (1 of 2)



Example: Regression Line (2 of 2)

Statdisk

```
Correlation Results:
Correlation coeff, r:  0.812948
Critical r:  ±0.3120061
P-value (two-tailed):  0.000

Regression Results:
Y= b0 + b1x:
Y Intercept, b0:  80.93041
Slope, b1:  3.218561
```

The general form of the regression equation has a y-intercept of $b_0 = 80.9$ and slope $b_1 = 3.22$.

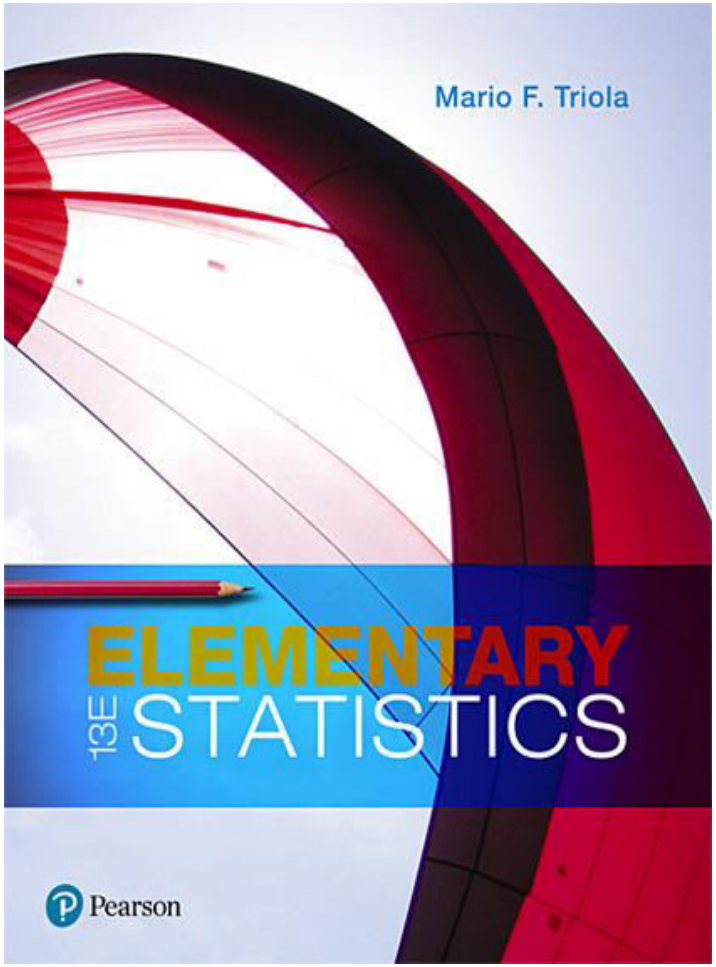
The equation of the regression line is $\hat{y} = 80.9 + 3.22x$.

Using variable names, the equation is:

Height = 80.9 + 3.22 (Shoe Print Length)

Elementary Statistics

Thirteenth Edition



Chapter 2

Exploring Data with Tables and Graphs

Exploring Data with Tables and Graphs

2-1 Frequency Distributions for Organizing and Summarizing Data

2-2 Histograms

2-3 Graphs that Enlighten and Graphs that Deceive

2-4 Scatterplots, Correlation, and Regression

Key Concept

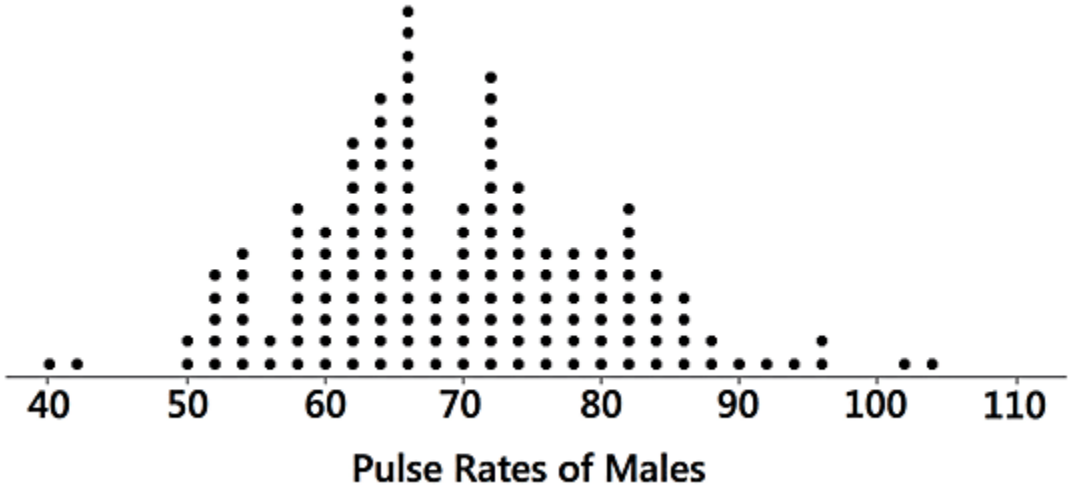
Introduce other common graphs that foster understanding of data.

Discuss some graphs that are deceptive because they create impressions about data that are somehow misleading or wrong.

Technology now provides us with powerful tools for generating a wide variety of graphs.

Graphs that Enlighten: Dotplots (1 of 2)

- **Dotplots**
  - A graph of **quantitative** data in which each data value is plotted as a point (or dot) above a horizontal scale of values. Dots representing equal values are stacked.

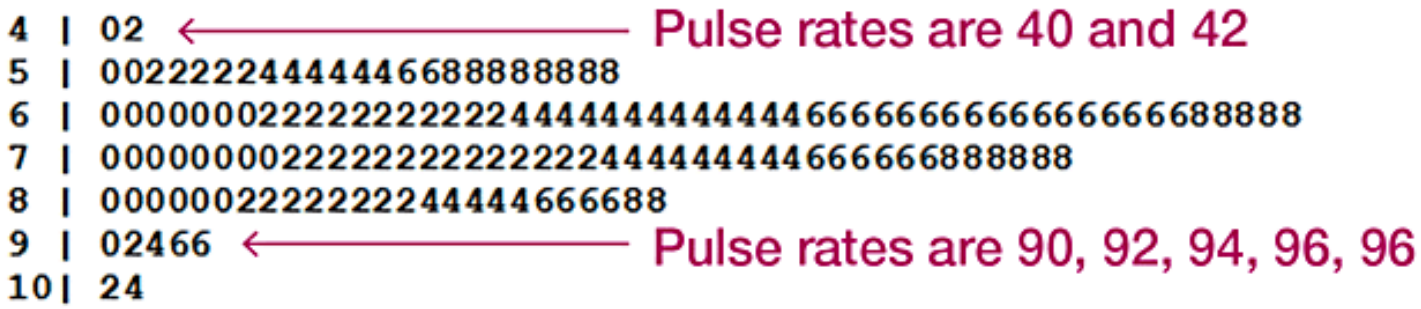


Graphs that Enlighten: Dotplots (2 of 2)

- **Dotplots**
  - **Features of a Dotplot**
    - Displays the shape of distribution of data.
    - It is usually possible to recreate the original list of data values.

Stemplots (1 of 2)

- **Stemplots (or stem-and-leaf plot)**
  - Represents **quantitative** data by separating each value into two parts: the stem (such as the leftmost digit) and the leaf (such as the rightmost digit).

