**Lecture 1**

**Mean, Median, and Mode**

There are four terms you will need to fully understand.Mean, median, and mode are three kinds of "averages".

The ***sample mean*** is the average and is computed as the sum of all the observed outcomes from the sample divided by the total number of events.

        
where n is the sample size and the x correspond to the observed value.

 **Example**

Suppose you randomly sampled six numbers:

        34, 43, 81, 106, 106 and 115

We compute the sample mean by adding and dividing by the number of samples, 6.

             34 + 43 + 81 + 106 + 106 + 115
              = 80.83
                                   6

The ***mode*** of a set of data is the number with the highest frequency.  In the above example 106 is the mode.

 The ***median*** is the middle score.  If we have an even number of events we take the average of the two middles.

Suppose you randomly selected 10 house prices.

        2.7,   2.9,   3.1,   3.4,   3.7,  4.1, 4.3,   4.7, 4.7, 4.8

      3.7 + 4.1
       = 3.9 which is the median
            2

****

**Standard deviation and other measures of dispersion**

**Standard Deviation,Variance, Coefficient of Variation and standard error**

The mean, mode, and median do a nice job in telling where the center of the data set is, but often we are interested in more.  How far the data is spread apart?  This is what the variance and standard deviation do.

**Standard deviation equation for an entire population:**

Deviation just means how far from the normal.

The Standard Deviation (**root mean square**)is a measure of how spread out numbers is.

Its symbol is **σ** (the Greek letter sigma)



Σ is the summation (or total)

(sigma)







**Standard deviation** equation for **a sample of a population** (The best estimate of the true standard deviation)**,**it can be shown that this underestimate is corrected by using **N-1** instead of **N**(which is called "Bessel's correction)**:**

The symbols also change to reflect that we are working on a sample instead of the whole population:











**N -1**is the number of values in the sample minus 1

**Note**:

**σ** is calledthe**population standard deviation** for **N**

**S** is called the **sample standard deviation** for **N-1**

This quantity is the **population standard deviation**. The formula is valid *only* if the eight values we began with form the *complete* population. If they instead were a random sample, drawn from some larger, "parent" population, then we should have used 7 (which is *N* − 1) instead of 8 (which is *N*) and then the quantity thus obtained would have been called the **sample standard deviation**.

Here are the amounts of 5 numbers**Drawn from larger numbers**:

4, 2, 5, 8, 6.

1. Calculate the mean:









2. Calculate for each value in the sample:











3. Calculate :







4. Calculate the standard deviation:



=2.24

## Variance

The **Variance** is defined as:

The average of the **squared** differences from the Mean Var.=**σ2**

And the **standard deviation** is the **square root** of the **Variance**

**Variance and Standard Deviation: Step by Step**

1. Calculate the mean.
2. Write a table that subtracts the mean from each observed value.
3. Square each of the differences.
4. Add this column.
5. Divide by N -1 where N is the number of items in the sample. This is the *variance*.
6. To get the *standard deviation* we take the square root of the variance.

  **Example**

Given10 randomly **selected** data.

        44,   50,   38,   96,   42,   47,   40,   39,   46,   50

       Mean  =  49.2

  Now  = 2600.4

        2600.4
         =  288.7        10 - 1

Hence the variance is 289 and the standard deviation is the square root of  289 = 17.

The sample variance will be denoted by s2 and the population variance will be denoted by 2. The variance and standard deviation describe how spread out the data is.  If the data all lies close to the mean, then the standard deviation will be small, while if the data is spread out over a large range of values, s will be large.

The *coefficient of variation* which is the standard deviation divided by the mean times 100%. Note the mean can also defined by the symbol 

                      
        CV  =    100%
                      

In the above example, it is

         17
        100%   =  34.6%
        49.2

This tells us that the standard deviation is 34.6% of the mean.

## Why Would We Take a Sample?

Mostly because it is **easier** and **cheaper**.

Imagine you want to know what the whole country thinks ... you can't ask millions of people, so instead you ask maybe 1,000 people.

But when we take a sample, we **lose** some accuracy.

**Standard error**

The variability of the sample means is called the **standard error of the mean** or the **standard deviation of the mean**.

Standard Error of the Mean (SEM) **= S/**$\sqrt{N}$

**Example**X = 10, 20, 30, 40, 50
Mean (xm) = (x1 + x2 + x3 + .... + xn)/N = 30
SD = √(1/(N-1) \* ((x1 - xm)2 + (x2 - xm)2 + ... + (xn - xm)2))
= √(1/(5-1)((10-30)2+(20-30)2+(30-30)2+(40-30)2+(50-30)2)) = 15.811

To Find Standard Error:
Standard Error = SD / √(N) = 7.0711

 **Why squarethe differences?**

If we just added up the differences from the mean ... the negatives would cancel the positives:

|  |  |  |
| --- | --- | --- |
| C:\Documents and Settings\MidyaTech\Desktop\Standard Deviation and Variance_files\standard-deviation-why-a.gif |   |  |

(4 + 4 - 4 – 4) / 4 =0

So that won't work. How about we use [**absolute values**](http://www.mathsisfun.com/numbers/absolute-value.html)?

(|4| + |4| + |-4| + |-4|) / 4 = (4 + 4 + 4 + 4) / 4 = 4

That looks good, but what about this case:

|  |  |  |
| --- | --- | --- |
| C:\Documents and Settings\MidyaTech\Desktop\Standard Deviation and Variance_files\standard-deviation-why-b.gif |   |  |

(|7| + |1| + |-6| + |-2|) / 4 = (7+1+6+2)/4 = 4

It also gives a value of 4, even though the differences are more spread out!

So let us try squaring each difference (and taking the square root at the end):

√ (42+42+42+42) / 4 =√64 / 4 = 4

√ (72+12+62+22) / 4 = √90 / 4 = 4.74

That is **nice!** The Standard Deviation is **bigger** when the **differences** are **more spread** out ... just what we want!