## Statistics and parameters

Tables, histograms and other charts are used to summarize large amounts of data. Often, an even more extreme summary is desirable. ***Statistics*** and ***parameters*** are numbers that characterize different aspects of the data

**Terminology:** A number that summarizes ***population*** data is usually referred to as a ***parameter***, while a number that summarizes ***sample*** data is called a ***statistic***.

**Comment:** There are two significant differences between *population parameters* and *sample statistics.*

- Population parameters are (more or less) constant, while sample statistics depend on the particular sample chosen—i.e., sample statistics *vary with the sample.*
- Sample statistics are *known* because we can compute them from the (available) sample data, while population parameters are often *unknown*, because data for the entire population is often unavailable.

## Measures of *central tendency*

The ‘center’ of a list of numbers can be characterized in several ways. The two most common measures of ‘middle’ are the average (or mean) and the median.

- The ***mean*** of a set of numbers is the sum of all the values divided by the number of values in the set.
- The ***median*** of a set of number is the middle number, when the numbers are listed in ascending (or descending) order. The median splits the data into two equally sized sets—50% of the data lies below the median and 50% lies above.

  (If the number of numbers in the set is *even*, then the median is the average of the two middle values.)

The mean and median are meant to describe the *typical* value in the data. Another statistic that is often used to describe the typical value is the ***mode***—this is the *most frequently occurring* value in the data.

**Example.** Find the mean, median and mode of the following set of numbers:

$$\{12, 5, 6, 8, 12, 17, 7, 6, 14, 6, 5, 16\}.$$

- The ***mean*** (average).

$$\frac{12 + 5 + 6 + 8 + 12 + 17 + 7 + 6 + 14 + 6 + 5 + 16}{12} = \frac{114}{12} = 9.5.$$

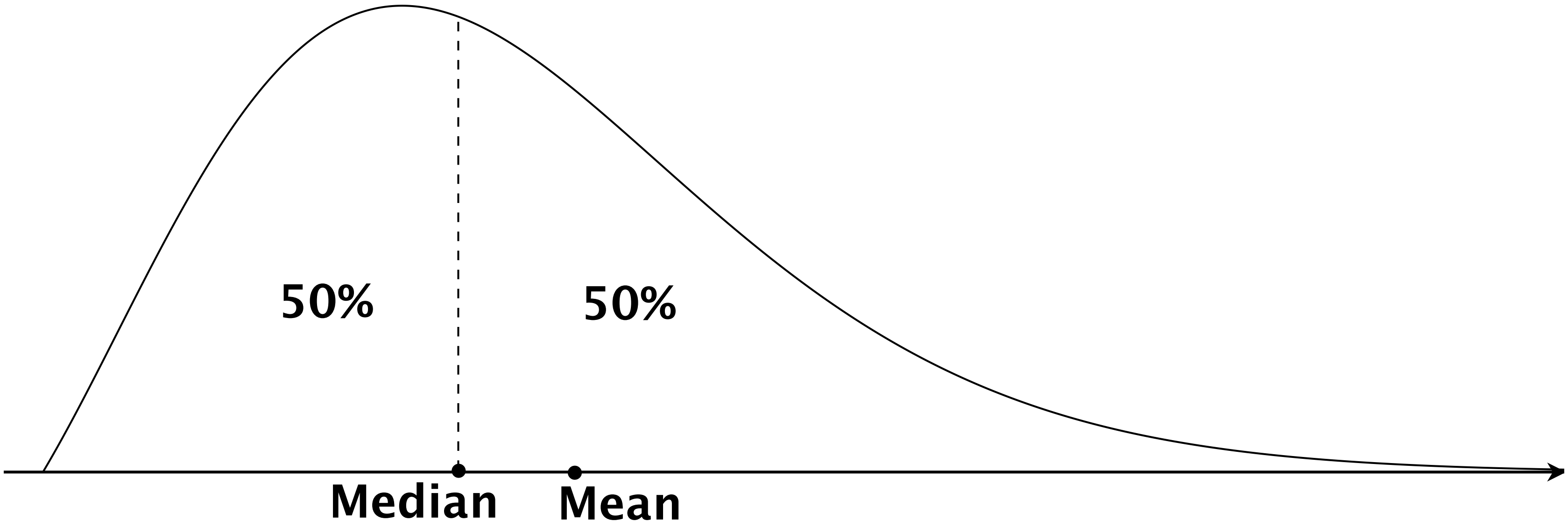
- The ***median***. Arrange the data in ascending order, and find the average of the middle two values in this case, since there are an even number of values:

$$5, 5, 6, 6, 6, 7, 8, 12, 12, 14, 16, 17 \longrightarrow \text{median} = \frac{7+8}{2} = 7.5.$$

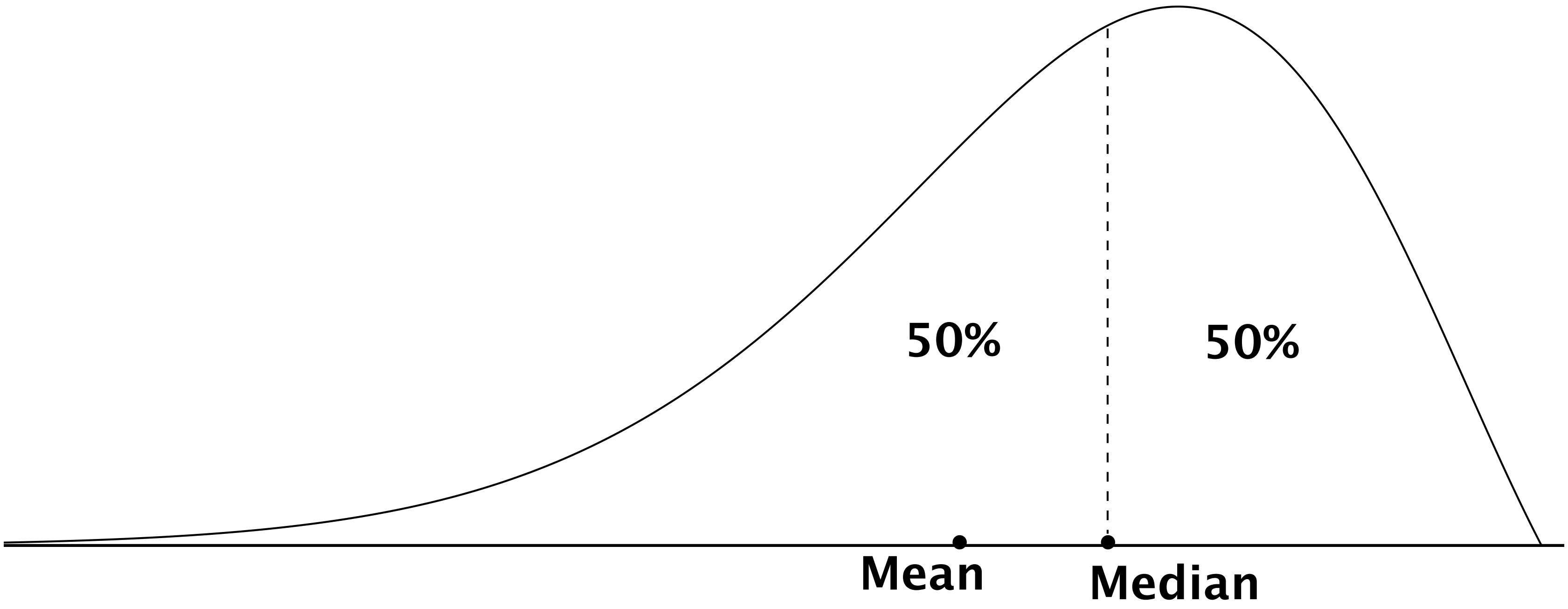
- The ***mode*** is 6, because 6 occurs most frequently (three times).