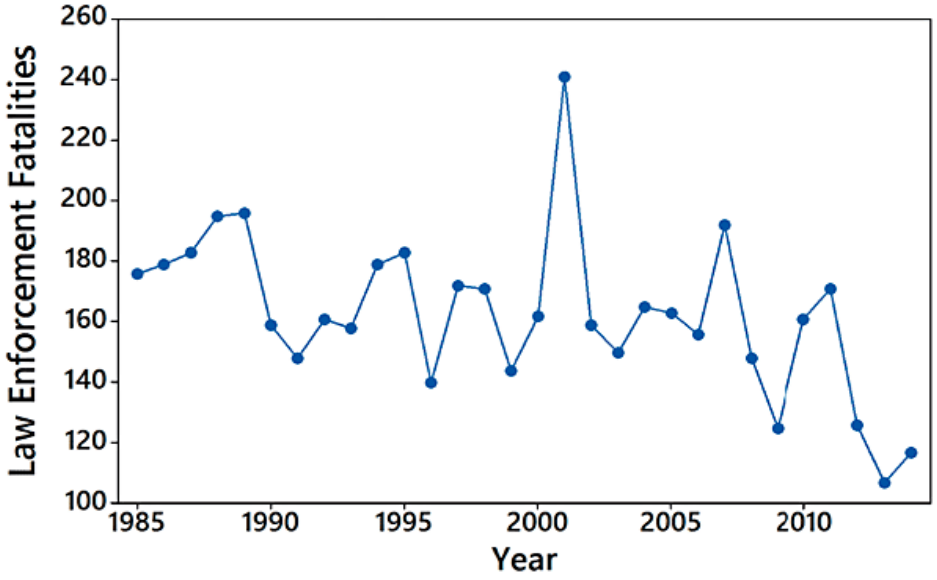


Stemplots (2 of 2)

- **Stemplots (or stem-and-leaf plot)**
  - **Features of a Stemplot**
    - Shows the shape of the distribution of the data.
    - Retains the original data values.
    - The sample data are sorted (arranged in order).

Time-Series Graph (1 of 2)

- **Time-Series Graph**
  - A graph of **time-series data,** which are quantitative data that have been collected at different points in time, such as monthly or yearly



Time-Series Graph (2 of 2)

- **Time-Series Graph**
  - **Feature of a Time-Series Graph**
    - Reveals information about trends over time.

Bar Graph (1 of 2)

- **Bar Graphs**
  - A graph of bars of equal width to show frequencies of categories of **categorical** (or qualitative) data. The bars may or may not be separated by small gaps.

Bar Graph (2 of 2)

- **Bar Graphs**
  - **Feature of a Bar Graph**
    - Shows the relative distribution of categorical data so that it is easier to compare the different categories.

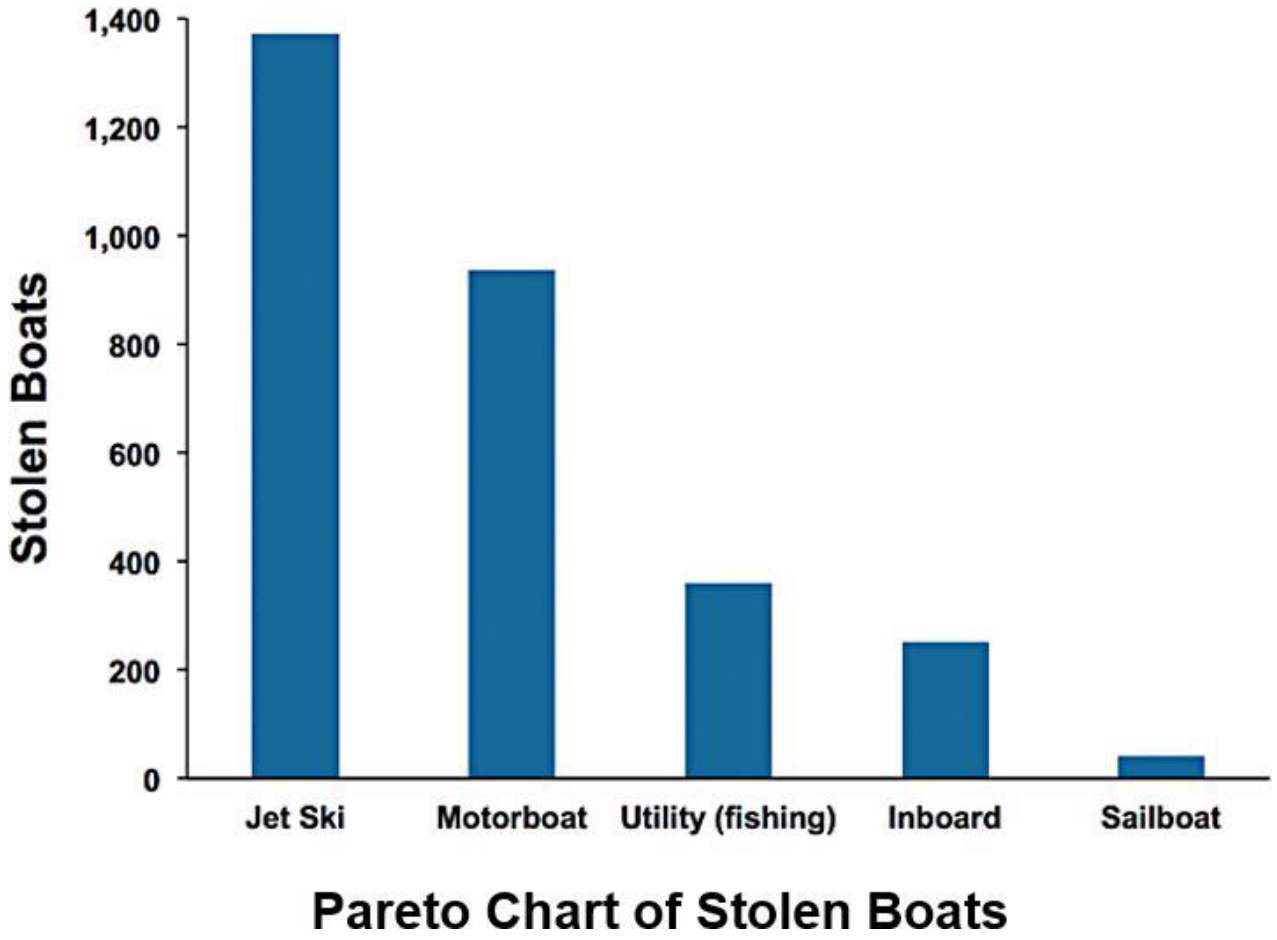
Pareto Chart (1 of 3)

- **Pareto Charts**
  - A Pareto chart is a bar graph for categorical data, with the added stipulation that the **bars are arranged in descending order** according to frequencies, so the bars decrease in height from left to right.

Pareto Chart (2 of 3)

- **Pareto Charts**
  - **Features of a Pareto Chart**
    - Shows the relative distribution of categorical data so that it is easier to compare the different categories.
    - Draws attention to the more important categories.

