II. Descriptive Statistics

B. Numerical Descriptions for Univariate Data

In this section
- Measures of Central Tendency
- Measures of Position
- Measures of Dispersion (Variability)

1. Measures of Central Tendency

In this section we will mainly deal with statistics. Remember a statistic is a numerical characteristic of a sample. These are descriptive statistics since they will summarize the data in the sample.

A measure of central tendency is a measure of average or typical value. The three most common measures of central tendency is what we will discuss: the mean, median, and mode.

Sample Mean – when most people use the word average, they are talking about the mean

\[ \bar{X} = \frac{\sum x}{n} \]

where \( \bar{X} \) is the sample mean
\( \sum \) is the sum
\( x \) is the data values
\( n \) is the sample size

The mean is the sum of the data values divided by the sample size.

Example 1
Select 4 students and ask “how many brothers and sisters do you have?”
Suppose the sample yields the following Data: 2,3,1,3
Calculate the mean.
Do you think the mean is a good measure of center for this data?

Example 2
Suppose we had selected a 5th person for our sample who had 10 siblings.
New Data: 2,3,1,3,10
Calculate the mean.
Do you think the mean is a good measure of center for this data?
Important characteristics of the sample mean are:

- \( \overline{X} \) is sensitive to extreme scores
- \( \overline{X} \) is not necessarily a possible value

An applied example of the mean not being used when extreme values are present is income. If you hear anyone talk about average income, they should say median income. The median is a much better measure for center in this case than the mean. Consider if I wanted to estimate average income for this class. If a billionaire was in the class, what effect would that have on the mean? It would make it very high and not a good measure for a typical value.

**Sample Median** - the middle score

Procedure for calculating \( \tilde{X} \) (denotes the sample median):

- rank data from smallest to largest
- if \( n \) is odd, median is the middle score
- if \( n \) is even, median is the mean of the two middle scores

**Example 3**
Back to number of siblings Data: 2,3,1,3
Solve for the median.

**Example 4**
New Data: 2,3,1,3,10
Solve for the median.

Important characteristics of the median are:

- \( \tilde{X} \) is not sensitive to extreme scores
- exactly half of the data is below \( \tilde{X} \) and exactly half of the data is above \( \tilde{X} \)

Because of the characteristics of the mean and median, if extreme scores exist in a data set the median is a better measure of central tendency.

If extreme scores are unlikely, the mean varies less from sample to sample than the median and is a better measure of center.

**Sample Mode** - the most frequent score

**Example 5**
Data: 2,3,1,3
New Data: 2,3,1,3,10
Calculate the mode for the above data sets.
There are some major weaknesses with the mode. For example suppose that in the New Data the 10 was changed to a 2. Then what is the mode? You can say it has two modes, both 2 and 3 or you can say the mode does not exist. Even worse, suppose one of the values of 3 was instead a 4. Then you can say the mode does not exist or that all the data points are the mode. Another issue is that if one of the 3 values is changed to a 1 then the mode is 1. Notice changing one value totally changes the mode making it unstable.

The one advantage of using the mode is that it can be used for qualitative data and that is when it should be used. For example if I say the average student in this class is female, what does that tell you? It just means that there are more females than males or the mode is female. The mean or median cannot be used for qualitative data.

Important characteristics of the mode are:
- does not always exist/can be more than one
- unstable
- can be used with qualitative data

2. Measures of Position

There are many measures of position that can be calculated. We will focus on the quartiles. Let me first mention the idea of a percentile which is also a measure of position. A common place that you will see percentiles are when school children take standardized tests. The students will get scores as percentiles. For example, suppose John scores in the 95th percentile in math. Do you know what this means? Technically, it means that John scored better than 95% of people who took the test. Does this say how well he did on the test? Not really, he could have done poorly as long as 95% of his peers did worse. This is a measure of position because it tells where his score is relative to everyone else.

Quartiles – divide the data into four equally sized parts

First Quartile, $Q_1$ : same as the 25th percentile
25% of the data lies below $Q_1$ and 75% of the data lies above $Q_1$

Second Quartile, $Q_2$ : same as the 50th percentile (this is also the median)
50% of the data lies below $Q_2$ and 50% of data lies above $Q_2$

Third Quartile, $Q_3$ : same as the 75th percentile
75% of the data lies below $Q_3$ and 25% of the data lies above $Q_3$
Procedure for calculating $Q_1$, $Q_2$, and $Q_3$:

- Order the data from smallest to largest
- Find the median. This is $Q_2$
- $Q_1$ is the median of the lower half of the data; that is, it is the median of the data falling below $Q_2$ (not including $Q_2$)
- $Q_3$ is the median of the upper half of the data (same as above)

**Five Number Summary** – the low score, $Q_1$, $Q_2$, $Q_3$, and the high score

Example 6
Find the quartiles for the following data sets
Data set #1: 1, 2, 3, 4, 5, 6
Data set #2: 1, 2, 3, 4, 5, 6, 7
Data set #3: 1, 2, 3, 4, 5, 6, 7, 8
Data set #4: 1, 2, 3, 4, 5, 6, 7, 8, 9

Below in the form of a stem-and-leaf plot is data I obtained when I was in graduate school. There was a group of faculty and graduate students in a meeting and they asked everyone how much money they had in their pockets. The data along with the five number summary for each follows:

<table>
<thead>
<tr>
<th>Students</th>
<th>Faculty</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 0 1 3 5 5 6 7 8</td>
<td>0 1 0 5 5</td>
</tr>
<tr>
<td>1 0</td>
<td>2 0 4 5 8 8</td>
</tr>
<tr>
<td>2</td>
<td>3 1</td>
</tr>
<tr>
<td>3</td>
<td>4 3</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>7 3</td>
</tr>
</tbody>
</table>

Low = 0  Low = 10
$Q_1$ = 1  $Q_1$ = 15
$Q_2$ = 5  $Q_2$ = 25
$Q_3$ = 7  $Q_3$ = 31
High = 10  High = 73

Example 7
Make sure you understand how to get the five number summary from the data for students and faculty. Also, are there any extreme values in this data set?
3. **Measures of Dispersion (Variability)**

A measure of dispersion is a measure of the spread of the data points or we can also say a measure of how much variability there is in the data set. I will start with a scenario to illustrate the importance of a measure of dispersion descriptively.

<table>
<thead>
<tr>
<th>Distribution #1</th>
<th>Distribution #2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2 5 5</td>
<td>2 5 5</td>
</tr>
<tr>
<td>3 5 5 5 5 5</td>
<td>3 5 5 5 5</td>
</tr>
<tr>
<td>4 5 5</td>
<td>4 5 5</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

Calculate the mean, median, and mode for distribution #1 and then distribution #2. Compare these two distributions based on the measures of center. What you should conclude is that all measures of center for the two distributions are the same. The solutions for the measures of center follow:

<table>
<thead>
<tr>
<th>Distribution #1</th>
<th>Distribution #2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\bar{X} = 35$</td>
<td>$\bar{X} = 35$</td>
</tr>
<tr>
<td>$\tilde{X} = 35$</td>
<td>$\tilde{X} = 35$</td>
</tr>
<tr>
<td>Mode = 35</td>
<td>Mode = 35</td>
</tr>
</tbody>
</table>

If you compare just the measures of center it appears the two distributions are the same, but are they? They are not the same. Distribution #2 has more variability than distribution #1. This can be reported with a measure of dispersion. That is why the most common way to numerically describe a data set is with a measure of center and a measure of dispersion.

**Sample Range** - represents the distance between the high and low value

Range = High Score - Low Score

**Example 8**

The years of experience for five faculty members will be utilized to illustrate the calculation of the measures of dispersion.

The data are as follows: 1, 30, 22, 10, 5

Calculate the range.

Important characteristics of the range are:
- easy to compute
- totally sensitive to extreme scores
- not a good measure of dispersion in most cases
**Interquartile Range (IQR)** – this is a measure of dispersion which measures the range of middle 50% of the data

\[ IQR = Q_3 - Q_1 \]

Important characteristics of IQR are:
- better than a typical range because it is not based on extremes
- still is not the best measure of center in most cases

**Sample Variance** - measures the average squared distance the data values are from \( \bar{X} \)

\[ S^2 = \frac{\sum (X - \bar{X})^2}{n-1} = \frac{n\left(\sum X^2\right) - \left(\sum X\right)^2}{n(n-1)} \]

where \( S^2 \) is the sample variance
\( \sum \) is the sum
\( X \) is the data values
\( \bar{X} \) is the sample mean
\( n \) is the sample size

Notice there are two ways to calculate \( S^2 \). The first way \( S^2 = \frac{\sum (X - \bar{X})^2}{n-1} \) is the conceptual formula and you can see that this equation comes straight from the definition.

The second equation \( S^2 = \frac{n\left(\sum X^2\right) - \left(\sum X\right)^2}{n(n-1)} \) is the algebraic equivalent of the first and is the one I suggest you use when calculating variance. This equation is much easier to deal with when doing the math by hand.

A common question with variance is why do we calculate the average squared distance and not the average distance directly? The answer to this question comes from the definition of the mean. Think about \( \sum (X - \bar{X}) \). Notice all I did was take the squared off the numerator of the conceptual formula for the variance. If you do this sum it will always be equal to zero. The reason is because some of the \( X \) values will be above the mean and some below the mean because the mean is the center. When you subtract from a \( X \) value bigger than the mean you will get a positive and when you subtract from a \( X \) value smaller than the mean you will get a negative. When these are all added together, since the mean is the center, the positives and negatives will balance out and you will get a sum of zero.
Example 9
Data: 1, 30, 22, 10, 5
Calculate the variance.

Important characteristics of the variance are:
• all the data points are directly involved in the calculation
• variance is in squared units which makes it hard to interpret

Sample Standard Deviation - measures the average distance the data values are from \( \bar{X} \)

When we calculate variance which is the average squared distance the data values are from the mean, we really would like to get rid of the squared because the outcome is not in the same units as the original data. So how do we get rid of a squared? We simply take the square root. Therefore, the square root of the variance gives us standard deviation.

\[
S = \sqrt{S^2}
\]

Example 10
Data: 1, 30, 22, 10, 5
Calculate the standard deviation.

Important characteristics of the standard deviation are:
• standard deviation uses the same units as the data
• standard deviation is the most common measure of dispersion