The relative positions of the mean and median provides information about how the data is distributed...

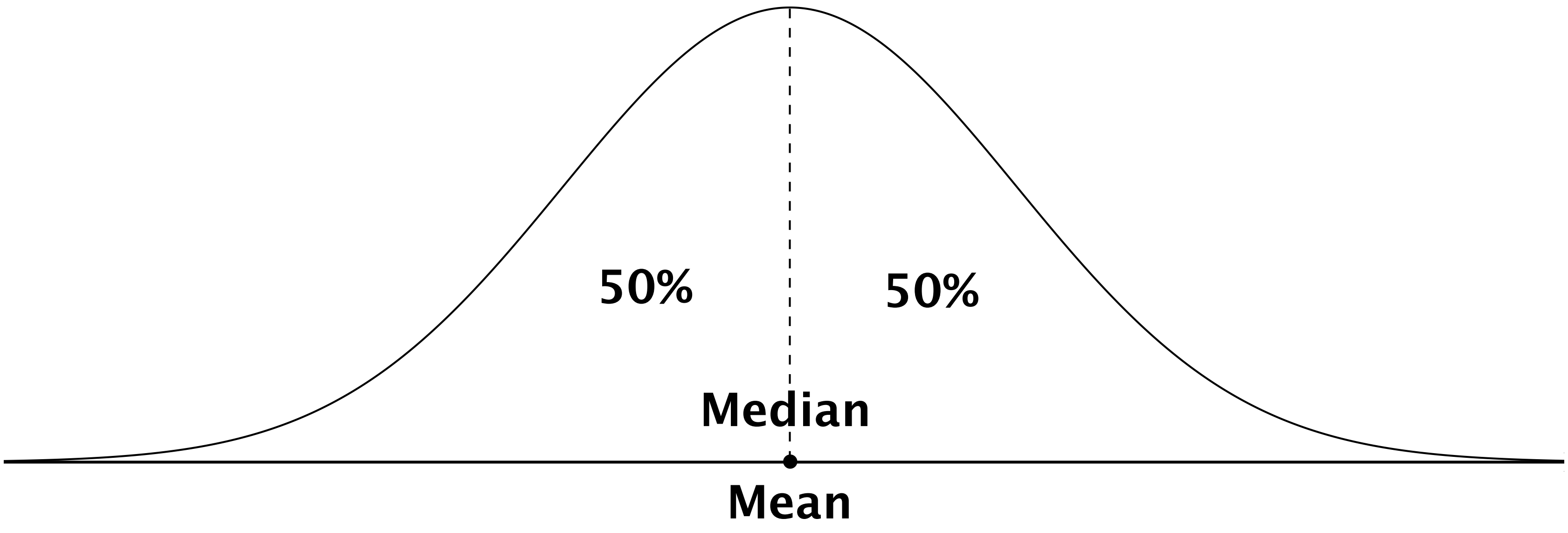
In the histogram below, the mean is bigger than the median, and the histogram has a longer tail on the right. We say that the data is *skewed to the right*.



