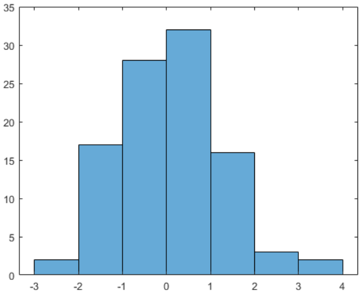
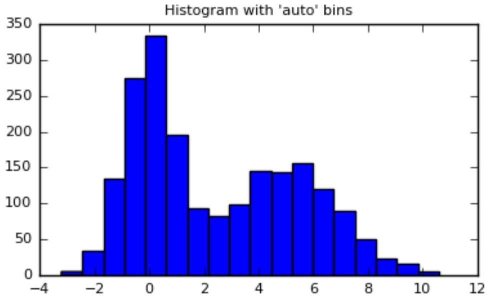
Chapter 2: Descriptive Statistics

2.2: Histograms

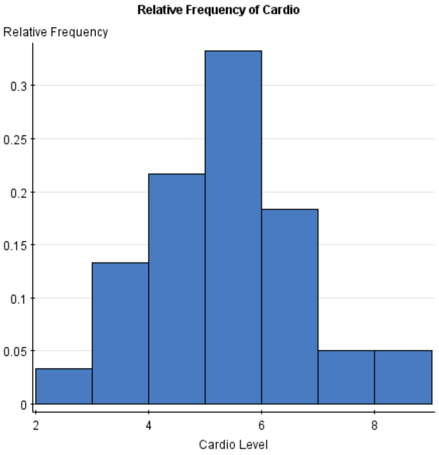
Graphics for Quantitative Data

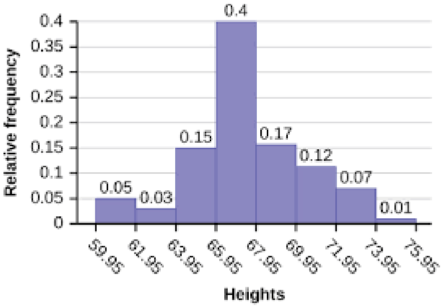
- **Histogram** - a graph consisting of bars of equal width drawn adjacent (they touch) to each other. The horizontal scale represents classes of quantitative data values and the vertical scale represents frequencies. The heights of the bars correspond to the frequency values.





- **Relative Frequency Histogram** - a graph with the same shape and horizontal scale as a histogram, but the vertical scale is marked with relative frequencies.