Pareto Chart (3 of 3)



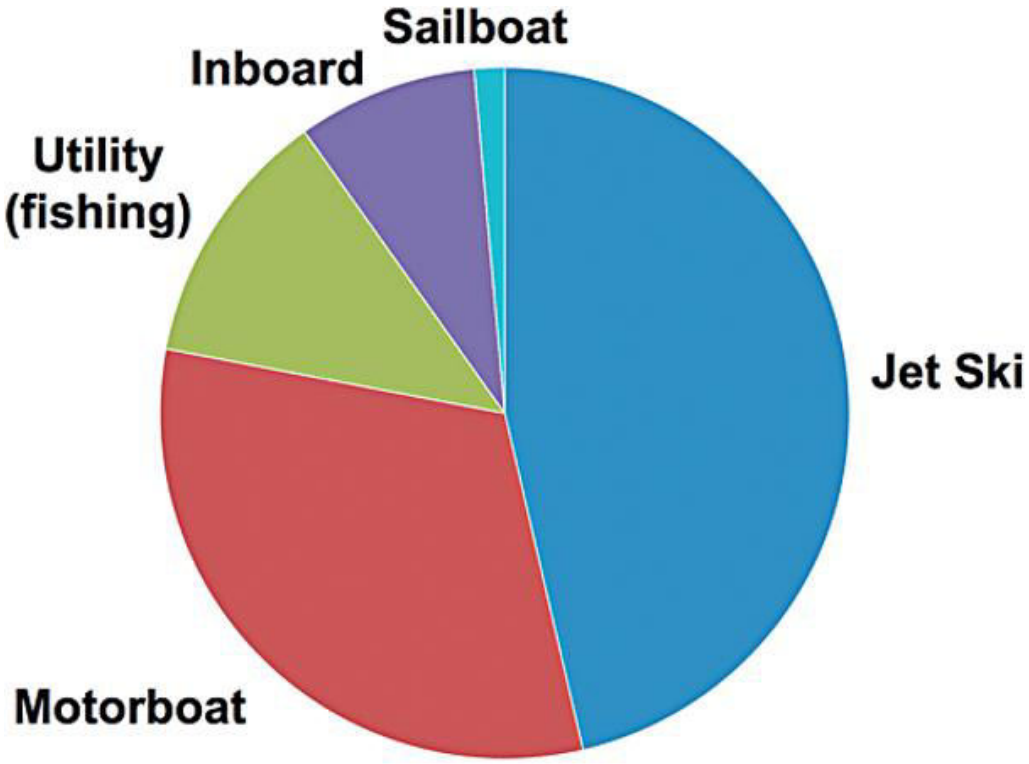
Pie Chart (1 of 3)

- **Pie Charts**
  - A very common graph that depicts categorical data as slices of a circle, in which the size of each slice is proportional to the frequency count for the category

Pie Chart (2 of 3)

- **Pie Charts**
  - **Feature of a Pie Chart**
    - Shows the distribution of categorical data in a commonly used format.

Pie Chart (3 of 3)



Pie Chart of Stolen Boats

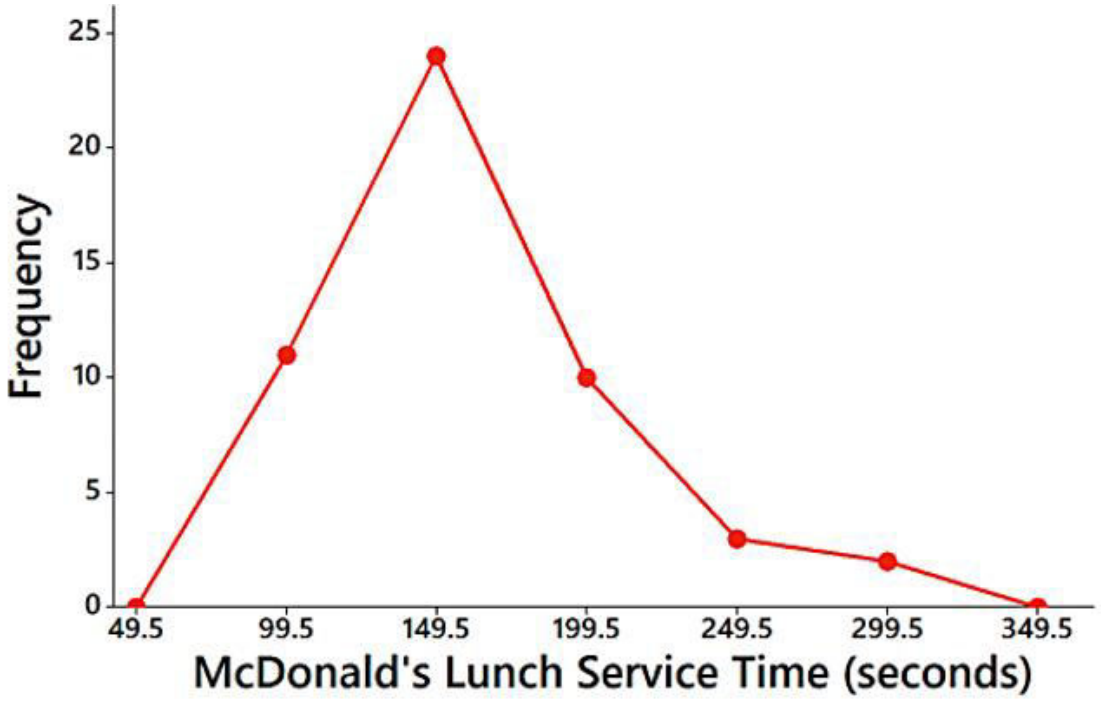
Frequency Polygon (1 of 3)

- **Frequency Polygon**
  - A graph using line segments connected to points located directly above class midpoint values
  - A frequency polygon is very similar to a histogram, but a frequency polygon uses line segments instead of bars.

Frequency Polygon (2 of 3)

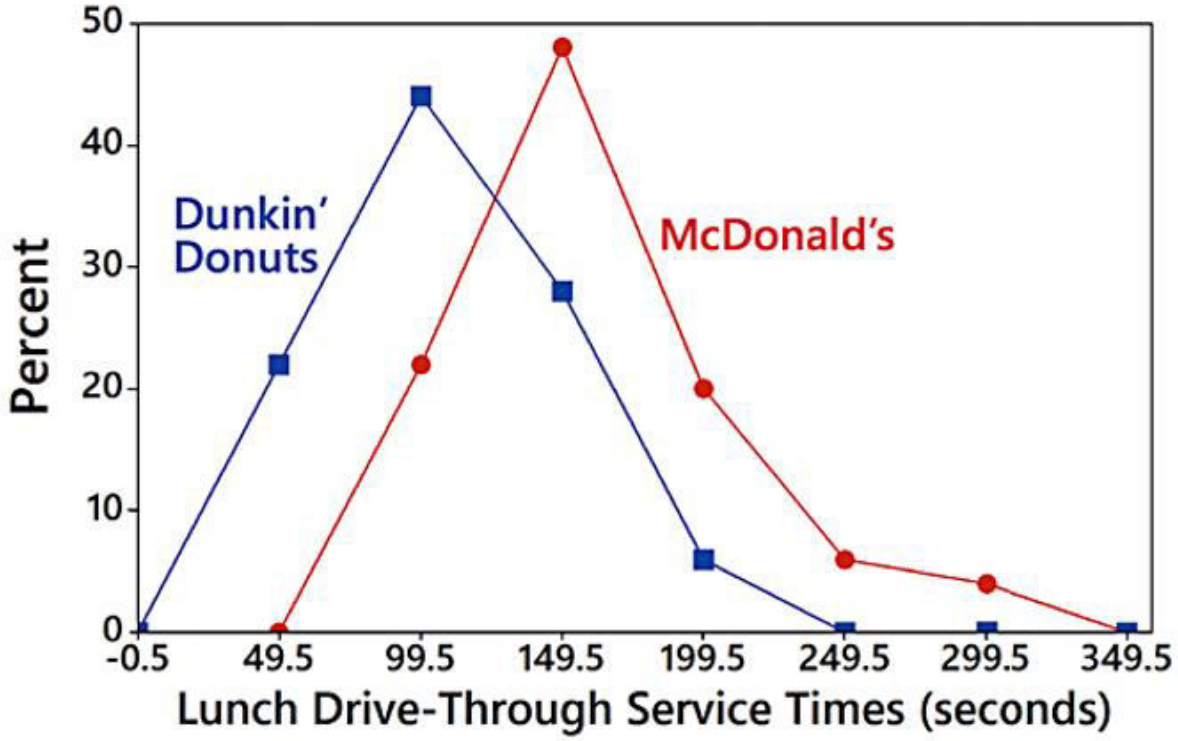
- **Frequency Polygon**
  - A variation of the basic frequency polygon is the relative frequency polygon, which uses relative frequencies (proportions or percentages) for the vertical scale.