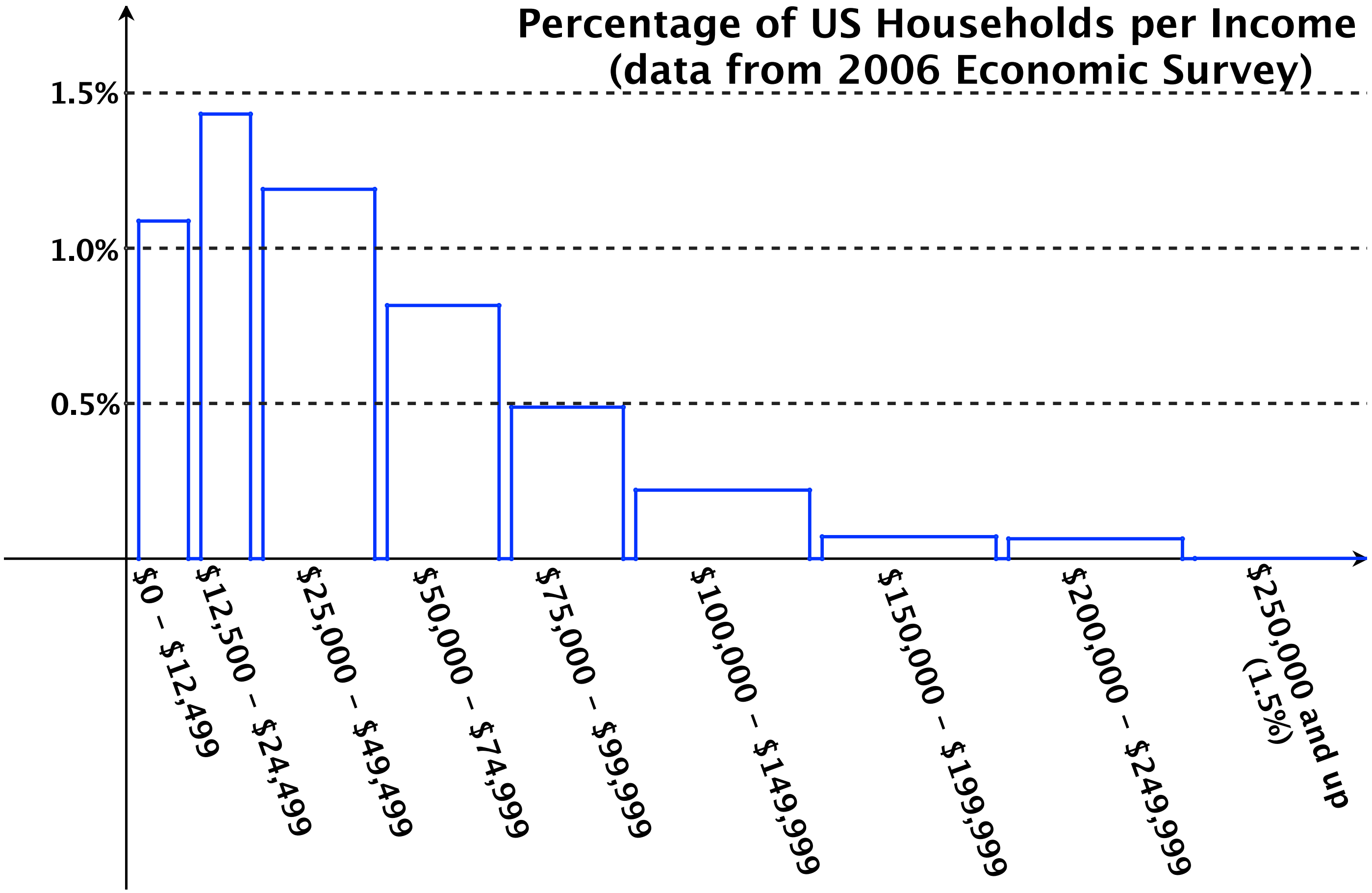
In the histogram below, the mean is smaller than the median, and the histogram has a longer tail on the left. We say that the data is *skewed to the left*.



