Measures of central tendency

Questions such as: "how many calories do I eat per day?" or "how much time do I spend talking per day?" can be hard to answer because the answer will vary from day to day. It's sometimes more sensible to ask "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* that might be considered typical. Each measure has its own strengths and weaknesses

*: some exclusions apply

Measures of central tendency

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; or we can apply our measures to a subset or sample of the population, to get an estimate of the central tendency for the population.

Measures of central tendency

A **sample** is a subset of the population, for example, we might collect data on the number of home runs hit by Miguel Cabrera in a random sample of 20 games. 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*.

Sometimes we can look at the entire population, not just a subset. For example, since Babe Ruth has now retired, so we might collect data on the number of home runs he hit in his career. If we calculate the mean, median and mode using the data collected from the entire population, the results are called the *population mean*, *population median* and *population mode*.

Mean

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

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

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

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

Mean

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 $\bar{x}$.

$$\frac{1+3+5+5+3}{5} = \frac{17}{5} = 3.4$$

Mean

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, 3, 3, 3, 3, 4, 4, 4, 4, 4

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

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

Note that the mean does not necessarily have to be one of the values observed in our data; in this case it is a value that could *never* be observed.

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
		$\bar{x} = \frac{\text{Sum}}{20} = \frac{50}{20} = 2.5$

Calculating the mean more efficiently

In general: If the frequency/relative frequency table for our sample of size n looks like the one below (where the observations are denoted O_i, the corresponding frequencies by f_i and the relative frequencies by f_i/n):

Observation O_i	Frequency f_i	Relative Frequency f_i/n
O_1	f_1	f_1/n
O_2	f_2	f_2/n
O_3	f_3	f_3/n
$\vdots$	$\vdots$	$\vdots$
O_R	f_R	f_R/n

then:

Calculating the mean more efficiently

$$\bar{x} = \frac{O_1 \cdot f_1 + O_2 \cdot f_2 + \cdots + O_R \cdot f_R}{n} =$$

$$O_1 \cdot \frac{f_1}{n} + O_2 \cdot \frac{f_2}{n} + O_3 \cdot \frac{f_3}{n} + \cdots O_R \cdot \frac{f_R}{n}$$

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

Outcome O_i	Frequency f_i	Outcome $\times$ Frequency $O_i \times f_i$
O_1	f_1	$O_1 \times f_1$
O_2	f_2	$O_2 \times f_2$
O_3	f_3	$O_3 \times f_3$
$\vdots$	$\vdots$	$\vdots$
O_R	$.f_R$	$O_R \times f_R$
		$\frac{\text{SUM}}{n} = \bar{x}$

Calculating the mean more efficiently

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

Outcome O_i	Frequency f_i	Relative Frequency f_i/n	Outcome $\times$ Relative Frequency $O_i \times f_i/n$
O_1	f_1	f_1/n	$O_1 \times f_1/n$
O_2	f_2	f_2/n	$O_2 \times f_2/n$
O_3	f_3	f_3/n	$O_3 \times f_3/n$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
O_R	$.f_R$	f_R/n	$O_R \times f_R/n$
			SUM $= \bar{x}$

You can of course choose any method for calculation from the three methods listed above. The easiest method to use will depend on how the data is presented.

Calculating the mean more efficiently

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.

Outcome	Frequency
1	3
2	4
3	5
4	6
5	3
6	3
7	2

Outcome	Frequency
8	1
10	1
11	1
12	1
15	1
18	1
$\mu =$	?

