Chapter 2: Descriptive Statistics

2.3, 2.4: Measures of the Location of the Data, Boxplots

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.

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

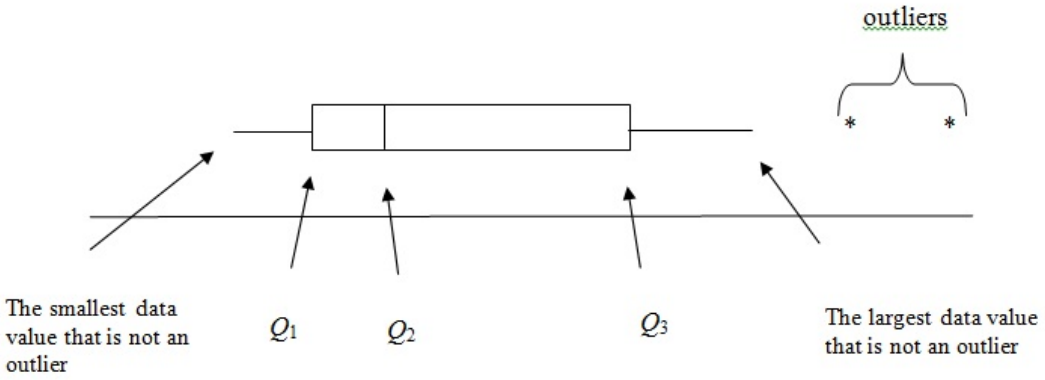
Percentiles are used to give the relative standing of an observation. Many standardized exams, such as the SAT college entrance exam, use percentiles to provide students with an understanding of how they scored on the exam in relation to all other students who took the exam.

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.

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.



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.

- **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 1: 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 1: Table **??** 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 2. Find the population mean or sample mean as indicated.

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

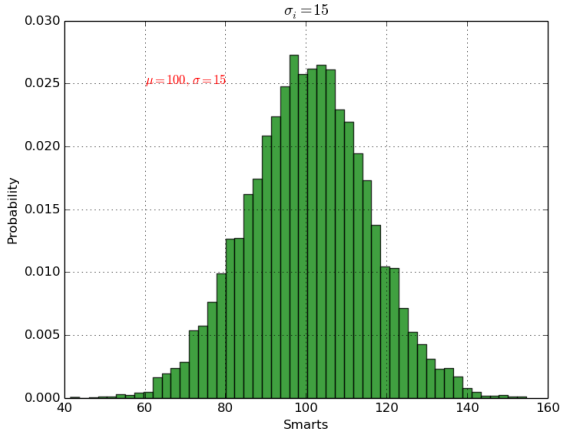
(b) Population: 3, 6, 10, 12, 14

(c) 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

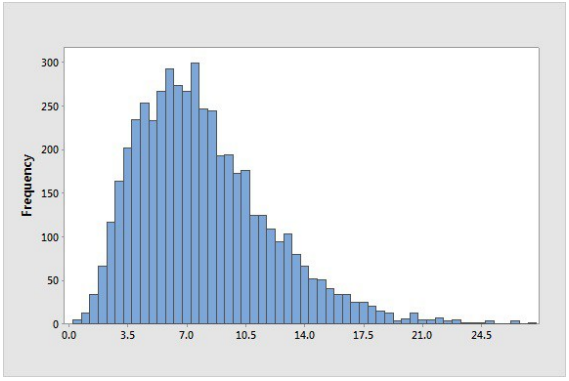
Shape of a Distribution

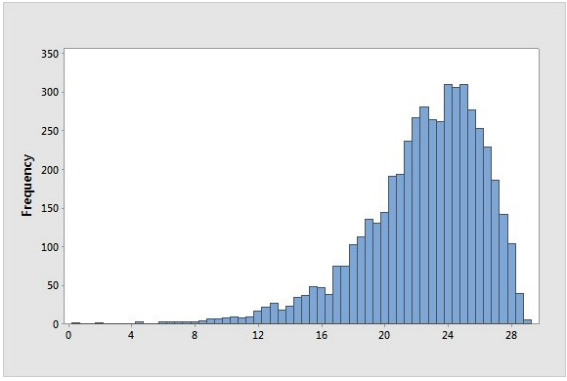
- **Bell-shaped Distribution** - the highest frequency occurs in the middle and frequencies tail off to the left and the right of the middle.



- **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.





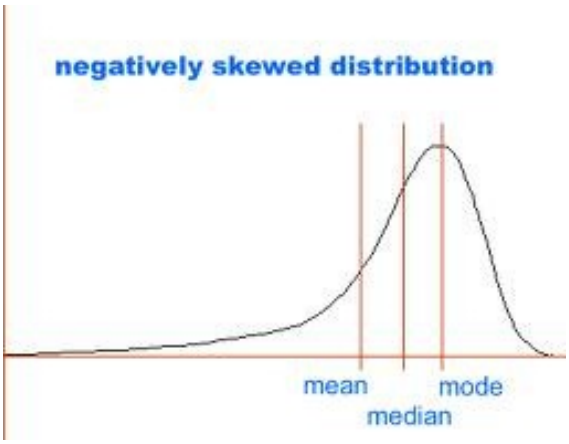
In distributions that are symmetric, the mean and the median are close in value.

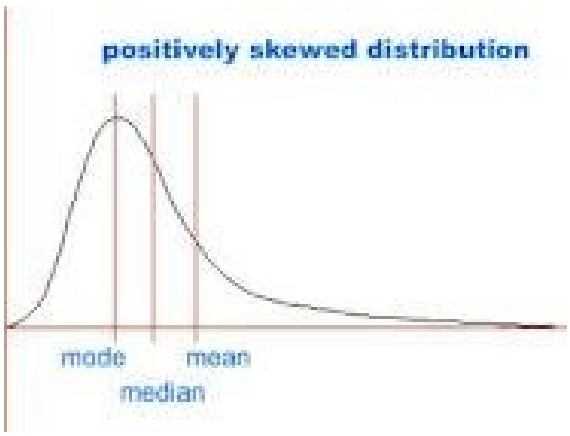
When data are either skewed right or skewed left, there are extreme values in the tail, which tend to pull the mean in the direction of the tail.

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 2: 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 3. 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

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.

$$z = \frac{x - \mu}{\sigma}$$

Example 4: 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 5. Find the population variance and standard deviation or the sample variance and standard deviation as indicated.

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

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