03 Measures of Middle and Spread

3.1 Measures of central tendency: mode, median, mean, midrange

Mode

The mode is the value that occurs most frequently in the data. Spreadsheet programs such as Microsoft Excel or OpenOffice.org Calc can determine the mode with the function MODE.

=MODE(data)

In the Fall of 2000 the statistics class gathered data on the number of siblings for each member of the class. One student was an only child and had no siblings. One student had 13 brothers and sisters. The complete data set is as follows:

1,2,2,2,2,2,3,3,4,4,4,5,5,5,7,8,9,10,12,12,13

The mode is 2 because 2 occurs more often than any other value. Where there is a tie there is no mode.

For the ages of students in that class

18,19,19,20,21,21,21,22,22,22,22,22,23,23,24,24,25,25,26

...there is no mode: there is a tie between 21 and 22, hence there no single must frequent value. Spreadsheets will, however, usually report a mode of 21 in this case. Spreadsheets often select the first mode in a multi-modal tie.

If all values appear only once, then there is no mode. Spreadsheets will display #N/A or #VALUE to indicate an error has occurred - there is no mode. No mode is NOT the same as a mode of zero. A mode of zero means that zero is the most frequent data
value. Do not put the number 0 (zero) for "no mode." An example of a mode of zero might be the number of children for students in statistics class.

**Median**

The median is the central (or *middle*) value in a data set. If a number sits at the middle, then it is the median. If the middle is between two numbers, then the median is half way between the two middle numbers.

For the sibling data...

1,2,2,2,2,2,3,3,4,4,|4|,5,5,5,7,8,9,10,12,12,13

...the median is 4.

Note the data must be in order (sorted) before you can find the median. For the data 2, 4, 6, 8 the median is 5: (4+6)/2.

The median function in spreadsheets is MEDIAN.

= MEDIAN(data)

**Mean (average)**

The mean, also called the arithmetic mean and also called the average, is calculated mathematically by adding the values and then dividing by the number of values (the sample size n).

If the mean is the mean of a population, then it is called the population mean μ. The letter μ is a Greek lower case "m" and is pronounced "mu."

If the mean is the mean of a sample, then it is the sample mean x. The symbol x is pronounced "x bar."

\[
\text{population mean } \mu = \frac{\text{sum of the population data}}{\text{population size } N} = \frac{\Sigma X}{N}
\]

\[
\text{sample mean } x = \frac{\text{sum of the sample data}}{\text{sample size } n} = \frac{\Sigma x}{n}
\]
The sum of the data \( \sum x \) can be determined using the function \( \text{=SUM(data)} \). The sample size \( n \) can be determined using \( \text{=COUNT(data)} \). Thus \( \text{=SUM(data)/COUNT(data)} \) will calculate the mean. There is also a single function that calculates the mean. The function that directly calculates the mean is \( \text{=AVERAGE(data)} \).

**Resistant measures:** One that is not influenced by extremely high or extremely low data values. The median tends to be more resistant than mean.

**Population mean and sample mean**

If the mean is measured using the whole population then this would be the population mean. If the mean was calculated from a sample then the mean is the sample mean. Mathematically there is no difference in the way the population and sample mean are calculated.

**Midrange**

The midrange is the midway point between the minimum and the maximum in a set of data. The midrange is calculated from:

\[
\text{midrange} = \frac{\text{maximum} + \text{minimum}}{2}
\]

In a spreadsheet use the following formula:

\[
\text{=(MAX(data)+MIN(data))/2}
\]

### 3.2 Differences in the Distribution of Data

**Range**

Consider the following data:

<table>
<thead>
<tr>
<th>Data set 1</th>
<th>Data set 2</th>
<th>Data set 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>5, 5, 5, 5</td>
<td>2, 4, 6, 8</td>
<td>2, 2, 8, 8</td>
</tr>
<tr>
<td>mode</td>
<td>median</td>
<td>mean ( \mu )</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>8</td>
</tr>
</tbody>
</table>

Neither the mode, median, nor the mean reveal clearly the differences in the distribution of the data above. The mean and the median are the same for each data set. The mode is the same as the mean and the median for the first data set and is unavailable for the
last data set (spreadsheets will report a mode of 2 for the last data set). A single number that would characterize how much the data is spread out would be useful.