Calculating the mean more efficiently

Outcome	Frequency
1	3
2	4
3	5
4	6
5	3
6	3
7	2

Outcome	Frequency
8	1
10	1
11	1
12	1
15	1
18	1
$\mu =$	5.34375

$$\mu = \frac{1 \cdot 3 + 2 \cdot 4 + 3 \cdot 5 + 4 \cdot 6 + 5 \cdot 3 + 6 \cdot 3 + 7 \cdot 2 + 8 \cdot 1 + 10 \cdot 1 + 11 \cdot 1 + 12 \cdot 1 + 15 \cdot 1 + 18 \cdot 1}{32}$$

$$= \frac{3 + 8 + 15 + 24 + 15 + 18 + 14 + 8 + 10 + 11 + 12 + 15 + 18}{32} = \frac{171}{32} = 5.34375$$

Estimating the mean from a histogram

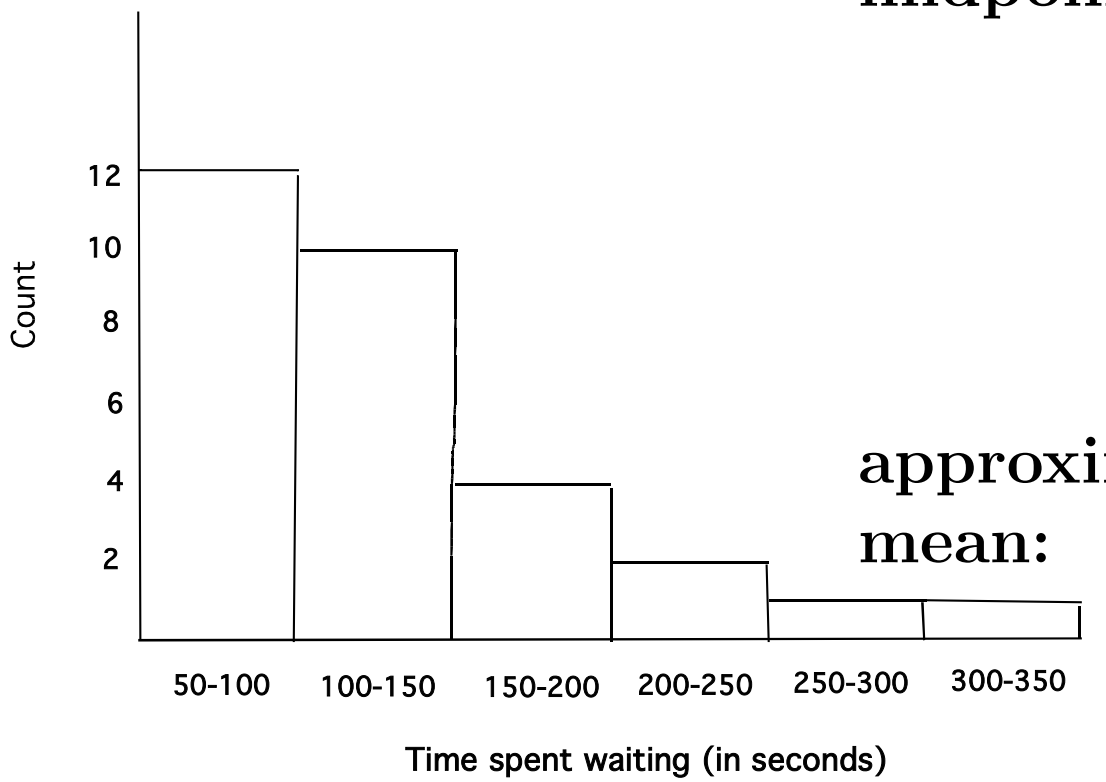
If we are given a histogram (showing frequencies) or a frequency table where the data is already grouped into categories, and we do not have access to the original data, we can still estimate the mean using the midpoints of the intervals which serve as categories for the data. Suppose there are k categories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

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

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

Estimating the mean from a histogram

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.



midpoints:

approximation of sample mean:

Estimating the mean from a histogram

midpoints: $\dfrac{50+100}{2} = 75$ $\dfrac{100+150}{2} = 125$ $\dfrac{150+200}{2} = 175$ $\dfrac{200+250}{2} = 225$ $\dfrac{250+300}{2} = 275$ $\dfrac{300+350}{2} = 325$

