

Section 8.2: Measures of central tendency

When thinking about questions such as: “how many calories do I eat per day?” or “how much time do I spend talking per day?”, we quickly realize that the answer will vary from day to day and often modify our question to something like “how many calories do I consume on a typical day?” or “on average, how much time do I spend talking per day?”. In this section we will study three ways of **measuring central tendency in data**, the mean, the median and the mode. Each measure give us a single value (the mode may give more than one) that might be considered typical. As we will see however, any one of these values can give us a skewed picture if the given data has certain characteristics.

A **population** of books, cars, people, polar bears, all games played by Babe Ruth throughout his career etc.... is the entire collection of those objects. For any given variable under consideration, each member of the population has a particular value of the variable associated to them, for example the number of home runs scored by Babe Ruth for each game played by him during his career. These values are called **data** and we can apply our measures of central tendency to the entire population, to get a single value (maybe more than one for the mode) measuring central tendency for the entire population. When we calculate the mean, median and mode using the data from the entire population, we call the results the population mean, the population median and the population mode.

A **sample** is a subset of the population, for example, we might collect the data on the number of home runs scored in a random sample of 20 games played by Babe Ruth. If we calculate the mean, median and mode using the data from a sample, the results are called the sample mean, sample median and sample mode.

The Mean: The **population mean** of m numbers x_1, x_2, \dots, x_m (the data for every member of a population of size m) is denoted by μ and is computed as follows:

$$\mu = \frac{x_1 + x_2 + \dots + x_m}{m}.$$

The **sample mean** of the numbers x_1, x_2, \dots, x_n (data for a sample of size n from the population) is denoted by \bar{x} and is computed similarly:

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}.$$

Example Consider the following set of data, showing the number of times a sample of 5 students check their e-mail per day:

$$1, 3, 5, 5, 3.$$

Here $n = 5$ and $x_1 = 1, x_2 = 3, x_3 = 5, x_4 = 5$ and $x_5 = 3$.

Calculate the sample mean \bar{x} .

Example The following data shows the results for the number of books that a random sample of 20 students were carrying in their book bags:

0, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 4, 4, 4, 4, 4

Then the **mean** of the sample is the average number of books carried per student:

$$\bar{x} = \frac{0 + 1 + 1 + 2 + 2 + 2 + 2 + 2 + 2 + 2 + 2 + 2 + 2 + 2 + 3 + 3 + 3 + 3 + 4 + 4 + 4 + 4 + 4}{20} = 2.5$$

Not that the mean here cannot be an observation in our data.

Calculating the mean more efficiently: We can calculate the mean above more efficiently here by using frequencies. We can see from the calculation above that

$$\bar{x} = \frac{0 + (1 \times 2) + (2 \times 8) + (3 \times 4) + (4 \times 5)}{20} = 2.5$$

The frequency distribution for the data is:

# Books	Frequency	# Books × Frequency
0	1	0 × 1
1	2	1 × 2
2	8	2 × 8
3	4	3 × 4
4	5	4 × 5
		$\bar{x} = \frac{\text{Sum}}{20}$

The general case can be dealt with as follows: If our frequency/relative frequency table for our sample of size n , looks like the one below, (where the observations are denoted 0_i , the corresponding frequencies, f_i and the relative frequencies f_i/n):

Observation	Frequency	Relative Frequency
0_i	f_i	f_i/n
0_1	f_1	f_1/n
0_2	f_2	f_2/n
0_3	f_3	f_3/n
\vdots	\vdots	\vdots
0_R	f_R	f_R/n

then,

$$\bar{x} = \frac{0_1 \cdot f_1 + 0_2 \cdot f_2 + \dots + 0_R \cdot f_R}{n} = 0_1 \cdot \frac{f_1}{n} + 0_2 \cdot \frac{f_2}{n} + 0_3 \cdot \frac{f_3}{n} + \dots + 0_R \cdot \frac{f_R}{n}$$

We can also use our table with a new column to calculate:

Outcome	Frequency	Outcome \times Frequency
0_i	f_i	$0_i \times f_i$
0_1	f_1	$0_1 \times f_1$
0_2	f_2	$0_2 \times f_2$
0_3	f_3	$0_3 \times f_3$
\vdots	\vdots	\vdots
0_R	f_R	$0_R \times f_R$
		$\frac{\text{SUM}}{n} = \bar{x}$

Alternatively we can use the relative frequencies, instead of dividing by the n at the end.