The range is one way to capture the spread of the data. The range is calculated by subtracting the smallest value from the largest value. In a spreadsheet:

=MAX(data) − MIN(data)

The range still does not characterize the difference between set 2 and 3: the last set has more data further away from the center of the data distribution. The range misses this difference.

To capture the spread of the data we use a measure related to the average distance of the data from the mean. We call this the standard deviation. If we have a population, we report this average distance as the population standard deviation. If we have a sample, then our average distance value may underestimate the actual population standard deviation. As a result the formula for sample standard deviation adjusts the result mathematically to be slightly larger. For our purposes these numbers are calculated using spreadsheet functions.

**Standard deviation**

One way to distinguish the difference in the distribution of the numbers in data set 2 and data set 3 above is to use the standard deviation.

<table>
<thead>
<tr>
<th>Data</th>
<th>mean µ</th>
<th>stdev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data set 1</td>
<td>5, 5, 5</td>
<td>0.00</td>
</tr>
<tr>
<td>Data set 2</td>
<td>2, 4, 6, 8</td>
<td>2.58</td>
</tr>
<tr>
<td>Data set 3</td>
<td>2, 2, 8, 8</td>
<td>3.46</td>
</tr>
</tbody>
</table>

The function that calculates the sample standard deviation is:

=STDEV(data)

In this text the symbol for the sample standard deviation is usually sx.
In this text the symbol for the population standard deviation is usually σ.

The symbol sx usually refers the standard deviation of single variable x data. If there is y data, the standard deviation of the y data is sy. Other symbols that are used for standard deviation include s and σx. Some calculators use the unusual and confusing notations σxn−1 and σxn for sample and population standard deviations.
In this class we always use the sample standard deviation in our calculations. The sample standard deviation is calculated in a way such that the sample standard deviation is slightly larger than the result of the formula for the population standard deviation. This adjustment is needed because a population tends to have a slightly larger spread than a sample. There is a greater probability of outliers in the population data.

**Coefficient of variation CV**

The Coefficient of Variation is calculated by dividing the standard deviation (usually the sample standard deviation) by the mean.

\[ \text{CV} = \frac{\text{STDEV(data)}}{\text{AVERAGE(data)}} \]

Note that the CV can be expressed as a percentage: Group 2 has a CV of 52% while group 3 has a CV of 69%. A deviation of 3.46 is large for a mean of 5 (3.46/5 = 69%) but would be small if the mean were 50 (3.46/50 = 7%). So the CV can tell us how important the standard deviation is relative to the mean.

**Rules of thumb regarding spread**

As an approximation, the standard deviation for data that has a symmetrical, heap-like distribution is roughly one-quarter of the range. If given only minimum and maximum values for data, this rule of thumb can be used to estimate the standard deviation.

At least 75% of the data will be within two standard deviations of the mean, regardless of the shape of the distribution of the data.

At least 89% of the data will be within three standard deviations of the mean, regardless of the shape of the distribution of the data.

If the shape of the distribution of the data is a symmetrical heap, then as much as 95% of the data will be within two standard deviations of the mean.

Data beyond two standard deviations away from the mean is considered "unusual" data.

**Basic statistics and their interaction with the levels of measurement**

<table>
<thead>
<tr>
<th>Level of measurement</th>
<th>Appropriate measure of middle</th>
<th>Appropriate measure of spread</th>
</tr>
</thead>
<tbody>
<tr>
<td>nominal</td>
<td>mode</td>
<td>none or number of categories</td>
</tr>
<tr>
<td>ordinal</td>
<td>median</td>
<td>range</td>
</tr>
<tr>
<td>interval</td>
<td>median or mean</td>
<td>range or standard deviation</td>
</tr>
</tbody>
</table>
At the interval level of measurement either the median or mean may be more appropriate depending on the specific system being studied. If the median is more appropriate, then the range should be quoted as a measure of the spread of the data. If the mean is more appropriate, then the standard deviation should be used as a measure of the spread of the data.

Another way to understand the levels at which a particular type of measurement can be made is shown in the following table.

<table>
<thead>
<tr>
<th>Levels at which a particular statistic or parameter has meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Level of measurement</strong></td>
</tr>
<tr>
<td>Nominal</td>
</tr>
<tr>
<td>sample size</td>
</tr>
</tbody>
</table>

For example, a mode, median, and mean can be calculated for ratio level measures. Of those, the mean is usually considered the best measure of the middle for a random sample of ratio level data.