Frequency Polygon (3 of 3)



Frequency Polygon of McDonald's Lunch Service Times

Relative Frequency Polygon



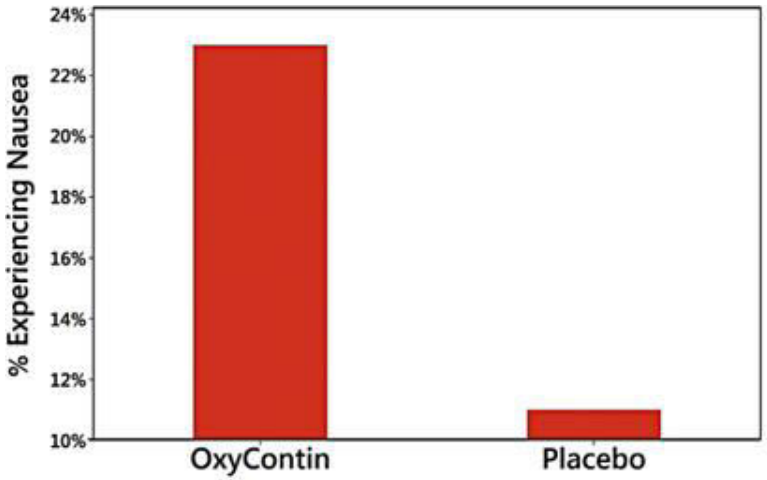
Relative Frequency Polygons for McDonald's and Dunkin' Donuts

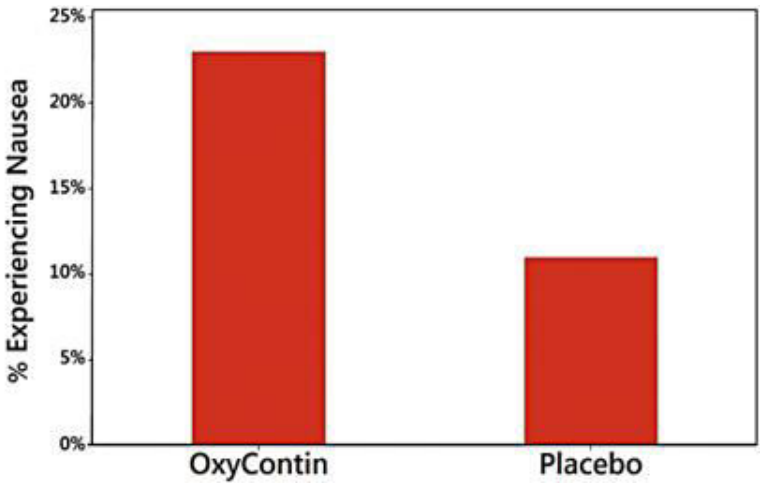
Graphs That Deceive (1 of 4)

- **Nonzero Vertical Axis**
  - A common deceptive graph involves using a vertical scale that starts at some value greater than zero to exaggerate differences between groups.

Graphs That Deceive (2 of 4)

- **Nonzero Vertical Axis**





Always examine a graph carefully to see whether a vertical axis begins at some point other than zero so that differences are exaggerated.

Graphs That Deceive (3 of 4)

- **Pictographs**
  - Drawings of objects, called **pictographs**, are often misleading. Data that are one-dimensional in nature (such as budget amounts) are often depicted with two-dimensional objects (such as dollar bills) or three-dimensional objects (such as stacks of coins, homes, or barrels).

Graphs That Deceive (4 of 4)

- **Pictographs**
  - By using pictographs, artists can create false impressions that grossly distort differences by using these simple principles of basic geometry:
    - When you double each side of a square, its area doesn't merely double; it increases by a factor of **four**.
    - When you double each side of a cube, its volume doesn't merely double; it increases by a factor of **eight.**

Pictographs

1970: 37% of U.S. adults smoked.

2013: 18% of U.S. adults smoked.

Concluding Thoughts (1 of 2)

In addition to the graphs we have discussed in this section, there are many other useful graphs - some of which have not yet been created. The world needs more people who can create original graphs that enlighten us about the nature of data.

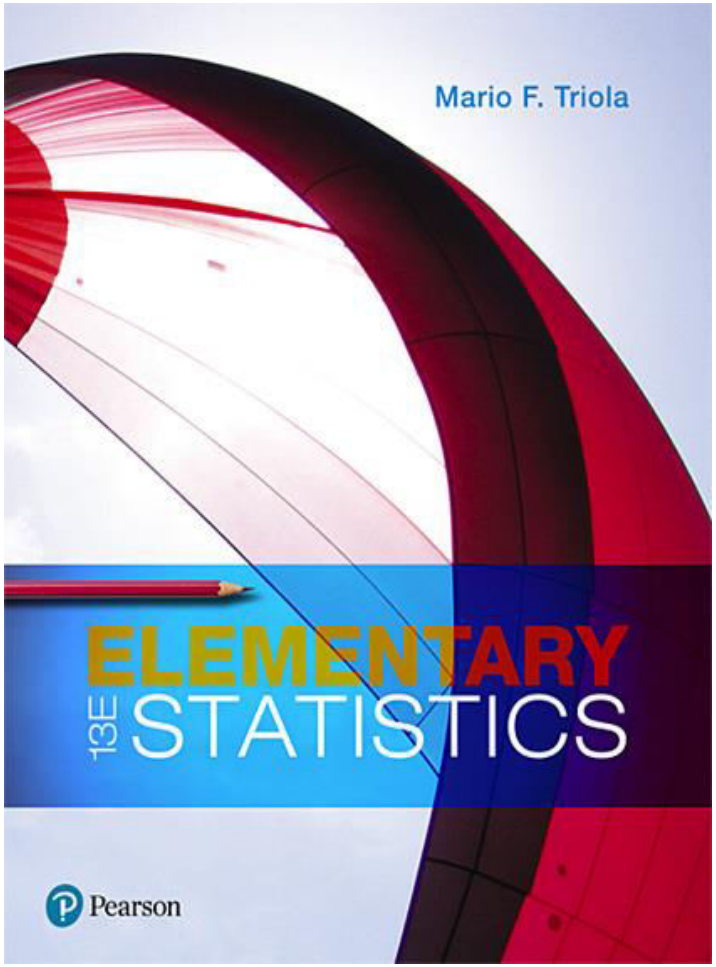
Concluding Thoughts (2 of 2)

In **The Visual Display of Quantitative Information,** Edward Tufte offers these principles:

- For small data sets of 20 values or fewer, use a table instead of a graph.
- A graph of data should make us focus on the true nature of the data, not on other elements, such as eye-catching but distracting design features.
- Do not distort data; construct a graph to reveal the true nature of the data.
- Almost all of the ink in a graph should be used for the data, not for other design elements.

Elementary Statistics

Thirteenth Edition



Chapter 2

Exploring Data with Tables and Graphs

Exploring Data with Tables and Graphs

2-1 Frequency Distributions for Organizing and Summarizing Data

2-2 Histograms

2-3 Graphs that Enlighten and Graphs that Deceive

2-4 Scatterplots, Correlation, and Regression

Key Concept

While a frequency distribution is a useful tool for summarizing data and investigating the distribution of data, an even better tool is a **histogram**, which is a graph that is easier to interpret than a table of numbers.

Histogram

- **Histogram**
  - A graph consisting of bars of equal width drawn adjacent to each other (unless there are gaps in the data)

    The horizontal scale represents classes of quantitative data values, and the vertical scale represents frequencies. The heights of the bars correspond to frequency values.