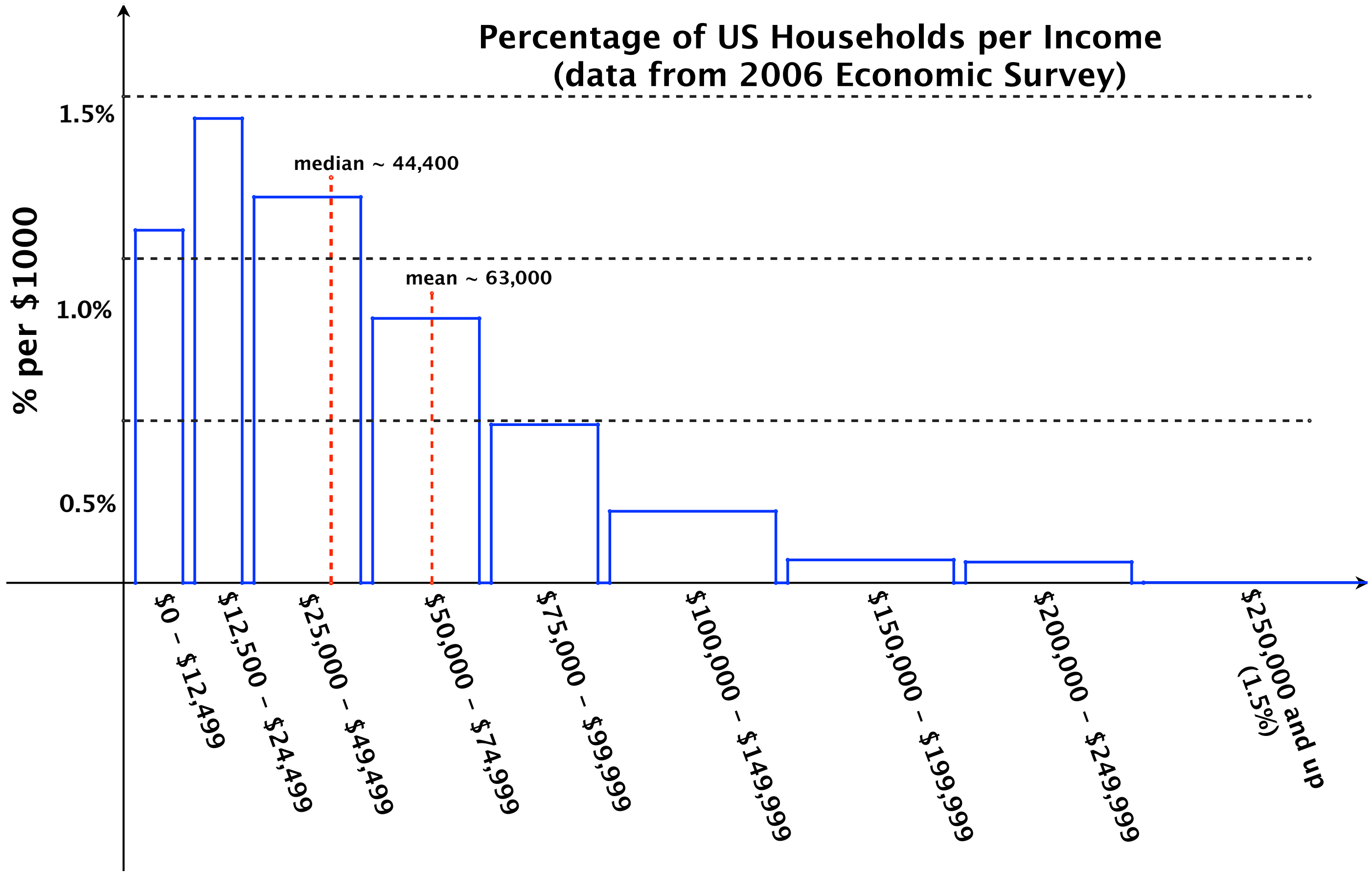
If the mean and median are (more or less) equal, then the tails of the distribution are (more or less) the same, and the data has a (more or less) symmetric distribution around the mean/median, as depicted below.



**Example:** The histogram below describes the distribution of household income in the United States. The data comes from the Wikipedia's summary of the *2006 Economic Survey*.



The distribution of household income is skewed to the right. The median household income in the survey was about $44,400. The mean income was about $63,000 (my estimate).



