# Measures of Central Tendency

If we attempt to describe the "center" of a data set, that is a value around which most of the observations (data items) cluster around or center around, or a value that is most typical or representative of the data set, we are referring to a measure of central tendency. Several measures of central tendency will be defined and examples of their computation will follow. In addition, some of the pros and cons of each measure will be demonstrated. It is important to note that since these measures of central tendency will measure "center" in different ways, it is unwise to make a judgment about a data set based solely on any single measure of center.

---

The **mean** of a data set is the sum of the observations divided by the number of observations.
(the mean is what we commonly call the average)
Symbolically, we say the mean $= \frac{\Sigma x}{n} = \frac{\text{sum of the observations}}{\text{number of observations}}$

Example: For the data set 5, -2, 9, 7, the mean $= \frac{5 + (-2) + 9 + 7}{4} = \frac{19}{4} = 4.75$

---

The **median** of a data set is the middle value (after ordering the observations).

Case 1: If the number of observations is odd, then the median is the middle observation.
Case 2: If the number of observations is even, then the median is the mean of the two middle observations.

Example (case 1): For the data set 12, -3, 9, 1, 7, ordering gives -3, 1, 7, 9, 12, so the median is 7 (middle value).

Example (case 2): For the data set 12, -3, 9, 1, 7, 15, ordering gives -3, 1, 7, 9, 12, 15.
Observe that the middle is between 7 and 9, so the median is $\frac{7 + 9}{2} = \frac{16}{2} = 8$.
We can easily locate the median's position by sight (particularly with a small data set), but we may use $\frac{n + 1}{2}$ to compute the median's position.
In case 1, with 5 observations, the median position is $(5 + 1)/2 = 6/2 = 3$.
So the median is the 3rd ordered observation which is 7.
In case 2, with 6 observations, the median position is $(6 + 1)/2 = 7/2 = 3.5$.
So the median is the mean (average) of the 3rd and 4th ordered observations which is $(7 + 9)/2 = 16/2 = 8$

---

The **mode** of a data set is the most frequently occurring observation.
(If all observations occur with the same frequency, then the there is no mode)
(If more than one observation has the greatest frequency, then each of these observations is a mode)

| Examples: | Data set | Mode |
|---|---|---|
| | 2, 3, 4, 5, 5, 6 | 5 |
| | 7, 8, 8, 8, 9, 9, 10, 12, | 8 |
| | 1, 2, 3, 4, 5, 6 | no mode |
| | 7, 8, 8, 9, 9, 13, 15 | 8,9 |

---

The **midrange** of a data set is the mean of the lowest and highest observations.

Symbolically, we say midrange $= \frac{L + H}{2}$

Example: For the data set 7, -5, 0, -2, 4, 11 the midrange $= \frac{L + H}{2} = \frac{-5 + 11}{2} = \frac{6}{2} = 3$

Now we will illustrate some of the pros and cons of each measure of center.
First, let's examine the mean and median.
An appealing aspect of the mean (average) is that each observation is directly involved in its calculation.
However, it can be influenced by an outlier (an observation far from the others) as we will see.
Suppose that the annual salaries (in thousands of dollars) for five employees of a small company are shown below for two different years (with mean and median salary calculated for both years).

| | MEAN | MEDIAN |
|---|---|---|
| Last Year: 21, 22, 23, 24, 25 | 23 | 23 |
| This Year: 20, 21, 22, 23, 54 | 28 | 22 |

Note that for last year the mean and median are both 23 thousand dollars. While the mean and median measure center differently, they can sometimes (as in this case) turn out to be the same value (here because of symmetry). However, the mean salary for this year is substantially greater than the median for this year.

Examining this year's salaries illustrates a clear weakness of the mean - it can be influenced by an outlier (54). The company's workers may rightfully argue that 28 thousand dollars is not representative or typical of this year's salaries since almost all workers make substantially less than 28 thousand dollars. The inflation of the mean to 28 is caused by the outlier of 54 being directly involved in the calculation.
(Also note that the mean is pulled in the direction of the outlier)
This also illustrates a strength of the median – it is not influenced by an outlier. No matter how large (or small) the outlier, the median (middle number after ordering) is still 22 thousand dollars, a figure that is more representative or typical of the workers' salaries than is 28 thousand. In cases where the mean (average) is pulled far from center by an outlier, the median is a more appropriate measure of center.