Outcome	Frequency	Relative Frequency	Outcome \times Relative Frequency
0_i	f_i	f_i/n	$0_i \times f_i/n$
0_1	f_1	f_1/n	$0_1 \times f_1/n$
0_2	f_2	f_2/n	$0_2 \times f_2/n$
0_3	f_3	f_3/n	$0_3 \times f_3/n$
\vdots	\vdots	\vdots	\vdots
0_R	f_R	f_R/n	$0_R \times f_R/n$
			SUM = \bar{x}

You can of course choose your favorite method for calculation from the three methods listed above.

Example The number of goals scored by the 32 teams in the 2014 world cup are shown below:

18, 15, 12, 11, 10, 8, 7, 7, 6, 6, 6, 5, 5, 5, 4, 4, 4, 4, 4, 4, 3, 3, 3, 3, 3, 2, 2, 2, 2, 1, 1, 1.

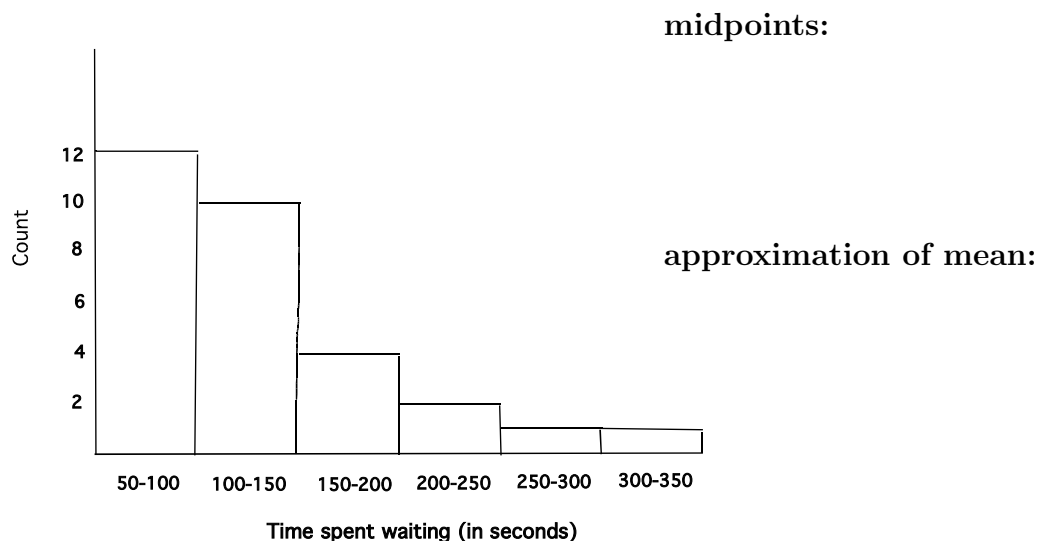
Make a frequency table for the data and taking the soccer teams who played in the world cup as a population, calculate the population mean, μ .

Estimating the mean from a histogram If we are given a histogram (showing frequencies) or a frequency table where the data is already grouped into categories and do not have access to the original data, we can estimate the mean using the midpoints of the intervals which serve as categories for the data. Suppose there are k categories (shown as the bases of the rectangles) with midpoints m_1, m_2, \dots, m_k respectively and the frequencies of the corresponding intervals are f_1, f_2, \dots, f_k , then the mean of the data set is approximately

$$\frac{m_1 f_1 + m_2 f_2 + \dots + m_k f_k}{n}$$

where $n = f_1 + f_2 + \dots + f_k$.

Example Approximate the mean for the set of data used to make the following histogram, showing the time (in seconds) spent waiting by a sample of customers at Gringotts Wizarding bank.



The Median

The Median of a set of quantitative data is the middle number when the measurements are arranged in ascending order.

To Calculate the Median: Arrange the n measurements in ascending order. We denote the median of the data by M .

1. If n is odd, M is the middle number.
2. If n is even, M is the average of the two middle numbers.

Example The number of goals scored by the 32 teams in the 2014 world cup are shown below:

18, 15, 12, 11, 10, 8, 7, 7, 6, 6, 6, 5, 5, 5, 4, 4, 4, 4, 4, 4, 3, 3, 3, 3, 3, 2, 2, 2, 2, 1, 1, 1.