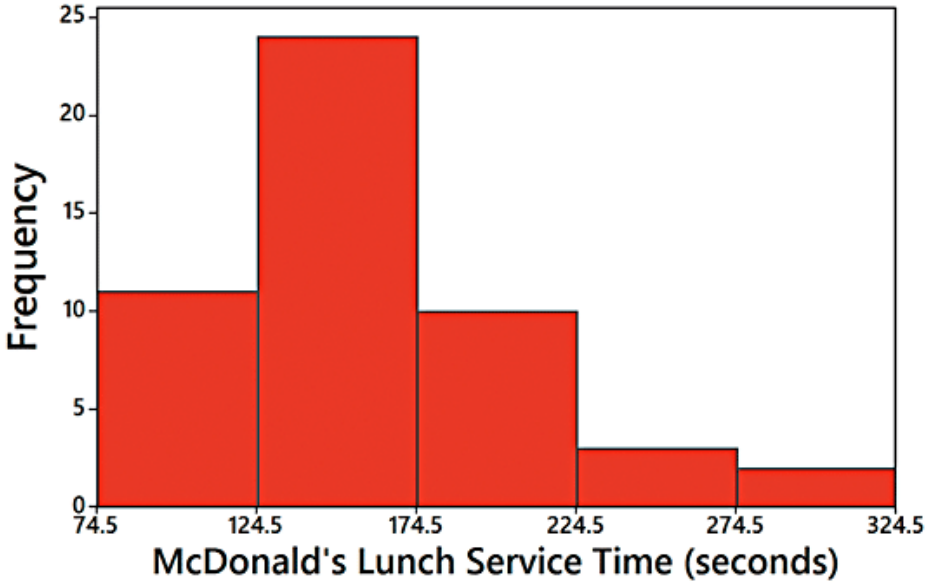
Important Uses of a Histogram

- Visually displays the shape of the **distribution** of the data
- Shows the location of the **center** of the data
- Shows the **spread** of the data
- Identifies **outliers**

Relative Frequency Histogram

- **Relative Frequency Histogram**
  - It has the same shape and horizontal scale as a histogram, but the vertical scale is marked with relative frequencies instead of actual frequencies.

Time (seconds)	Frequency
75-124	11
125-174	24
175-224	10
225-274	3
275-324	2

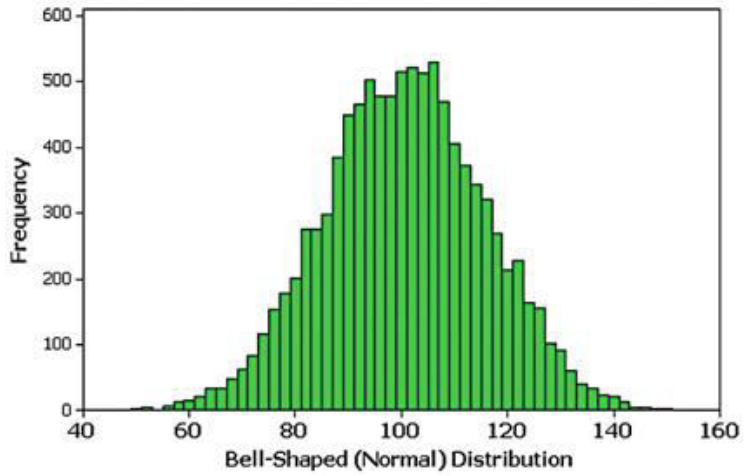


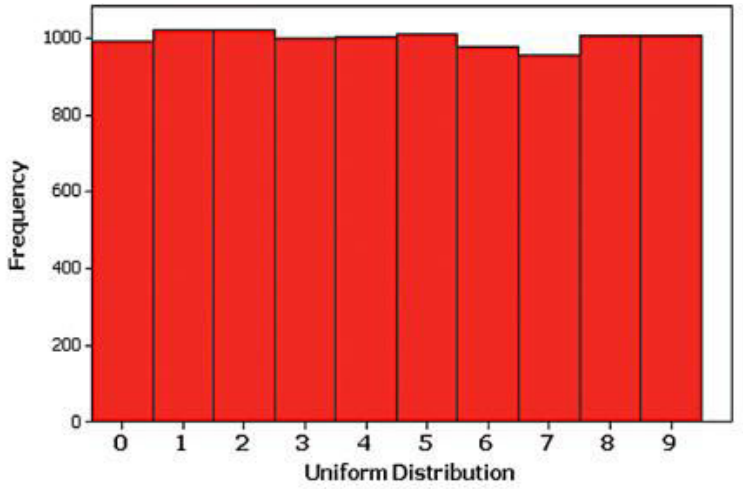
Critical Thinking Interpreting Histograms

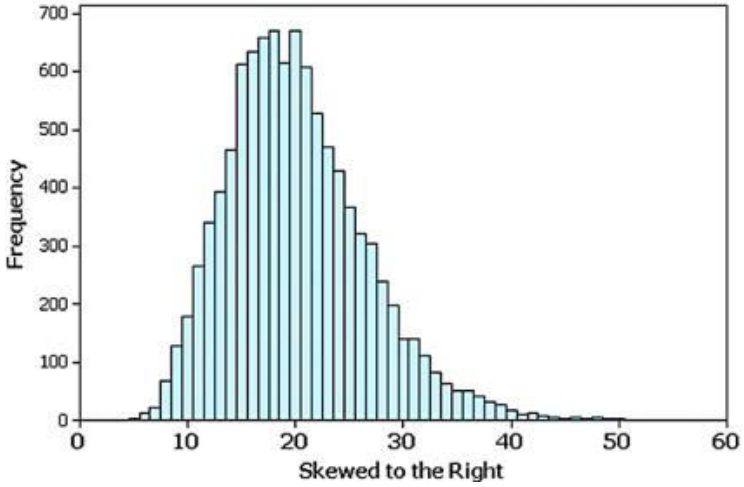
Explore the data by analyzing the histogram to see what can be learned about "**CVDOT**":

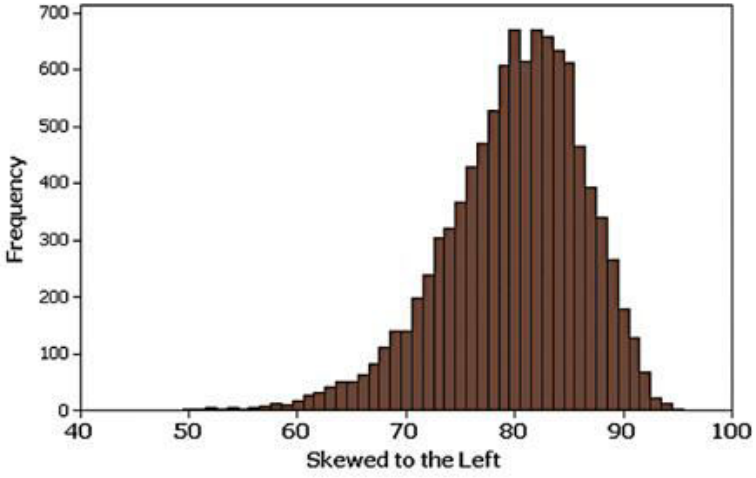
- the **C**enter of the data,
- the **V**ariation,
- the shape of the **D**istribution,
- whether there are any **O**utliers,
- and **T**ime.