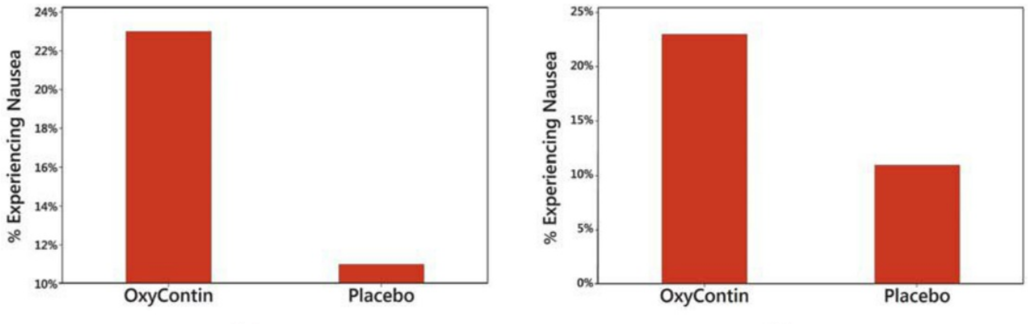




Graphs That Deceive

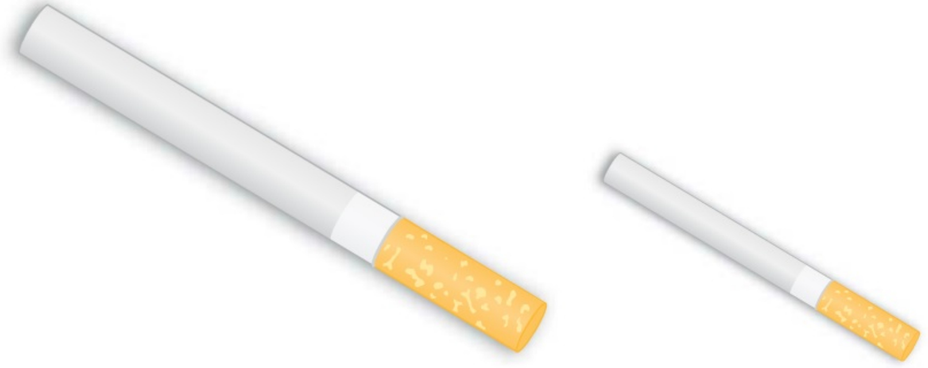
Graphs can be technically correct, but misleading. Two common types of bad graphs are *nonzero axis* graphs and *pictographs*.

- Nonzero axis: Some graphs are misleading because one or both of the axes begin at some value other than zero, so that differences are exaggerated.



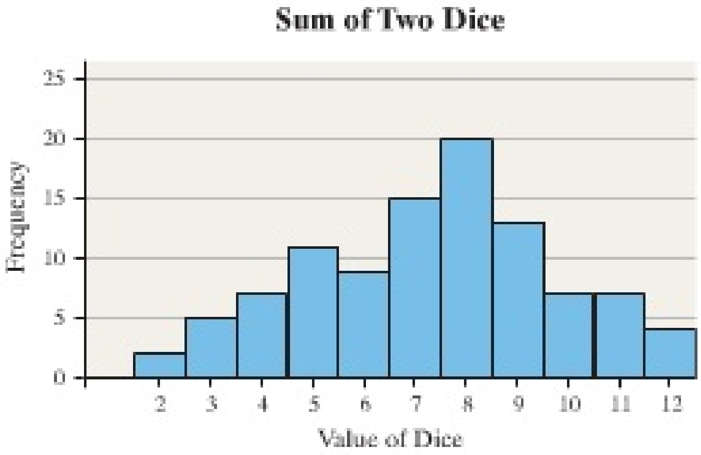
(a) (b)

- Pictographs: Drawings of objects are often misleading. Three-dimensional objects are commonly used to depict data, but drawings of such objects can create false impressions that distort differences.



1970: 37% of U.S. adults smoked. 2013: 18% of U.S. adults smoked.

Example 1. An experiment was conducted in which two fair dice were thrown 100 times. The sum of the pips showing on the dice was then recorded. The following frequency histogram gives the results?



(a) What was the most frequent outcome of the experiment?

(b) What was the least frequent?

(c) How many times did we observe a 7?

(d) How many more 5's were observed than 4's?

(e) Determine the percentage of time a 7 was observed.

- **Median** - the middle value when the original data values are arranged in order of increasing (or decreasing) magnitude. The median is often denoted by M.
  - If the number of data values is odd, the median is the number located in the exact middle of the list.
  - If the number of data values is even, the median is found by computing the mean of the two middle numbers.

Example 2: Determine the Median with an Odd Number of Observations

- The data in the Table 1 represent the length (in seconds) of a random sample of songs released in the 1970s. Find the median length of the songs.

Example 3: Determine the Median with an Even Number of Observations

- Find the median of the data in Table 2.

Table 1: Table for Finding the Median with Odd Number of Observations.

Song	Length
“Sister Golden Hair”	201
“Black Water”	257
“Free Bird”	284
“The Hustle”	208
“Southern Nights”	179
“Stayin’ Alive”	222
“We Are Family”	217
“Heart of Glass”	206
“My Sharona”	240

Table 2: Student Test Scores.

Student	Score
Michelle	82
Ryanne	77
Bilal	90
Pam	71
Jennifer	62
Dave	68
Joel	74
Sam	84
Justine	94
Juan	88

Recall that the median divides the lower 50% of a set of data from the upper 50%. This is a special case of the general concept of percentiles.

- **Percentiles** - denoted P_k, a value in a set of data such that k percent of the observations are less than or equal to the value.

  $$P_x = \text{percentile of } x = \frac{\text{number of values less than } x}{\text{total number of values}} \times 100$$

  Percentiles are used to give the relative standing of an observation. Many standardized exams, such as the SAT college entrance exam, use percentiles to provide students with an understanding of how they scored on the exam in relation to all other students who took the exam.

