4 – Central Tendency & Spread

Two characteristics of quantitative data – central tendency and spread – can each be represented by single numerical values. The value used for each, however, must be determined by the specific data set.

Measures of central tendency are numerical values that represent a quantitative dataset by answering the question, "Where is the data centered?" There are four such measures you should know, and each one's most appropriate use:

(1) The mean is typically referred to as the average. This is the most commonly used measure of central tendency.

(2) The median is the middle value that splits a ranked dataset into two equal portions.

(3) The mode is the most frequently occurring value in a dataset.

(4) The midrange is the value midway between the highest and lowest values.

Measures of spread (aka, variation) are numerical values that represent a quantitative dataset by answering the question, "How spread out is the data?" There are three of spread measures you should know:

(1) Variance is a statistic that is computed using each data point's deviation from the mean. (for example, if the mean of a dataset is 12.5, then a data point of 7.5 has a deviation of 5).

(2) Standard Deviation is computed directly from the variance. Most statistical computations and techniques are based upon the standard deviation.

(3) Range is the difference between the maximum and minimum values of a dataset.

The symbols that represent some of the measures of central tendency and spread are different for sample statistics and population statistics. Those symbols are:

	Mean	Variance	Standard Deviation.
Sample	$\bar{X}$ (pronounced "x-bar")	s^2	s
Population	μ (lower case Greek "mu")	σ^2 (lower case Greek "sigma")	σ

Here is how each of these statistics can be found:

Mean: The mean is computed by adding up all of the values and then dividing by the number of data points, or $\bar{X} = \frac{sum\ of\ all\ values}{number\ of\ values}$

Median: If *N* is odd, then the median is the value in the $\frac{N+1}{2}$ position. If *N* is even, then the median is the number halfway between the value in the $\frac{N}{2}$ position and the value in the $\frac{N}{2} + 1$ position.

Mode: Finding the mode is a matter of inspecting the dataset and counting the frequency of each value's occurrence. There may be more than one mode or none at all if no value is repeated.

Midrange: This statistic is computed by adding the highest and lowest values, and then dividing by two.

Variance: The procedure to compute sample variance is reflected in the formula $s^2 = \frac{sum\ of\ (deviation\ of\ each\ value\ from\ mean)^2}{N-1}$. In short, after the mean is computed, the amount that each value deviates from the mean is squared, and then they are all added up. The result is divided by *N* – 1. For the population mean, we use the same procedure, but we divide instead by *N*. $\sigma^2 = \frac{sum\ of\ (deviation\ of\ each\ value\ from\ mean)^2}{N}$. You will not be asked to compute variance on any quiz or exam.

Standard Deviation: The standard deviation is simple, once the variance has been computed. Just take the square root of the variance. You will not be asked to compute standard deviation on any quiz or exam.

Range: This statistic is computed by subtracting the smallest value in a dataset from the largest. It is rarely used.

Example 1: In order to understand the different measures of central tendency and spread, we compute them for two different datasets and compare. Here are two datasets which contain responses to two past Math 175 student surveys that asked, "How long would it take you to drive home from here?"

The values of the measures of central tendency and spread appear below and can be used to make generalizations about the data sets, as well as how they compare.

Fall 2012

10	20	35	90
10	20	40	110
10	20	45	150
10	25	45	240
15	25	45	240
15	30	45	270
15	30	60	1500
20	30	60	174000

Spring 2012

4	10	15	30	75
5	10	15	30	90
7	10	20	30	90
10	10	20	35	135
10	15	20	40	260
10	15	25	40	270
10	15	25	45	290
10	15	25	45	330
10	15	30	45	330

	Fall 2012	**Spring 2012**
Mean ($\bar{X}$)	5540.00	57.69
Median	32.5	20
Mode(s)	10, 20, 45	10
Midrange	87005	167
Variance (σ^2)	945044333.87	8012.04
Standard Deviation (σ)	30741.57	89.51
Range	173990	326

Clearly, the mean or median would be the best measures of central tendency for the spring 2012 data set and median is best for the fall 2012 measure of central tendency.

When data is in a frequency distribution, direct computations of the various measures of central tendency and spread are impossible because the raw data is unknown. In these cases, the technique for making these computations is to create a virtual data set that follows the frequency distribution.

To illustrate this, consider this frequency distribution below from a psychology experiment which recorded the amount of time it took 20 children to color a set of drawings.

The virtual dataset that we use to compute central tendency and spread is composed of the midpoints of each class, each represented by their class's frequency. For the first class, the midpoint is 27. So, we write 27 four times in our new dataset, and so on. The process is shown here:

Minutes	Frequency
25-29	4
30-34	4
35-39	7
40-44	5

27	32	37	42
27	32	37	42
27	32	37	42
27	37	37	42
32	37	37	42

Mean	35.25	Variance	29.67
Median	37	Standard Deviation	5.45
Mode	37	Range	15
Midrange	34.5		

In some of the upcoming material, we will be referring to important intervals some number of *standard deviations from the mean*. This interval is created from the two statistics, mean and standard deviation, and represents the addition and subtraction of the standard deviation to and from the mean, some specific number of times.

In example 1 above, the mean and standard deviation of the spring 2012 data were computed to be 57.69 and 89.51 respectively. So, the interval that represents *within two standard deviations of the mean* can be found by subtracting 89.51 from 57.69 twice and separately adding 89.51 twice to 57.69.

That interval goes from 57.69 – 2×89.51 = -121.33 to 57.69 + 2×89.51 = 236.71, which is expressed (-121.33, 236.71). In this case, since 40 of the 45 data points (90%) are in this interval, we can say that 90% of the data are within two standard deviations of the mean.

There is a famous theorem that uses this idea:

> **Chebychev's Theorem**: For any data set, at least $1 - \frac{1}{k^2}$ of the data points will lie within *k* standard deviations of the mean, where *k* is any real number (not necessary an integer) larger than 1.
>
> In the example above, k = 2 and so $1 - \frac{1}{k^2} = 1 - \frac{1}{2^2} = \frac{3}{4} = 75\%$.
>
> Above, we computed the portion of the spring 2012 data that lies within two standard deviations of the mean to be 90%, and because of the *at least* language in the theorem, it is confirmed that the theorem works in this case.