Common Distribution Shapes



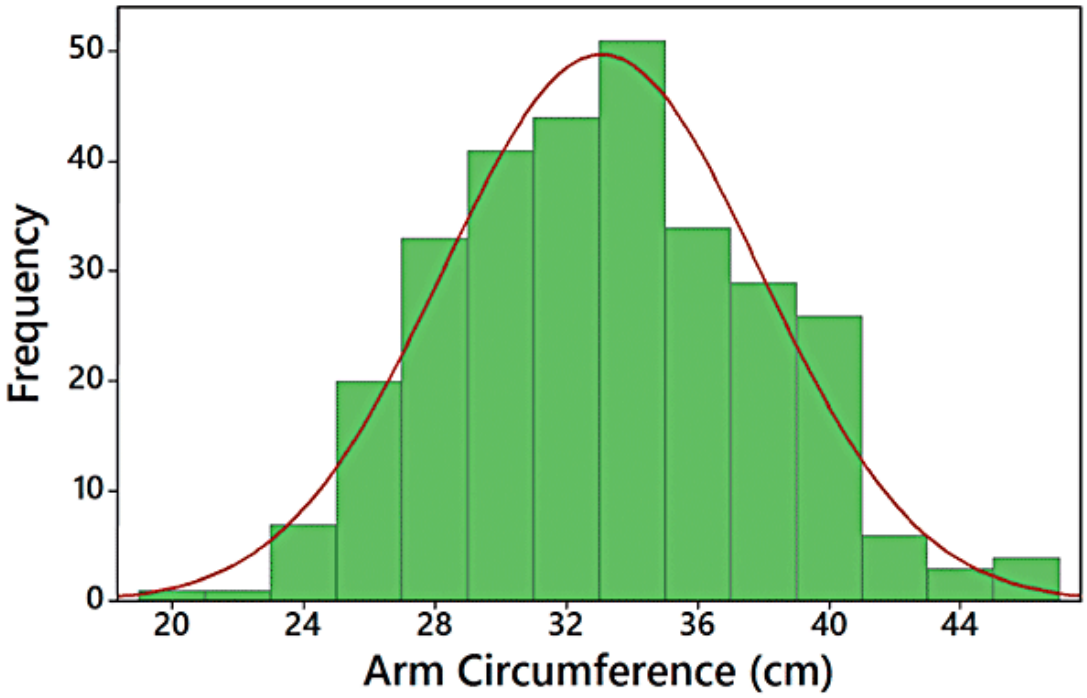






Normal Distribution

Because this histogram is roughly bell-shaped, we say that the data have a **normal distribution**.

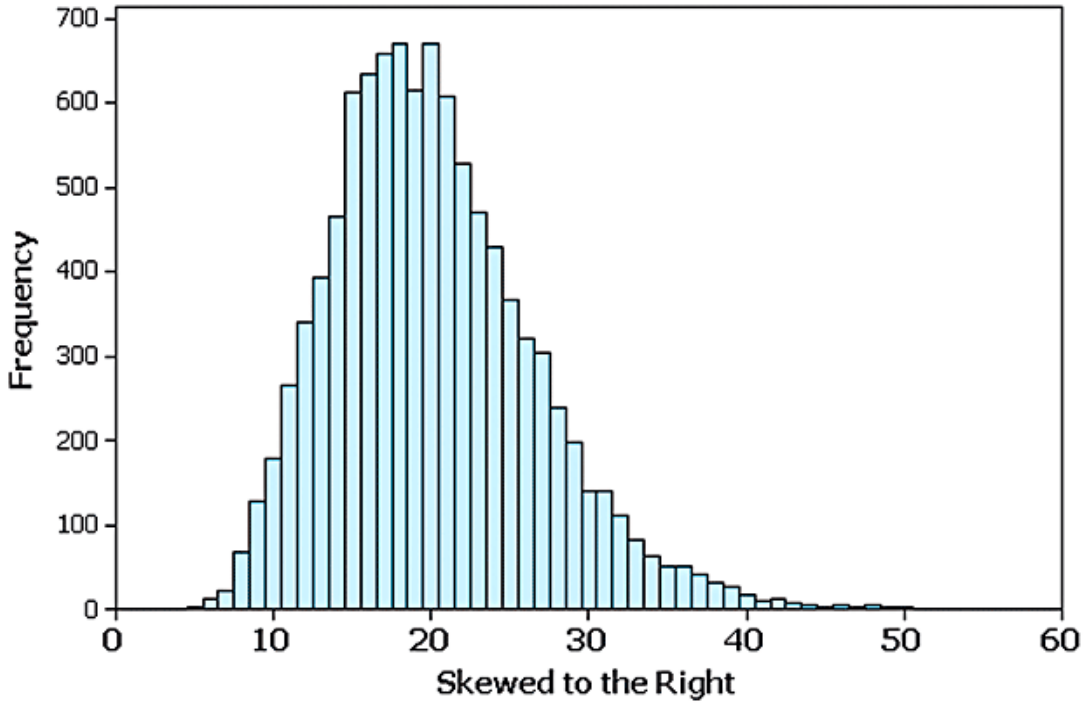


Skewness (1 of 3)

- **Skewness**
  - A distribution of data is **skewed** if it is not symmetric and extends more to one side than to the other.

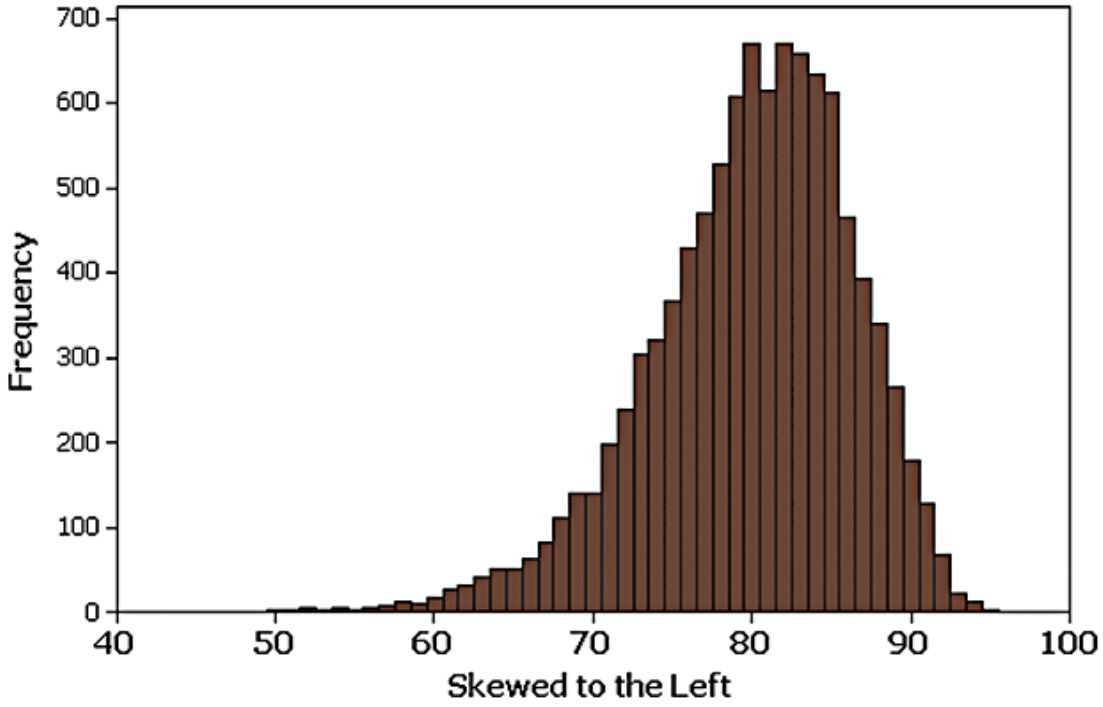
Skewness (2 of 3)

Data **skewed to the right (positively skewed)** have a longer right tail.



Skewness (3 of 3)

Data **skewed to the left (negative skewed)** have a longer left tail.