  A special case of percentiles are called quartiles.

- **Quartiles** - divides the data set into fourths. The 25th, 50th, and 75th percentiles denoted by Q_1, Q_2, and Q_3, respectively.

- The **interquartile range** (IQR) is the difference between Q_3 and Q_1, and describes the variability of the middle 50% of the data values.

  $$IQR = Q_3 - Q_1$$

  Finding Quartiles:

Step 1 Arrange the data in ascending order.

Step 2 Determine the median, M, or second quartile, Q_2.

Step 3 Determine the first and third quartiles, Q_1 and Q_3, by dividing the data set into two halves; the bottom half will be the observations below the location of the median and the top half will be the observations above the median. The first quartile is the median of the lower half and the third quartile is the median of the top half (average two values if necessary).

Example 4. The Highway Loss Data Institute routinely collects data on collision coverage claims. Collision coverage insures against physical damage to an insured individual's vehicle. The data in the table below represent a random sample of 18 collision coverage claims based on data obtained from the Highway Loss Data Institute for 2004 models.

Coverage Claims

$6,751	$9,908	$3,461	$2,336	$21,147	$2,332
$189	$1,185	$370	$1,414	$4,668	$1,953
$10,034	$735	$802	$618	$180	$1,657

(a) Find and interpret the first, second, and third quartiles for collision coverage claims.

(b) Determine and interpret the interquartile range of the collision claim data.

5-number summary - a subset of the data that consists of the minimum value, the first quartile, the median, the third quartile, and the maximum value.

Boxplots - also called box-and-whisker plots; constructed from the 5-number-summary; shows how far extreme values are from the bulk of the data.

- **Strengths:** give a direct look at location and spread; outliers identified; great for comparing.
- **Weaknesses:** symmetry and skewness can be judged, but not so much shape.

When performing any type of data analysis, we should always check for extreme observations in the data set.

- **Outlier** - a data point that is not consistent with the bulk of the data from that group. These values can occur by chance, because of error in the measurement, during data entry, or from sampling errors.

How to Draw a Boxplot and Identify Outliers

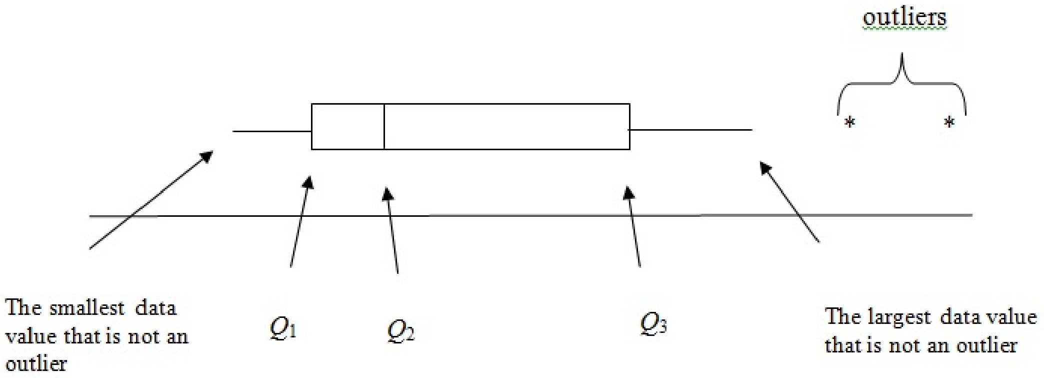
Step 1 Determine the first, Q_1, and third, Q_3 quartiles.

Step 2 Compute the IQR.

Step 3 Determine the fences. Fences serve as cutoff points for determining the outliers

$$\text{Lower Fence} = Q_1 - 1.5 \times (\text{IQR})$$
$$\text{Upper Fence} = Q_3 + 1.5 \times (\text{IQR})$$

Step 4 If the data value is less than the lower fence or greater than the upper fence, it is considered an outlier.