---

Still, the median is not without its weakness as will now be illustrated. While the median uses all observations in the ordering process, it then focuses only on the middle observation, and thus totally ignores all other observations.
Suppose that you have a chance to go to work for two different companies whose salaries (in thousands of dollars) are shown below. However, also suppose that you have no idea where you will end up in the company's salary structure. Would you rather go to work for company A or company B?
Company A: 10, 10, 21, 21, 21
Company B: 20, 20, 20, 98, 99
Would you rather work for company A since its median (21) is higher than company B's median (20)?
Most likely not! When asked this question, people instinctively choose company B because the lowest salary for company B is virtually as much as the highest salary for company A, and you have a 40% chance (2 out of 5) of making more than everyone in company A combined! But the median (ignoring all observations except the middle one) would never reveal this information. The lesson here is that we should not rely on only one measure of center to make a judgment about a data set.
A trimmed mean is calculated after removing the most extreme observation on each end of an ordered data set, thus directly involving almost all observations in its calculation but excluding the most extreme outliers.
Olympic events such as gymnastics or diving that depend on scoring from judges use a trimmed mean where a competitor's score is the mean (average) after the highest and lowest scores are trimmed (dropped).

---

Both the mode (most common observation) and midrange (average of the lowest and highest observations) are appealing in that they are easily calculated, but each has a weakness that is clearly demonstrated below.

Suppose that a student's test scores are as follows: 68, 68, 93, 97, 99
The mode is 68, but hardly typical of the student's performance.
A weakness of the mode is that the most common observation may be far from the typical observation.

Suppose that a student's test scores are as follows: 61, 90, 92, 92, 93
The midrange is $(61 + 93)/2 = 154/2 = 77$, but should such a student get a C because of only one bad test?
A weakness of the midrange is that it uses only the most extreme observations in its calculation.

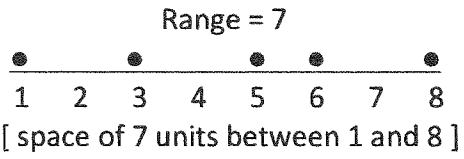
# MEASURES OF DISPERSION

While on a daily basis we tend to work more with measures of central tendency (we often speak or hear about median home price, median income, or we compute the average of our test scores), measures of dispersion (also called spread or variability) are also of great practical importance. Dispersion is sometimes desirable. For example, if a standardized test is administered and all students make the same score, it may be argued that the test was poorly constructed since it was unable to distinguish between students of different ability/knowledge levels. However, dispersion is undesirable in many cases. For example, if a restaurant advertises its hamburger patties to weigh an average of 4 ounces, customers may be driven away by an inconsistent product ( they may be pleased if their burger weighs 5 ounces one day, but may not ever come back if it weighs only 3 ounces the next day). In addition, a great amount of dispersion, spread, or variability makes the job of estimation and prediction much more difficult.
For instance, if Sue has test scores of 80, 81, and 79 while Jan has scores of 80, 100, and 60, both women have the same average score of 80. However, our common sense tells us that it will be much more difficult to predict the next performance for Jan because her scores are much more dispersed (or spread out or varied) than are Sue's.
Therefore, it is essential that we be able to compute measures describing the dispersion of a data set.

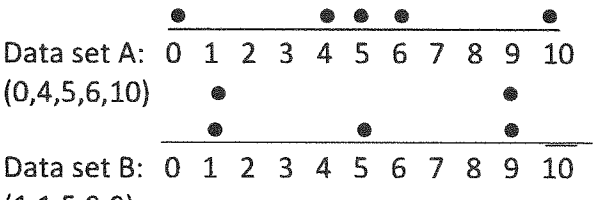
---

**RANGE = largest observation – smallest observation** or more simply: **RANGE = HIGH - LOW**

Example: For the data set 5, 1, 3, 8, 6,
Range = 8 – 1 = 7
Note that the range is simply the distance between the smallest and largest observation in the data set.
(See the dot plot at the right)




Advantages of the range as a measure of dispersion are that it is quick and easy to calculate, and it also has a simple and easy to see interpretation (the distance between the smallest and largest observations in a data set).
In this case, 8 – 1 = 7.
However, its main disadvantage (that it only involves the most extreme observations in its calculation) will be made clear by the example below. Suppose we have two data sets, A and B, displayed by the dot plots below.




