5 Mean, Median, Mode; Outliers

A measure of central tendency is an average: a single value intended to be typical of a data set. It’s a value intended to answer the question “Where is the middle of most of the data?”

→ These are for univariate data only
→ For a numerical data set, a measure of central tendency tells you about where the middle of the data lies on the number line.

Notation: Number of data points in a population: \( N \)
Number of data points in a sample: \( n \)

We will study three measures of central tendency: mean, median, and mode.

Mean

→ For quantitative data only

If the data are \( x_1, x_2, \ldots, x_k \), then the mean is \( \frac{x_1 + \cdots + x_k}{k} \) (for population or sample).

Notation: population mean: \( \mu \)
sample mean: \( \bar{x} \)

The mean has the same units as the data.

E.g. Find the mean of the following data.

\[
52, 52, 54, 57, 58, 63, 63, 65, 67, 67, 70, 71, 72, 73, 75, 76, 77, 93
\]

Solution: \( \mu = \frac{52 + 52 + 54 + 57 + \cdots + 77 + 93}{18} = \frac{1205}{18} \approx 66.94 \)

Think of \( \bar{x} \) as an estimate of \( \mu \). For any one sample, it might happen that \( \bar{x} \) is far from \( \mu \), but it can be proved that if many samples are taken from the same population, then we can expect the average value of \( \bar{x} \) to be very close to \( \mu \). This means that \( \bar{x} \) is an unbiased estimator for \( \mu \).

Median

→ For quantitative data only

The median \( Q_2 \) is the number halfway up a sorted list of data.

You will need to know that the middle position in a list with \( k \) entries is the \( \frac{k+1}{2} \) position. (This is not the same as the middle entry.)

To find the median of a dataset with \( k \) values:

1. Sort the data in ascending order.
2. If \( k \) is odd, then the median is the middle data value.
   If \( k \) is even, then the median is the mean of the two middle values.

E.g. Find the median of the data set 2, 7, 3, 4, 8.

1. Sort the data in ascending order to get 2, 3, 4, 7, 8
2. The middle position is obviously the third
3. The median is the value in the middle position, namely 4
Find the median of the following data.

52, 52, 54, 57, 58, 63, 65, 67, 70, 71, 72, 73, 75, 76, 77, 93

1. Sort the data in ascending order (already done)
2. There are \( k = 18 \) data points, so the middle position is the \( \frac{18+1}{2} = 9.5 \) position. Obviously there is no 9.5 position; the median is the average of the values in the 9th and 10th positions.
3. The median is thus \( \frac{67+67}{2} = 67 \)

Either the mean or the median may be called an **average**.

**Mode**

If there is a data value that occurs most frequently, it is called the **mode** of the data. If there are two values that occur most frequently, the data is **bimodal** and we report both values. If there are more than two values that occur most frequently, the data is **multimodal**, and we report no mode.

**Example** Find the mode(s) of each data set.

(a) a, a, b, c, c, c, d, d, e, f, g [mode: c]
(b) a, b, c, d, d, e, e, f, g [mode: d and e, bimodal]
(c) a, b, c, c, d, d, e, f, f, g, g, h [no mode: more than two points occur the largest number of times]
(d) a, b, c, d, e [no mode]

Note that the mode is the data value, not its frequency.

Bimodal data is often a sign that your measurements are really from two populations.

**Outliers**

There’s a problem with data points that are far away from most of the data.

**Example** A certain neighborhood contains 50 houses of which 49 are occupied. Oddly, the households in the neighborhood have net worths of exactly $1,000, $2,000, $49,000. What are the mean and median net worth amounts for the neighborhood? How will these change when Jeff Bezos, whose net worth is $160,000,000,000, moves into the 50th house? Which one better represents the neighborhood average?

Before Mr. Bezos moved in, the mean was \( \frac{1000+2000+\ldots+49000}{49} = $25,000 \) and the median was the same amount. Once he’s in the neighborhood, the mean becomes \( \frac{1000+2000+\ldots+49000+160000000000}{50} \approx $3,200,024,500 \). The median only changes to $25,500, so it is a much better “typical value” than the mean in this case.

A data value is a (suspected) **outlier** if it is extremely high or low compared to the rest of the data.

A histogram for a data set with an outlier often looks like the one below.
If there are at least two bars missing between the bulk of the data and a single outlying data point, we will consider the point to be an outlier.

Our only method for finding outliers is to look at a data plot or histogram. There are numerical methods, but we won’t study them in this course.

**How to choose the measure of central tendency to use**

As the example showed, the mean is strongly affected by outliers, but the median isn’t. To choose the measure of central tendency to use, go down the following list and use the first rule that fits.

1. If the data set contains qualitative data, use the mode.
2. If there is an outlier (or two) in a set of data, use the median.
3. In all other situations, use the mean.

We usually use the TI calculator’s 1-Var Stats function to find the mean and median of data. See the book for instructions.

**Weighted mean**

→For quantitative data only

Sometimes we have data from categories some of which are more important than others. E.g., when computing GPAs, four-credit courses are weighted more heavily than three-credit courses. In these cases, we must compute a *weighted mean*. If there are \( k \) categories, the weights of the categories are \( w_1, w_2, \ldots, w_k \), and the data values in those categories are \( x_1, x_2, \ldots, x_k \), then the weighted mean is

\[
\bar{x}_w = \frac{x_1w_1 + x_2w_2 + \cdots + x_kw_k}{w_1 + w_2 + \cdots + w_k}
\]

Note: Some authors use \( \bar{x} \) to denote weighted means.

E.g. Last semester, Jasmine got an A in Political Science, a B in English, a C in Biology, an A in Movie Appreciation, and a D in Math. All her classes were three-credit classes except Biology, which was a four-credit class. Assuming the usual scale (A = 4, B = 3, etc.), what is her GPA?

<table>
<thead>
<tr>
<th>Class</th>
<th>Grade</th>
<th>Grade Value (value)</th>
<th>Credits (weight)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poli Sci</td>
<td>A</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>English</td>
<td>B</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Biology</td>
<td>C</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Mov App</td>
<td>A</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Math</td>
<td>D</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>

\[
\text{GPA} = \frac{4\cdot3 + 3\cdot3 + 2\cdot4 + 4\cdot3 + 1\cdot3}{3 + 3 + 4 + 3 + 3} = 2.75
\]