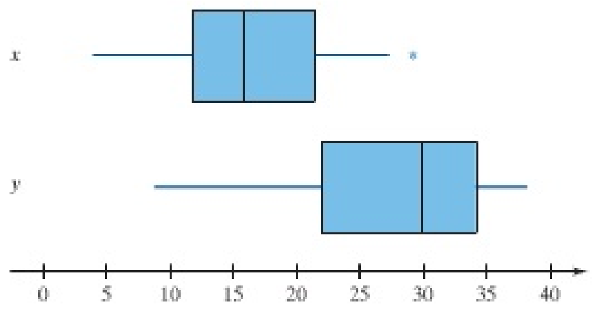

outliers
*
*
The smallest data value that is not an outlier
Q_1
Q_2
Q_3
The largest data value that is not an outlier

Example 4 cont. Using the Collision Coverage Claims data from page 16, answer the following:

(c) Find the outliers from the data.

(d) Draw the boxplot for this problem.

Example 5: Use the side-by-side boxplots shown to answer the questions that follow.



(a) To the nearest integer, what is the median of variable x?

(b) To the nearest integer, what is the first quartile of variable y?

(c) Which variable has more dispersion? Why?

(d) Does the variable x have any outliers? If so, what is the value of the outliers?

2.5: Measure of the Center of the Data

A **measure of center** is a value at the center or middle of a data set such as the mean, median, mode, or midrange.

- **Mode** - the value that occurs with the greatest frequency in a data set. A data set can have no mode (no value is repeated) or several modes. It is **bimodal** when two data values occur with the same frequency and it is **multimodal** when more than two values occur with the same frequency.

- **Arithmetic Mean** - the average value of a data set found by adding the data values and dividing the total by the number of data values. Consider the following notation:
  - $\sum$ - the *sum* of a set of data values.
  - x - the variable usually used to represent the individual data values.
  - n - the number of data values in a *sample* (sample size).
  - N - the number of data values in a *population* (population size).
  - $\bar{x} = \frac{\sum x}{n}$ - the mean of a set of *sample* values (sample mean).
    - Is this value a statistic or a parameter?
  - $\mu = \frac{\sum x}{N}$ - the mean of all values in a *population* (population mean).
    - Is this value a statistic or a parameter?

Table 3: Student Test Scores.

Student	Score
Michelle	82
Ryanne	77
Bilal	90
Pam	71
Jennifer	62
Dave	68
Joel	74
Sam	84
Justine	94
Juan	88

Example 6: Table 3 represents the first exam score of 10 students enrolled in a section of Introductory Statistics.

(a) Compute the population mean.

(b) Take a sample from this class by randomly selecting four students, then compute the sample mean.

Example 7. The following data represent the monthly cell phone bill for my wife's phone for six randomly selected months.

$35.34 $42.09 $39.43 $38.93 $43.39 $49.36

Compute the mean, median, and mode monthly cell phone bill.

Compare the Mean and Median

So far, we've discussed two measures of central tendency - mean and median. Which is better? It depends.

Resistant - if extreme values (very large or very small) relative to the data do not affect its value substantially, then, the data is said to be resistant.

So, "Why would I ever compute the mean?" The mean and median are close for symmetric data, and the median is a better measure of central tendency for skewed data. However, much of the statistical inference we perform is based on the mean.

Example 8. Find the population mean or sample mean as indicated.

(a) Sample: 20, 13, 4, 8, 10

(b) Sample: 83, 65, 91, 87, 84

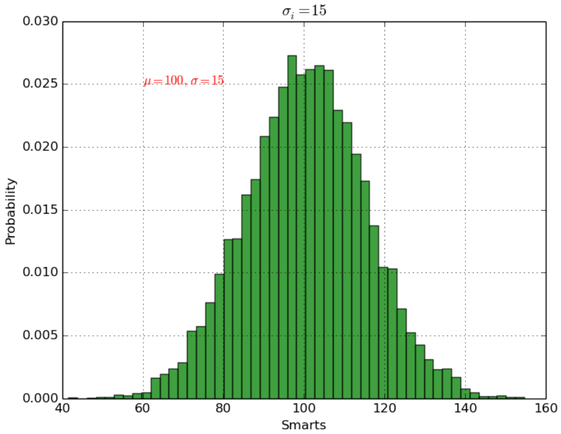
(c) Population: 3, 6, 10, 12, 14

(d) For Super Bowl XL, CBS television sold 65 ad slots for a total revenue of roughly $162.5 million. What was the mean price per ad slot?