Outcome	Frequency
75	12
125	10
175	4
225	2
275	1
325	1
Sample size	30

Estimating the mean from a histogram

$$\bar{x}_{\text{approx}} = \frac{75 \cdot 12 + 125 \cdot 10 + 175 \cdot 4 + 225 \cdot 2 + 275 \cdot 1 + 325 \cdot 1}{30}$$

$$= \frac{900 + 1250 + 700 + 450 + 275 + 325}{30} = \frac{3900}{30} = 130$$

This calculation only gives an approximation to the sample mean because I do not know the distribution of actual wait times within each bar (cf. the two histograms for Old Faithful eruption durations in the previous section's slides).

Estimating the mean from a histogram

We can calculate the minimum possible sample mean by assuming all the people in each bar are at the left hand edge. For example, all 12 people in the first bar waited 50 seconds. This gives a result of $\bar{x}_{\text{min}} = 105$.

We can also calculate the maximal possible sample mean by assuming all the people in each bar are at the right hand edge. This gives the result $\bar{x}_{\text{max}} = 155$.

Notice

$$\bar{x}_{\text{approx}} = \frac{\bar{x}_{\text{min}} + \bar{x}_{\text{max}}}{2}$$

and the actual sample mean, $\bar{x}$ satisfies the inequalities

$$\bar{x}_{\text{min}} \leqslant \bar{x} \leqslant \bar{x}_{\text{max}}$$

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 (or descending) 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.

The Median

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.

The data is in descending order. There are 32 events and half of 32 is 16. Sixteen elements from the right is 4, indicated in green in the list below. Sixteen elements from the left is 4, indicated in red in the list below. The median is $4 = \dfrac{4+4}{2}$.

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

The Median

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

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

Find the mean and median of the above set of data.

The data in ascending order is 0, 2, 5, 5, 75. The median is $\dfrac{0+2+5+5+75}{5} = \dfrac{87}{5} = 17.4$. There are $5 = 2 \cdot 3 - 1$ numbers so to find the median count in 3 from either end to get 5.

Notice that the median gives us a more representative picture here, since the mean is skewed by the outlier \$75.

The Mode

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:

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

Since 5 occurs twice and all the other events are unique, the mode is 5.

The Mode

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

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

Here is the frequency table:

18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1
1	0	0	1	0	0	1	1	1	0	1	2	3	3	6	5	4	3

To find the mode, look in the frequency table for the largest number(s) there. In this case 4 occurs 6 times and no other entry occurs this many times so the mode is 4.

The Mode

Notes 1) The mode need not be unique: if over the course of a week I drink 3,2,2,3,4,1,1 cups of coffee per day, the mode number of cups I drink is 1, and 2, and 3.

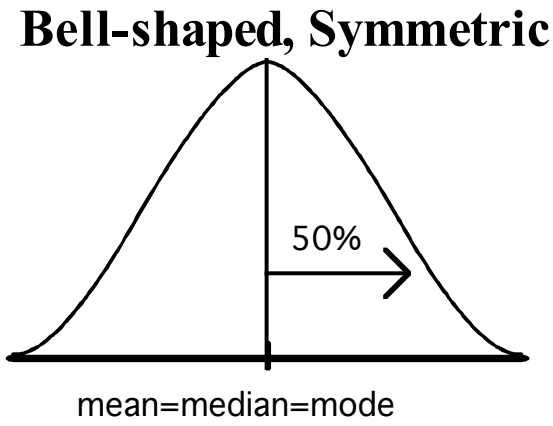
The mode can be computed for qualitative data. For example, if in this class we have 11 people with blue eyes, 6 with green eyes, 5 with hazel eyes, 4 with brown eyes and 1 with purple eyes, then the mode eye color is blue, but it makes no sense to talk about the mean or median eye colour.