## Comments:

- The mean and median describe the middle of the data in somewhat different ways:
  - The median divides the histogram into two halves of equal area.
  - The mean is the '*balancing point*' of the histogram.
- The mean is more sensitive to *outliers* — data values that are much bigger or much smaller than average.
- The median gives a more 'fair' sense of middle when the data is skewed in one direction or the other. On the other hand, the mean is easier to use in mathematical formulas.
- Both the median and the mean leave out a lot of information. E.g., they tell us nothing about the spread of the data, where we might find 'peaks' in the distribution, etc.

## Notation.

- The *population* mean (a parameter) is denoted by the Greek letter $\mu$ ('mu'). If there are several variables being studied, we put a subscript on the $\mu$ to tell us which variable it pertains to. For example, if we have data for population height ($h$) and population weight ($w$), the mean height would be denoted by $\mu_h$ and the mean weight by $\mu_w$.
- The mean of a set of *sample* data (a statistic) is denoted by putting a bar over the variable. E.g., if $\{h_1, h_2, h_3, \ldots, h_n\}$ is a sample of heights, then the average of this sample would be denoted by $\overline{h}$.
- We can use ***summation notation*** to simplify the writing of (long) sums:

$$h_1 + h_2 + h_3 + \cdots + h_n = \sum_{j=1}^{n} h_j.$$

For example we can write:

$$\begin{aligned}\overline{h} &= \frac{h_1 + h_2 + \cdots + h_n}{n} \\ &= \frac{1}{n}\,(h_1 + h_2 + \cdots + h_n) = \frac{1}{n}\sum_{j=1}^{n} h_j.\end{aligned}$$

**Comment:** The point of summation notation is to simplify expressions that involve sums with many terms, or in some cases, an unspecified number of terms. The operation of addition doesn't change so all the usual rules/properties of addition continue to hold. In particular

(i) $\displaystyle\sum_{j=1}^{n}(h_j + w_j) = \sum_{j=1}^{n} h_j + \sum_{j=1}^{n} w_j$

*and*

(ii) $\displaystyle\sum_{j=1}^{n}(a\,h_j) = a\left(\sum_{j=1}^{n} h_j\right)$

## Measuring the spread of the data

The mean and median describe the middle of the data distribution. To get a better sense of where the data lies, statisticians also use '*measures of dispersion*'.

- The ***range*** is the distance between the smallest and largest values in the data.
- The ***interquartile range*** is the distance between the value separating the bottom 25% of the data from the rest and the value separating the top 25% of the data from the rest. In other words, it is the *range* of the middle 50% of the data.
- The ***standard deviation*** is *something like* the average distance of the data to its mean, or the.

**Example:** In the histogram describing household income distribution, about 25% of all households have incomes below $22,400 and about 25% of all households have incomes above $78,000, so the interquartile range is $78,000 - 22,400 = 65,600$.

## The standard deviation

As mentioned above, the standard deviation of a set of numbers is something like the average distance of the numbers from their mean. Technically, it is a little bit more complicated than that.

Suppose that $x_1, x_2, x_3, \ldots, x_n$ are numbers and $\overline{x}$ is their mean.

An obvious candidate for measuring spread is the *average distance* to the mean, (or the average *deviation* from the mean):

$$\frac{1}{n}\sum_{j=1}^{n}(x_j - \overline{x}).$$

The problem with this is *cancellation*—positive terms and negative terms in the sum can cancel each other out... How much cancellation?

$$\frac{1}{n}\sum_{j=1}^{n}(x_j - \overline{x}) = 0$$

To fix this problem, statisticians use the ***standard deviation***, which is defined by

$$SD_x = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(x_j - \overline{x})^2}.$$

In words, the SD is the ***root of the mean of the squared deviations of the numbers from their mean***.

(*) Squaring the deviations fixes the cancellation problem...

(*) ... but exaggerates both very small deviations (making them smaller) and very large deviations (making them bigger).

