MEASURES OF CENTRAL TENDENCY AND DISPERSION

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.

Example

Suppose you randomly sampled six acres in the Desolation Wilderness for a non-indigenous weed and came up with the following counts of this weed in this region:

34, 43, 81, 106, 106 and 115

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

$$\frac{34 + 43 + 81 + 106 + 106 + 115}{6} = 80.83$$

We can say that the sample mean of non-indigenous weed is 80.83.

The *mode* of a set of data is the number with the highest frequency. In the above example 106 is the mode, since it occurs twice and the rest of the outcomes occur only once.

The *population mean* is the average of the entire population and is usually impossible to compute. We use the Greek letter μ for the population mean.

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 and *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 in the South Lake Tahoe area. Your are interested in the typical house price. In $100,000 the prices were

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 = $4.08 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

$$\frac{3.7 + 4.1}{2} = 3.9$$

The median house price is $390,000. This better reflects what house shoppers should expect to spend.

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 have 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 the Ches Tahoe 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.

The sample standard deviation will be denoted by s and the population standard deviation will be denoted by the Greek letter σ.

The sample variance will be denoted by s^2 and the population variance will be denoted by σ^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. Having outliers will increase the standard deviation.

Range

The range is the difference between the highest and lowest scores in a data set and is the simplest measure of spread. So we calculate range as:

Range = maximum value - minimum value

For example, let us consider the following data set:

23 56 45 65 59 55 62 54 85 25

The maximum value is 85 and the minimum value is 23. This results in a range of 62, which is 85 minus 23. Whilst using the range as a measure of spread is limited, it does set the boundaries of the scores. This can be useful if you are measuring a variable that has either a critical low or high threshold (or both) that should not be crossed. The range will instantly inform you whether at least one value broke these critical thresholds. In addition, the range can be used to detect any errors when entering data. For example, if you have recorded the age of school children in your study and your range is 7 to 123 years old you know you have made a mistake!

Quartiles and Interquartile Range

Quartiles tell us about the spread of a data set by breaking the data set into quarters, just like the median breaks it in half. For example, consider the marks of the 100 students below, which have been ordered from the lowest to the highest scores, and the quartiles highlighted in red.

Order	Score	Order	Score	Order	Score	Order	Score	Order	Score
1st	35	21st	42	41st	53	61st	64	81st	74
2nd	37	22nd	42	42nd	53	62nd	64	82nd	74
3rd	37	23rd	44	43rd	54	63rd	65	83rd	74
4th	38	24th	44	44th	55	64th	66	84th	75
5th	39	25th	45	45th	55	65th	67	85th	75
6th	39	26th	45	46th	56	66th	67	86th	76
7th	39	27th	45	47th	57	67th	67	87th	77
8th	39	28th	45	48th	57	68th	67	88th	77
9th	39	29th	47	49th	58	69th	68	89th	79
10th	40	30th	48	50th	58	70th	69	90th	80
11th	40	31st	49	51st	59	71st	69	91st	81
12th	40	32nd	49	52nd	60	72nd	69	92nd	81
13th	40	33rd	49	53rd	61	73rd	70	93rd	81
14th	40	34th	49	54th	62	74th	70	94th	81
15th	40	35th	51	55th	62	75th	71	95th	81
16th	41	36th	51	56th	62	76th	71	96th	81
17th	41	37th	51	57th	63	77th	71	97th	83
18th	42	38th	51	58th	63	78th	72	98th	84
19th	42	39th	52	59th	64	79th	74	99th	84
20th	42	40th	52	60th	64	80th	74	100th	85

The **first quartile** (Q1) lies between the 25th and 26th student's marks, the **second quartile** (Q2) between the 50th and 51st student's marks, and the **third quartile** (Q3) between the 75th and 76th student's marks. Hence:

First quartile (Q1) = (45 + 45) ÷ 2 = **45**
Second quartile (Q2) = (58 + 59) ÷ 2 = **58.5**
Third quartile (Q3) = (71 + 71) ÷ 2 = **71**

In the above example, we have an even number of scores (100 students, rather than an odd number, such as 99 students). This means that when we calculate the quartiles, we take the sum of the two scores around each quartile and then half them (hence Q1= (45 + 45) ÷ 2 = 45) . However, if we had an odd number of scores (say, 99 students), we would only need to take one score for each quartile (that is, the 25th, 50th and 75th scores). You should recognize that the second quartile is also the median.

Quartiles are a useful measure of spread because they are much less affected by outliers or a skewed data set than the equivalent measures of mean and standard deviation. For this reason, quartiles are often reported along with the median as the best choice of measure of spread and central tendency, respectively, when dealing with skewed and/or data with outliers. A common way of expressing quartiles is as an interquartile range. The interquartile range describes the difference between the third quartile (Q3) and the first quartile (Q1), telling us about the range of the middle half of the scores in the distribution. Hence, for our 100 students:

Interquartile range = Q3 - Q1
= 71 - 45
= 26

However, it should be noted that in journals and other publications you will usually see the interquartile range reported as 45 to 71, rather than the calculated range.