Assessing Normality with Normal Quantile Plots (1 of 5)

Criteria for Assessing Normality with a Normal Quantile Plot

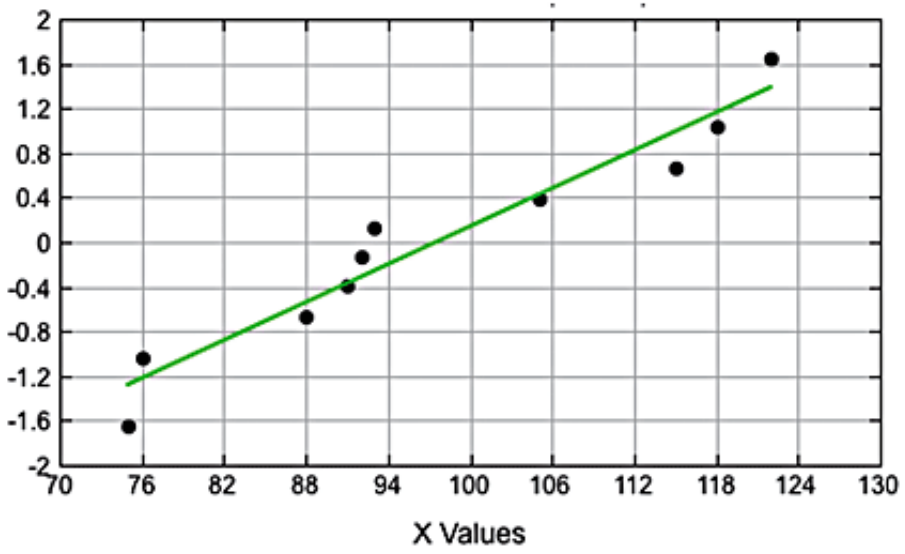
- **Normal Distribution:** The pattern of the points in the normal quantile plot is reasonably close to a straight line, and the points do not show some systematic pattern that is not a straight-line pattern.

Assessing Normality with Normal Quantile Plots (2 of 5)

Criteria for Assessing Normality with a Normal Quantile Plot

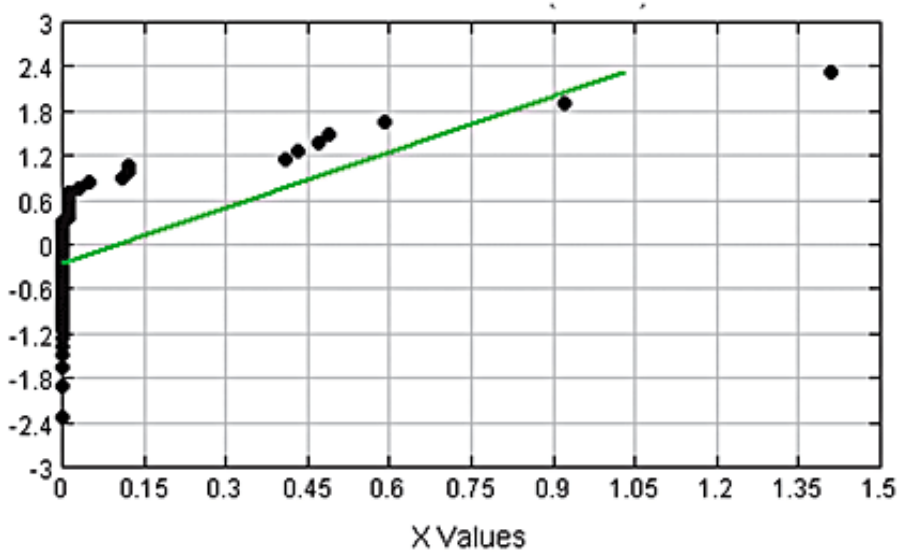
- **Not a Normal Distribution:** The population distribution is **not** normal if the normal quantile plot has either or both of these two conditions:
  - The points do not lie reasonably close to a straight-line pattern.
  - The points show some systematic pattern that is not a straight-line pattern.

Assessing Normality with Normal Quantile Plots (3 of 5)



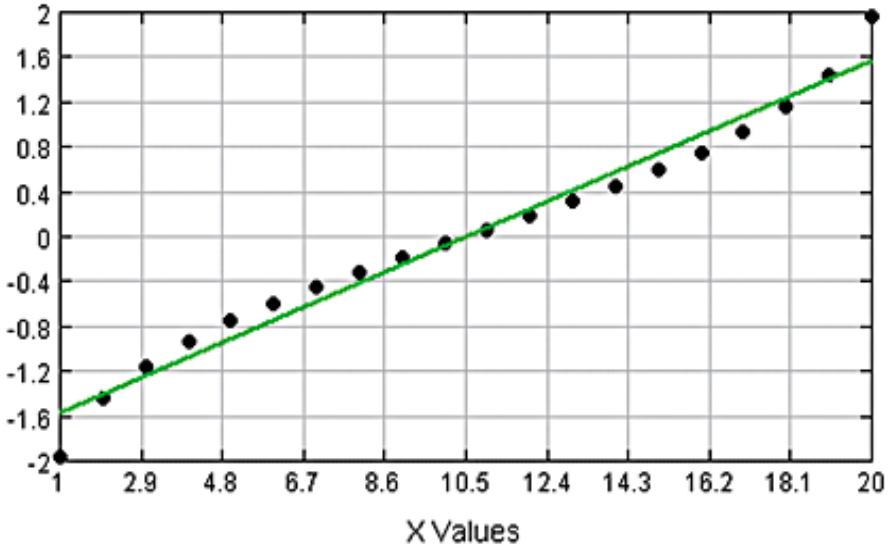
Normal Distribution: The points are reasonably close to a straight-line pattern, and there is no other systematic pattern that is not a straight-line pattern.

Assessing Normality with Normal Quantile Plots (4 of 5)



Not a Normal Distribution: The points do not lie reasonably close to a straight line.

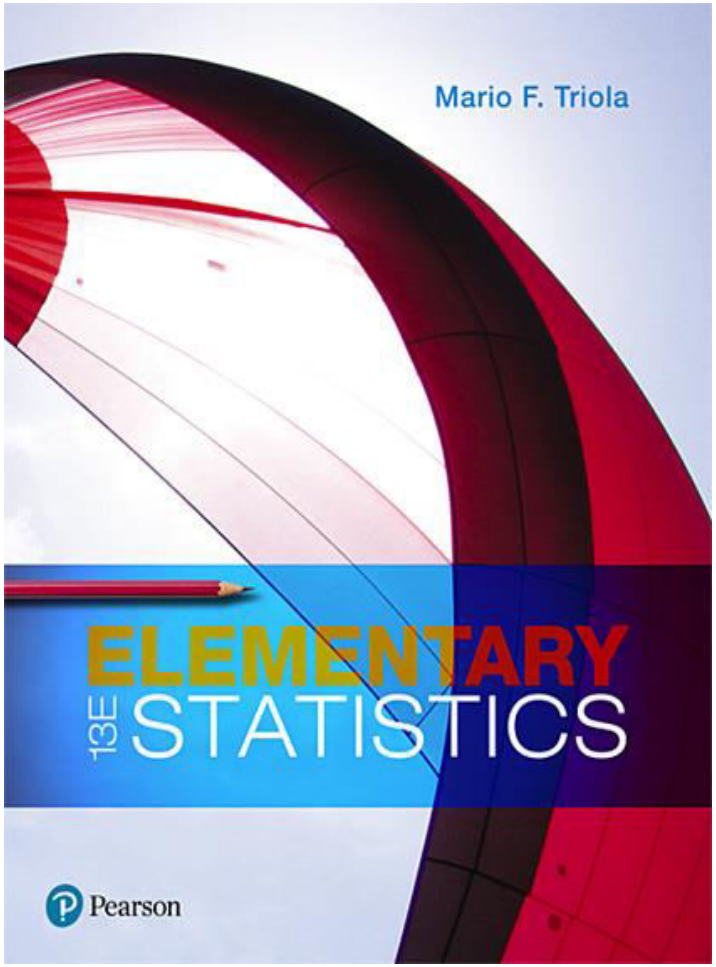
Assessing Normality with Normal Quantile Plots (5 of 5)



Not a Normal Distribution: The points show a systematic pattern that is not a straight-line pattern.