(*) Taking the square root of the average squared deviation fixes this problem.

(*) If a lot of the data is far from the mean, then many of the $(x_j - \overline{x})^2$ terms will be quite large, so the mean of these terms will be large and the SD of the data will be large.

(*) In particular, outliers can make the SD bigger. (Outliers have an even bigger effect on the *range* of the data.)

(*) On the other hand, if the data is all clustered close to the mean, then all of the $(x_j - \overline{x})^2$ terms will be fairly small, so their mean will be small and the SD will be small.

**Example:** Find the SD of the set $\{x_j\} = \{2, 4, 5, 8, 5, 11, 7\}$.

- Step 1. Find the mean: $\overline{x} = \frac{1}{7}\sum_{j=1}^{7} x_j = \frac{42}{7} = 6.$
- Step 2. Find the mean of the squared deviations of the numbers from their mean:

$$\frac{1}{7}\sum_{j=1}^{7}(x_j - 6)^2 = \frac{52}{7}.$$

- Step 3. $SD_x = \sqrt{52/7} \approx 2.726.$

(*) Useful shortcut (for calculations done by hand):

$$\frac{1}{n}\sum_{j=1}^{n}(x_j-\overline{x})^2 = \left(\frac{1}{n}\sum_{j=1}^{n}x_j^2\right)-(\overline{x})^2$$

so

$$SD_x = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(x_j-\overline{x})^2} = \sqrt{\left(\frac{1}{n}\sum_{j=1}^{n}x_j^2\right)-(\overline{x})^2}$$

Check with example:

$\{x_j\} = \{2,4,5,8,5,11,7\}$ and $\overline{x}=6$:

$$\left(\frac{1}{7}\sum_{j=1}^{7}x_j^2\right)-\overline{x}^2 = \frac{304}{7}-36 = \frac{52}{7} \quad \checkmark$$

**Example:** Find the SD of the set $\{y_j\} = \{21, 39, 52, 78, 51, 112, 74\}$.

- Step 1. Find the mean: $\overline{y} = \dfrac{1}{7}\displaystyle\sum_{j=1}^{7} y_j = 61.$

- Step 2. Find the mean of the squared deviations of the numbers from their mean:

$$\frac{1}{7}\sum_{j=1}^{7}(y_j - 61)^2 = \frac{5324}{7}.$$

- Step 3. $SD_y = \sqrt{5324/7} \approx 27.578.$

**Observation:** The mean and the standard deviation are both sensitive to scale. In more detail, if you change the scale of the data—you multiply each of the numbers in the data by the same amount—the mean and the standard deviation will change by the same factor.

For example, if we double each of the numbers in the previous example, we obtain the set $\{u_j\} = \{42, 78, 104, 156, 102, 224, 148\}$, i.e., $u_j = 2y_j$

The mean and standard deviation of this new set of numbers are

$$\overline{u} = \frac{42 + 78 + 104 + 156 + 102 + 224 + 148}{7} = \frac{854}{7} = 122 = 2 \cdot \overline{y}$$

and

$$SD_u = \sqrt{\frac{(42 - 122)^2 + \cdots + (148 - 122)^2}{7}} \approx 55.157 = 2 \cdot SD_y$$

## SD vs. SD$^+$

One of the most important uses of *sample statistics* is to ***estimate*** the corresponding *population parameters*.

- The mean of a ***representative*** sample is a good estimate of the mean of the population that the sample represents.
- The SD of a ***representative*** sample is a good estimate of the SD of the population that the sample represents, as long as the sample is *large*.
- If a sample is relatively small, statisticians use the SD$^+$ of the sample to estimate the SD of the population. If $n$ is sample size, then

$$SD^+ = \sqrt{\frac{n}{n-1}} \cdot SD$$

- The SD$^+$ is usually called the *sample standard deviation*.
- Many calculators with statistical functions compute the SD$^+$ instead of the SD because in practice, this is what most people want.