2.6: Skewness and the Mean, Median, and Mode

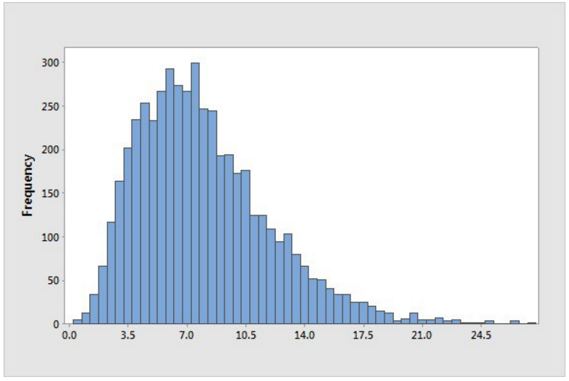
Shape of a Distribution

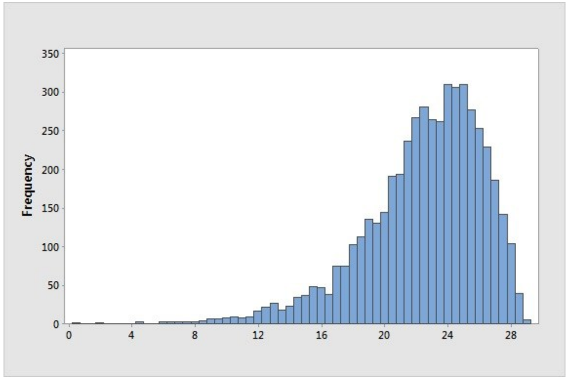
- **Bell-shaped Distribution** - the highest frequency occurs in the middle and frequencies tail off to the left and the right of the middle.



- **Right Skewed** - the tail extends to the right of the peak longer than to the left.

- **Left Skewed** - the tail extends to the left of the peak longer than to the right.





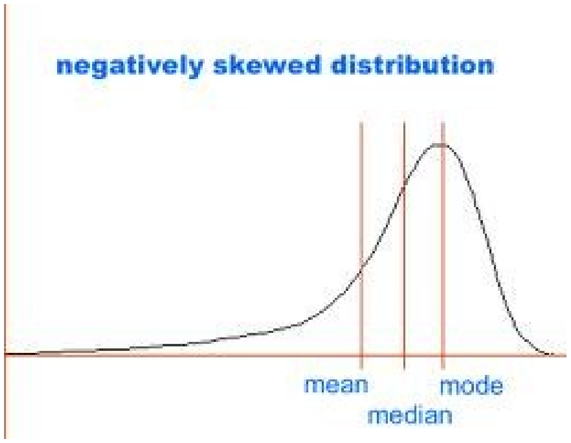
In distributions that are symmetric, the mean and the median are close in value.

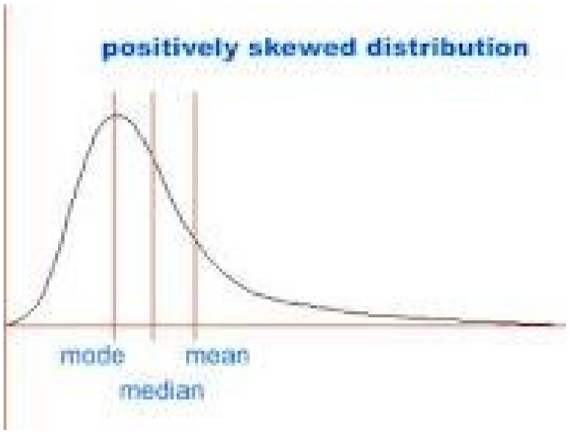
When data are either skewed right or skewed left, there are extreme values in the tail, which tend to pull the mean in the direction of the tail.

For example in skewed-right distributions, there are large observations in the right tail. These observations tend to increase the value of the mean, while having little effect on the median.

Table 4: Relation Between the Mean, Median, and Distribution Shape

Distribution Shape	Mean versus Median
Skewed left	Mean substantially smaller than median
Symmetric	Mean roughly equal to median
Skewed right	Mean substantially larger than median





2.7: Measures of the Spread of the Data

Variation - the degree to which the data are spread out.