Elementary Statistics

Thirteenth Edition



Chapter 2

Exploring Data with Tables and Graphs

Exploring Data with Tables and Graphs

2-1 Frequency Distributions for Organizing and Summarizing Data

2-2 Histograms

2-3 Graphs that Enlighten and Graphs that Deceive

2-4 Scatterplots, Correlation, and Regression

Key Concept

When working with large data sets, a **frequency distribution** (or **frequency table**) is often helpful in organizing and summarizing data. A frequency distribution helps us to understand the nature of the **distribution** of a data set.

Frequency Distribution

- Frequency Distribution (or Frequency Table)
  - Shows how data are partitioned among several categories (or **classes**) by listing the categories along with the number (frequency) of data values in each of them.

Definitions (1 of 2)

- Lower class limits
  - The smallest numbers that can belong to each of the different classes
- Upper class limits
  - The largest numbers that can belong to each of the different classes
- Class boundaries
  - The numbers used to separate the classes, but without the gaps created by class limits

Definitions (2 of 2)

- Class midpoints
  - The values in the middle of the classes Each class midpoint can be found by adding the lower class limit to the upper class limit and dividing the sum by 2.
- Class width
  - The difference between two consecutive lower class limits in a frequency distribution

Procedure for Constructing a Frequency Distribution (1 of 2)

1. Select the number of classes, usually between 5 and 20.
2. Calculate the class width.

$$\text{Class width} \approx \frac{(\text{maximum data value}) - (\text{minimum data value})}{\text{number of classes}}$$

Round this result to get a convenient number. (It's usually best to round **up**.)

Procedure for Constructing a Frequency Distribution (2 of 2)

3. Choose the value for the first lower class limit by using either the minimum value or a convenient value below the minimum.

4. Using the first lower class limit and class width, list the other lower class limits.

5. List the lower class limits in a vertical column and then determine and enter the upper class limits.

6. Take each individual data value and put a tally mark in the appropriate class. Add the tally marks to get the frequency.

Example: McDonald's Lunch Service Times (1 of 5)

Using the McDonald's lunch service times in the first table, follow the procedure shown on the next slide to construct the frequency distribution shown in the second table. Use five classes.

Drive-through Service Times (seconds) for McDonald's Lunches

107	139	197	209	281	254	163	150	127	308	206	187	169	83	127	133	140
143	130	144	91	113	153	255	252	200	117	167	148	184	123	153	155	154
100	117	101	138	186	196	146	90	144	119	135	151	197	171	190	169	

McDonald's Lunch Drive-Through Service Times

Time (Seconds)	Frequency
75-124	11
125-174	24
175-224	10
225-274	3
275-324	2

Example: McDonald's Lunch Service Times (2 of 5)

Step 1: Select 5 as the number of desired classes.

Step 2: Calculate the class width as shown below. Note that we round 45 up to 50,which is a more convenient number.

$$\text{Class width} \approx \frac{(\text{maximum data value}) - (\text{minimum data value})}{\text{number of classes}}$$

$$= \frac{308 - 83}{5} = 45 \approx 50 \quad \text{(rounded up to a more convenient number)}$$

Example: McDonald's Lunch Service Times (3 of 5)

Step 3: The minimum data value is 83, which is not a very convenient starting point, so go to a value below 83 and select the more convenient value of 75 as the first lower class limit.

Step 4: Add the class width of 50 to the starting value of 75 to get the second lower class limit of 125. Continue to add the class width of 50 until we have five lower class limits. The lower class limits are therefore 75, 125, 175, 225, and 275.

Example: McDonald's Lunch Service Times (4 of 5)

Step 5: List the lower class limits vertically, as shown below. From this list, we identify the corresponding upper class limits as 124, 174, 224, 274, and 324.

75-
125-
175-
225-
275-

Example: McDonald's Lunch Service Times (5 of 5)

Step 6: Enter a tally mark for each data value in the appropriate class. Then add the tally marks to find the frequencies shown in the table.

Time (Seconds)	Frequency
75-124	11
125-174	24
175-224	10
225-274	3
275-324	2

Relative Frequency Distribution (1 of 2)

- Relative Frequency Distribution or Percentage Frequency Distribution
  - Each class frequency is replaced by a relative frequency (or proportion) or a percentage.

$$\text{Relative frequency for a class} = \frac{\text{frequency for a class}}{\text{sum of all frequencies}}$$

$$\text{Percentage for a class} = \frac{\text{frequency for a class}}{\text{sum of all frequencies}} \times 100\%$$

Relative Frequency Distribution (2 of 2)

- Relative Frequency Distribution or Percentage Frequency Distribution
  - Each class frequency is replaced by a relative frequency (or proportion) or a percentage.

The sum of the percentages in a relative frequency distribution must be very close to 100% (with a little wiggle room for rounding errors).

Cumulative Frequency Distribution

- Cumulative Frequency Distribution
  - The frequency for each class is the sum of the frequencies for that class and all previous classes.

Cumulative Frequency Distribution of McDonald's Lunch Service Times

Time (Seconds)	Cumulative Frequency
Less than 125	11
Less than 175	35
Less than 225	45
Less than 275	48
Less than 325	50

Critical Thinking: Using Frequency Distributions to Understand Data

In statistics we are often interested in determining whether the data have a **normal distribution**.

1. The frequencies start low, then increase to one or two high frequencies, and then decrease to a low frequency.
2. The distribution is approximately symmetric. Frequencies preceding the maximum frequency should be roughly a mirror image of those that follow the maximum frequency.

Gaps

- The presence of gaps can show that the data are from two or more different populations.
- However, the converse is not true, because data from different populations do not necessarily result in gaps.

Example: Exploring Data: What Does a Gap Tell Us? (1 of 2)

The table shown is a frequency distribution of the weights (grams) of randomly selected pennies.

Weight (grams) of Penny	Frequency
2.40-2.49	18
2.50-2.59	19
2.60-2.69	0
2.70-2.79	0
2.80-2.89	0
2.90-2.99	2
3.00-3.09	25
3.10-3.19	8

Example: Exploring Data: What Does a Gap Tell Us? (2 of 2)

- Examination of the frequencies reveals a large **gap** between the lightest pennies and the heaviest pennies.
- This suggests that we have two different populations:
  - Pennies made before 1983 are 95% copper and 5% zinc.
  - Pennies made after 1983 are 2.5% copper and 97.5% zinc.

Comparisons

Combining two or more relative frequency distributions in one table makes comparisons of data much easier.

Example: Comparing McDonald's and Dunkin' Donuts (1 of 2)

The table shows the relative frequency distributions for the drive-through lunch service times (seconds) for McDonald's and Dunkin' Donuts.

Time (seconds)	McDonald's	Dunkin' Donuts
25-74		22%
75-124	22%	44%
125-174	48%	28%
175-224	20%	6%
225-274	6%	
275-324	4%	

Example: Comparing McDonald's and Dunkin' Donuts (2 of 2)

Time (seconds)	McDonald's	Dunkin' Donuts
25-74		22%
75-124	22%	44%
125-174	48%	28%
175-224	20%	6%
225-274	6%	
275-324	4%	

- Because of the dramatic differences in their menus, we might expect the service times to be very different.
- By comparing the relative frequencies, we see that there are major differences. The Dunkin' Donuts service times appear to be lower than those at McDonald's.