LESSON HANDOUT

STD Deviation and Skewness

Why

Standard (STD) deviation is a better way of achieving both the series aims of:

- Determining an individual values significance in the context of the whole, and;
- Classifying our data into areas of significance.

The reason that STD deviation is a better way to achieve this is because STD Deviation factors in the *underlying values* in the dataset rather than just relying on rank (as quartiles, percentiles and IQR do).

Skew provides a measure of the amount and direction of skew in our dataset allowing us to factor this into our analysis.

Standard Deviation

Std deviation is the *average distance to the mean* and is a measure of **spread**.
It is calculated by measuring each individual's values distance from the mean, and then averaging those distances. For a great explanation on how it is calculated watch this video from AP Stats guy. You can consider STD deviation as working in a team with the mean (because it's calculated based on the mean). If the mean is skewed, then so will the standard deviation.

STD Deviation in practice - Example

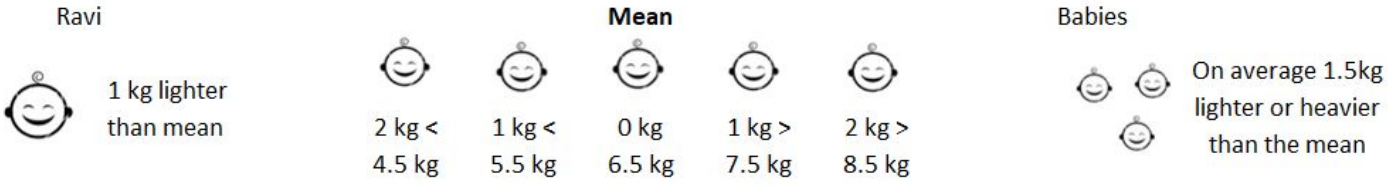
In this scenario, a doctor measuring the weight of an individual baby - Ravi - to check whether his development is on track. The doctor weighs Ravi, who is 3 months old, and finds that he weighs 5.5 kg.

How does the doctor determine if this baby is on track? The doctor has four options:

- **Option 1.** Make a qualitative assessment based upon their experience of weighing other babies
    - Problem - small sample size, subjective
- **Option 2.** Compare Ravi's weight to the average or median weight of 3 month-old babies in the country, in this case, the average is 6.5kg
    - Problem - Ravi weighs 1 kg less than the average but so what, how significant is this?

- **Option 3.** Rank Ravi's weight amongst all 3-month-old babies, Ravi is in the 38th percentile meaning he is heavier than 38% of all babies in the country.
    - Problem - Good start, it's still difficult to say whether this is significant, are all babies in the 38th percentile behind in their development?

- **Option 4.** Compare Ravi's distance from the average (1 kg) to the average distance to the mean.



    - Solution - We now know that, on average, babies are 1.5kg heavier or lighter than the mean baby weight meaning that Ravi's weight is not unusual.

Skewness

Skewness provides a measure of the amount and direction of skew in our dataset, allowing us to factor this into our analysis. Skew can be used to determine whether to use STD Deviation, or quartiles to segment our data. If our data is moderately or strongly skewed, then we should consider ***against*** using STD Deviation because it will not segment our data appropriately.

Calculating skew in excel is simple, use the formula: =SKEW (values to measure)

You can interpret skew using the following table:

SKEWNESS	Measures the amount and direction of skew				
	Highly left skewed	Moderately left skewed	Approx symetrical	Moderately right skewed	Highly right skewed
	- 1 or less	- 0.5 to - 1	- 0.5 to + 0.5	0.5 to 1	+ 1 or more

Using STD deviation to segment data

Provided the data is not skewed, we can classify our data into areas of significance using STD Deviation:

- Those values below or above 3 STD deviations from the mean are classified as very low and very high values respectively.
- Those values below or above 1 STD deviation from the mean are classified as low and high values respectively.
- Those values within 1 STD deviation from the mean are classified as medium.

- We can assign these values to our data using the approximate match in VLOOKUP (assigning approximate rather than exact match at the end of the formula).
  - Step 1. Create a classification table. Approximate match works in this case, by looking up the value in the second column when it is equal to the value in the first column.

Classification using approximate match

Value	Classification
Min value	Very low
Mean minus 3 STD deviations	Low
Mean minus 1 STD deviation	Medium
Mean plus 1 STD deviation	High
Mean plus 3 STD deviations	Very high

- Step 2. Create the VLOOKUP formula using an approximate match.

Using STD Deviation check an individual value's significance

The number of STD deviations an individual value is from the mean is a very useful indicator of that values's significance. This measure (number of STD deviations from the mean) is also called the **Z score**.

To calculate percentile rank in excel use the 'Percent Rank' function.

- = PERECENTRANK (range of cells, individual cell)