Basic Statistics

Mean, Mode, Median, and Standard Deviation

The Mean and Mode

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. We use x as the symbol for the sample mean. In math terms,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x$$

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

The *mode* of a set of data is the number with the highest frequency, one that occurs maximum number of times.

Median, and Trimmed Mean

One problem with using the mean, is that it often does not depict the typical outcome. If there is one outcome that is very far from the rest of the data, then the mean will be strongly affected by this outcome. Such an outcome is called an *outlier*. An alternative measure is the median. The *median* is the middle score. If we have an even number of events we take the average of the two middles. The median is better for describing the typical value. It is often used for income and home prices.

Example

Suppose you randomly selected 10 house prices. You are interested in the typical house price. In lakhs the prices are

2.7, 2.9, 3.1, 3.4, 3.7, 4.1, 4.3, 4.7, 4.7, 40.8

If we computed the mean, we would say that the average house price is 744,000. Although this number is true, it does not reflect the price for available housing in South Lake Tahoe. A closer look at the data shows that the house valued at 40.8 x 100,000 = 40.8 million skews the data. Instead, we use the median. Since there is an even number of outcomes, we take the average of the middle two (3.7 + 4.1)/2 = 3.9. Therefore, the median house price is 390,000. This better reflects what ahouse shopper should have to buy a house.

There is an alternative value that also is resistant to outliers. This is called the *trimmed mean* which is the mean after getting rid of the outliers or 5% on the top and 5% on the bottom. We can also use the trimmed mean if we are concerned with outliers skewing the data, however the median is used more often since more people understand it.

Example:

At a ski rental shop data was collected on the number of rentals on each of ten consecutive Saturdays:

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

To find the sample mean, add them and divide by 10:

$$\frac{44 + 50 + 38 + 96 + 42 + 47 + 40 + 39 + 46 + 50}{10} = 49.2$$

Notice that the mean value is not a value of the sample. To find the median, first sort the data:

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

Notice that there are two middle numbers 44 and 46. To find the median we take the average of the two.

$$\text{Median} = \frac{44 + 46}{2} = 45$$

Notice also that the mean is larger than all but three of the data points. The mean is influenced by outliers while the median is robust.

Variance and Standard deviation

The mean, mode, median, and trimmed mean do a nice job in telling where the center of the data set is, but often we are interested in more. For example, a pharmaceutical engineer develops a new drug that regulates iron in the blood. Suppose she finds out that the average sugar content after taking the medication is the optimal level. This does not mean that the drug is effective. There is a possibility that half of the patients have dangerously low sugar content while the other half has dangerously high content. Instead of the drug being an effective regulator, it is a deadly poison. What the pharmacist needs is a measure of how far the data is spread apart. This is what the variance and standard deviation do. First we show the formulas for these measurements. Then we will go through the steps on how to use the formulas.

We define the *variance* to be

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x-\bar{x})^2$$

and the *standard deviation* to be

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x-\bar{x})^2}$$

Variance and Standard Deviation: Step by Step

1. Calculate the mean, x.
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

The owner of a restaurant is interested in how much people spend at the restaurant. He examines 10 randomly selected receipts for parties of four and writes down the following data.

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

He calculated the mean by adding and dividing by 10 to get

x = 49.2

Below is the table for getting the standard deviation:

x	x - 49.2	$(x - 49.2)^2$
44	-5.2	27.04
50	0.8	0.64
38	11.2	125.44
96	46.8	2190.24

42	-7.2	51.84
47	-2.2	4.84
40	-9.2	84.64
39	-10.2	104.04
46	-3.2	10.24
50	0.8	0.64
Total		**2600.4**

Now

$$\frac{2600.4}{10 - 1} = 288.7$$

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

Since the standard deviation can be thought of measuring how far the data values lie from the mean, we take the mean and move one standard deviation in either direction. The mean for this example was about 49.2 and the standard deviation was 17. We have:

49.2 - 17 = 32.2

and

49.2 + 17 = 66.2

What this means is that most of the patrons probably spend between 32.20 and 66.20.