Quick inspection of the data sets shows that A has a larger range (10-0=10) than B (9-1=8). However, by observation one can reasonably state that B has greater overall variability than A. The range uses only the extremes (high and low) in its calculation. Looking at the remaining observations (4,5,6 for A) and (1,5,9 for B), B is much more variable or dispersed than A.

If we develop a measure of dispersion that directly involves each observation in its calculation, and that measures dispersion in terms of how far each observation is from the data set's center (mean), [ Data sets A and B both have the same mean (average) of 5 ], then this measure should indicate that data set B is more dispersed than data set A.
**The measure that we develop will be the variance, and its square root will be the standard deviation.**

---

## VARIANCE

To calculate the variance, for each observation we will ...

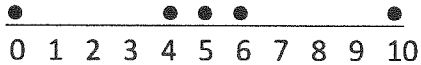
1) Determine its deviation from the mean: (observation – mean)
2) Square the deviation from the mean: $(\text{observation} - \text{mean})^2$
3) Sum the squared deviations from the mean: $\sum(\text{observation} - \text{mean})^2$
4) Average the squared deviations from the mean

Formulas for the variance are show below: ( N = population size, n = sample size)

Population variance = $\sigma^2 = \dfrac{\sum (\text{observation} - \text{mean})^2}{N}$ **Sample variance = $s^2 = \dfrac{\sum (\text{observation} - \text{mean})^2}{n-1}$**

You may naturally be wondering if we are averaging the squared deviations, why do we divide by n-1 instead of n when calculating the sample variance? This is done to make $s^2$ (sample variance) an unbiased estimator of $\sigma^2$ (population variance) meaning that while $s^2$ can vary from sample to sample (underestimating or overestimating $\sigma^2$), it will have an expected value (or average) equal to $\sigma^2$.

**The formula for sample variance was placed in bold print to emphasize that in statistics, we deal almost exclusively with samples because it is almost always too time consuming, costly, or impractical to access an entire population.**
Now let's consider data set A (from the previous page) to be a sample, and calculate its sample variance.




| observation | observation – mean | (observation – mean)$^2$ |
|---|---|---|
| 0 | 0 – 5 = -5 | (-5)(-5) = 25 |
| 4 | 4 – 5 = -1 | (-1)(-1) = 1 |
| 5 | 5 – 5 = 0 | (0)(0) = 0 |
| 6 | 6 – 5 = 1 | (1)(1) = 1 |
| 10 | 10 – 5 = 5 | (5)(5) = 25 |
| | 0 | 52 |

$$S^2 = \frac{\sum (\text{observation} - \text{mean})^2}{n-1} = \frac{52}{5-1} = \frac{52}{4} = 13$$

So the sample variance is 13. Note that the sample variance is a more sophisticated measure of dispersion than the range, thus unlike the range, it does not have a simple interpretation that can be easily seen on a number line. It is best to simply think of the variance (and its related measure the standard deviation) as descriptions of a data set's dispersion or variability from its center in the overall sense.

Note: Data set B (from the previous page), has a sample variance of 16 (which is greater than 13) thus verifying our suspicion that it has greater overall variability in terms of dispersion from its center (mean) than does A.

## STANDARD DEVIATION

**Standard deviation = square root of the variance**

Population standard deviation = square root of population variance = $\sigma$

**Sample standard deviation = square root of sample variance = s**

Note that the symbols for standard deviation are simply the symbols for variance without the exponent 2 since the square root of $x^2$ is simply x (where $x \geq 0$).
Once again, since in statistics we deal almost exclusively with samples, sample standard deviation is in bold.

Since the standard deviation is the square root of the variance, it is in the same units as the data.
For example, if the data is in feet, standard deviation is also in feet, but the variance would be in square feet.

To calculate the sample standard deviation for data set A ( the set shown above consisting of 0,4,5,6,10), we simply find the square root of its variance.

**Thus for data set A, $S = \sqrt{13} \approx 3.61$**

**Comments on standard deviation (which also apply to variance)**

1) Standard deviation $\geq 0$
2) Standard deviation = 0 only when all observations in the data set are identical.
3) Adding or subtracting the same value to each observation in a data set does not change the standard deviation

When using a TI-84/83 calculator to calculate standard deviation [directions for this can be found on page 84 of Elementary Statistics (5th edition) by Larson/Farber], be careful because the calculator output displays both $\sigma$ and s. So be certain as to whether you are dealing with a population or a sample.
Also, if variance is needed, just square the result displayed for standard deviation.