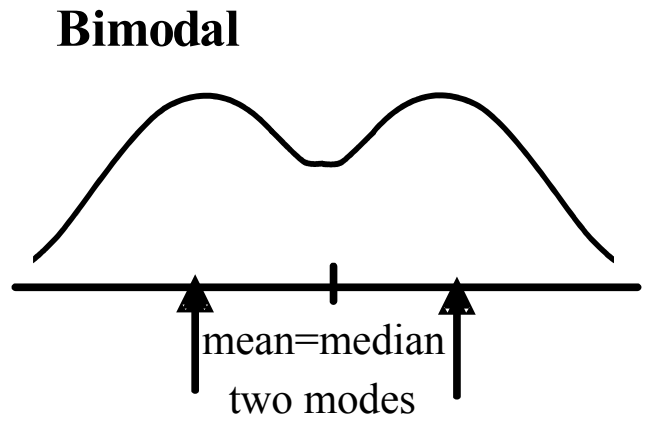
The Histogram and the mean, median and mode

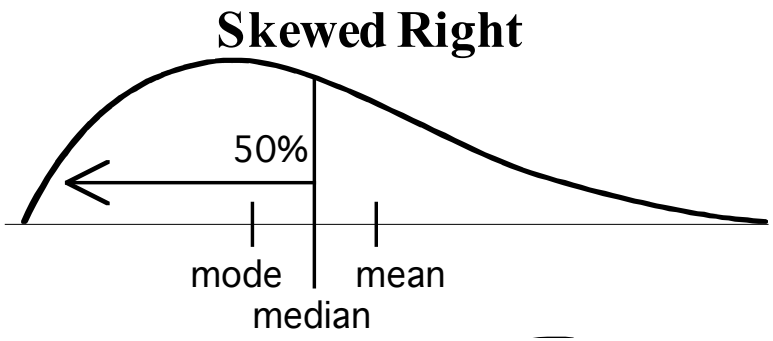
With large sets of data and narrow class widths, the histogram looks roughly like a smooth curve. The mean, median and mode, have a graphical interpretation in this case.

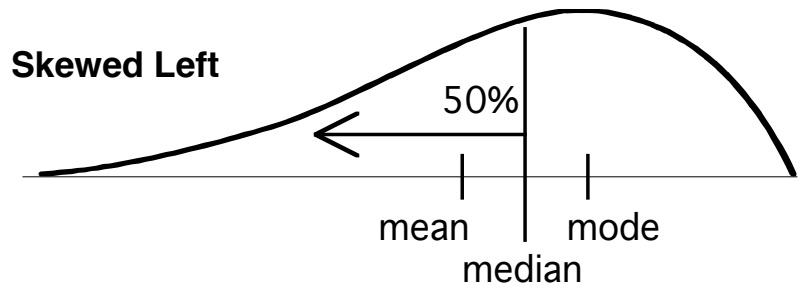
- The mean is the balance point of the histogram of the data (the point on the horizontal axis from where I would pick up the histogram to balance it on my finger)
- the median is the point on the horizontal axis such that half of the area under the histogram lies to the right of the median and half of the area lies to its left.
- The mode occurs at the data point where the graph reaches its highest point (maybe not unique).

The Histogram and the mean, median and mode









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.

Skewed Data

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.

Consider the previous example with data on the amount of money carried by a sample of five students:

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

We have already calculated mean = \$17.4, median = \$5.

Now consider the same set of data with the largest amount of money replaced by \$5,000:

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

What is the new mean and median? The median is the same, 5 but the mean is $(5000 + 2 + 5 + 0 + 5)/5 = 1002.4$

Comparing different measures

The mean, the median, and the mode represent three different methods for finding a central value of some data, or a "measure of central tendancy". 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:

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

What is the typical income for this group?

Comparing different measures

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

- The mean income is: \$41,000,
- The median income is: \$30,000,
- The mode 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 property tax increase, you might argue that income is too low to afford a tax increase and report the mode.