Measures of Dispersion

- **Range, R** - the difference between the maximum data value and the minimum data value.

  $$\text{range} = \text{max. value} - \text{min. value}$$

  The range is not resistant. The range is computed using only 2 values in the data set (the largest and the smallest). The variance and standard deviation, however, use all the data in the computations.

- **Standard deviation** - a measure of the spread of the data.

  Example. Consider the standard deviations for the following two sets of numbers, both with a mean of 100.

Set	Numbers	Mean	Standard Deviation
1	100, 100, 100, 100, 100	100	0
2	90, 90, 100, 110, 110	100	10

  - In Set 1, since all the values are the same, there is no variability, so the standard deviation is 0.
  - In Set 2, one value is the mean, the other four values are 10 points away from the mean, so the average distance away from the mean is close to 10.

  Calculating the Standard Deviation

  - Population Standard Deviation:

  $$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$

  - Sample Standard Deviation:

  $$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

Important Properties of the Standard Deviation

- The standard deviation is a measure of how much data values deviate from the mean.
- The value of the standard deviation can never be negative. It's zero if all the values are exactly the same.
- Larger values of standard deviation indicates greater amounts of variation.
- Outlier(s) can drastically change the value of the standard deviation.

- **Variance** - deviation about the mean; square of the standard deviation.
  - population variance: sum of the squared deviations about the population mean divided by the number of observations in the population, N.

    $$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

  - sample variance: the sum of the squared deviations about the sample mean and divided by $n-1$.

    $$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

Step 1 Calculate $\bar{x}$, the sample mean or μ, the population mean.

Step 2 For each observation, calculate the difference between the data value and the mean: (data value - mean).

Step 3 Square each difference calculated in Step 2: (data value - mean)^2.

Step 4 Sum the squared differences calculated in Step 3: add up all the values.

Step 5 Divide this sum by $n-1$ (for sample) or N (for population). The answer for this step is called the **variance**.

Step 6 Take the square root of the variance calculated in Step 5. This is called the **standard deviation**.

Example 9. The data represent the scores on the first exam of 10 students enrolled in a section of Introductory Statisitcs. (a) Compute the range. (b) Compute the variance. (c) Compute the standard deviation.

Student	Score
1. Michelle	82
2. Ryanne	77
3. Bilal	90
4. Pam	71
5. Jennifer	62
6. Dave	68
7. Joel	74
8. Sam	84
9. Justine	94
10. Juan	88

Example 10. The Insurance Institute for Highway Safety crashed the 2007 Audi A4 four times at 5 miles per hour. The costs of repair for each of the four crashes are as follows:

$976 $2038 $918 $1899

Compute the range, sample variance, and sample standard deviation cost of repair.

Comparing Values from Different Data Sets

- **z Score** - the number of standard deviations that a given value x is above or below the mean. The z-score is unitless, it has a mean of 0 and a standard deviation of 1.

  - Population z-score: $z = \frac{x-\mu}{\sigma}$

  - Sample z-score: $z = \frac{x-\bar{x}}{s}$

Example 11: Determine whether the New York Yankees or the Philadelphia Phillies had a relatively better run-producing season. The Yankees scored 968 runs in the American League, where the mean number of runs scores was $\mu = 793.9$ and the standard deviation was $\sigma = 73.5$ runs. The Phillies scored 892 runs in the National League where the mean number of runs scores was $\mu = 763.0$ and the standard deviation was $\sigma = 58.9$.

Example 12. In 2005, babies born after a gestation period of 32 to 35 weeks had a mean weight of 2,600 grams and a standard deviation of 660 grams. In the same year, babies born after a gestation period of 40 weeks had a mean weight of 3,500 grams and a standard deviation of 470 grams. Suppose a 34-week gestation period baby weighs 2,400 grams and a 40-week gestation period baby weighs 3,300 grams. Which baby weighs less relative to the gestation period?

Example 13. Find the population variance and standard deviation or the sample variance and standard deviation as indicated.

(a) Sample: 20, 13, 4, 8, 10

(b) Sample: 83, 65, 91, 87, 84

(c) Population: 3, 6, 10, 12, 14