Find the median of the above set of data.

Example A sample of 5 students were asked how much money they were carrying and the results are shown below:

\$75, \$2, \$5, \$0, \$5.

Find the mean and median of the above set of data. Notice that the median gives us a more representative picture here, since the mean is skewed by the outlier \$75.

The Mode

Definition The **mode** of a set of measurements is the most frequently occurring value; it is the value having the highest frequency among the measurements.

Example Find the mode of the data collected on the amount of money carried by the 5 students in the example above:

\$75, \$2, \$5, \$0, \$5.

You find that in some cases the mode is not unique.:

Example What is the mode of the data on the number of goals scored by each team in the world cup of 2006?

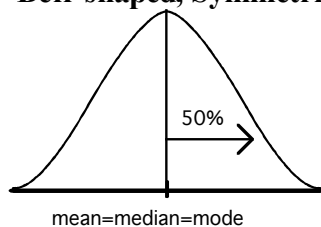
13, 12, 11, 10, 9, 9, 7, 5, 5, 5, 5, 5, 5, 4, 4, 3, 3, 3, 3, 3, 3, 2, 2, 2, 2, 2, 2, 1, 1, 1, 1, 0.

Note The mode can be computed for qualitative data. The mode is not often used as a measure of center for quantitative data.

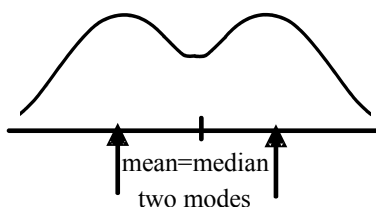
The Histogram and the mean, median and mode

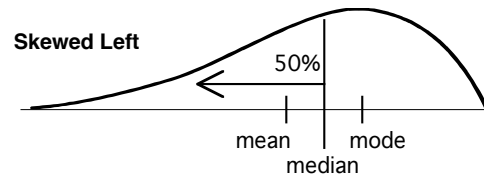
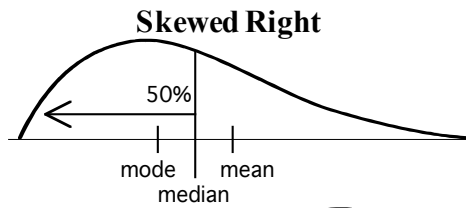
With large sets of data and narrow class widths, the histogram looks roughly like a smooth curve. The mean, median and mode, have a graphical interpretation in this case. The mean is **the balance point of the histogram of the data**, whereas **the median is the point on the x-axis such that half of the area under the histogram lies to the right of the median and half of the area lies to its left**. The **mode occurs at the data point where the graph reaches its highest point**. This of course may not be unique.

Bell-shaped, Symmetric



Bimodal





Skewed Data

Definition A data set is said to be **skewed** if one tail of the distribution has more extreme observations than the other tail.

The mean is sensitive to extreme observations, but the median is not (check out the example below).

Example Consider the data from the above example concerning the amount of money carried by the five students in the sample.

\$75, \$2, \$5, \$0, \$5.

We have already calculated the mean and the median of the data, which we found to be : mean = \$17.4, median = \$5.

Now consider the same set of data with the largest amount of money replaced by \$5,000, that is suppose our data was

\$5,000, \$2, \$5, \$0, \$5.

What is the new mean and median?

We can see from the histograms, that for data skewed to the right, the mean is larger than the median and for data skewed to left, the mean is less than the median.

Different Measures Can Give Different Impressions

The famous trio, the mean, the median, and the mode, represent three different methods for finding a so-called center value. These three values may be the same for a set of data but it is very likely that they will have three different values. When they are different, they can lead to different interpretations of the data being summarized.

Consider the annual incomes of five families in a neighborhood:

\$12,000 \$12,000 \$30,000 \$51,000 \$100,000

What is the typical income for this group?

The mean income is: \$41,000, The median income is: \$30,000, The modal income is: \$12,000.

If you were trying to promote that this is an affluent neighborhood, you might prefer to report the mean income.

If you were a Sociologist, trying to report a typical income for the area, you might report the median income.

If you were trying to argue against a tax increase, you might argue that income is too low to afford a tax increase and report the mode.