Data Science

Descriptive Statistics

Measure of Central Tendency

- Mean or Average
- Median
- Mode

Mean

- Most popular and commonly used measure
- Mean tells the mathematical average of all data
- It takes into account each and every value in dataset
- Sum of the deviations of each value from the mean is always zero.
- Can only be used with numeric data (both discrete and continuous).
- It is sensitive to outliers

Median

- It represents exact middle point of a sorted data set
- It can be used when data is having outliers
- It is also used with numerical data only.

Mode

- Mode is the most frequently occurring value or category in the data set.
- Normally, the mode is used for categorical data where we wish to know which is the most common category
- A data can have one or more than one mode.
- Mode can be used both numerical and categorical data

When to use which measure ?

Mode

- Mean -
  - Data is numerical, symmetric and have no outliers
- Median
  - When data is numerical and skewed or having outliers
- Mode
  - Used with categorical data, can be used with numerical data also

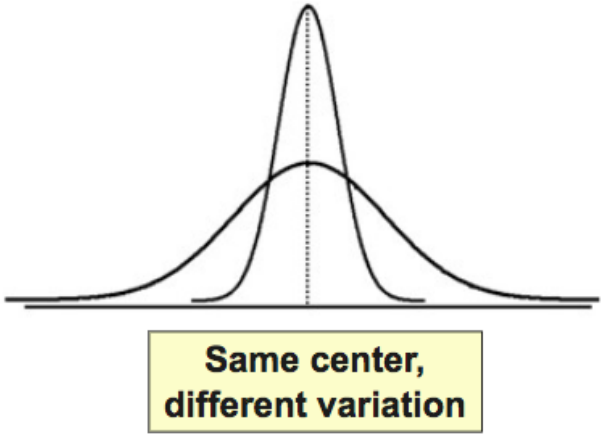
Measure of Spread/Variance

Measure of Spread

- It represents the amount of dispersion in a dataset i.e. how spread out are the values
- How far away the data points tend to fall from the center.
- This type of measure only applies to ordinal and numeric data that can be ranked

Measure of Spread

- A measure of spread gives us an idea of how well the mean represents the data.
- When a distribution has lower variability, the values in a dataset are more consistent.
- When the variability is higher, the data points are more dissimilar and extreme values become more likely.



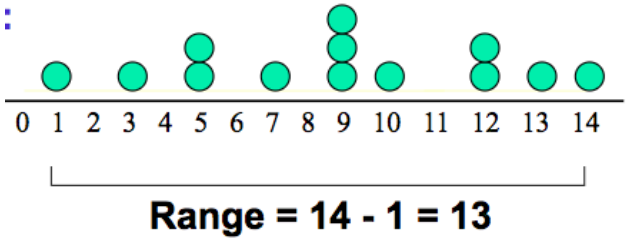
Measure of Spread

- Range
- IQR (Interquartile range)
- Variance
- Standard deviation

Range

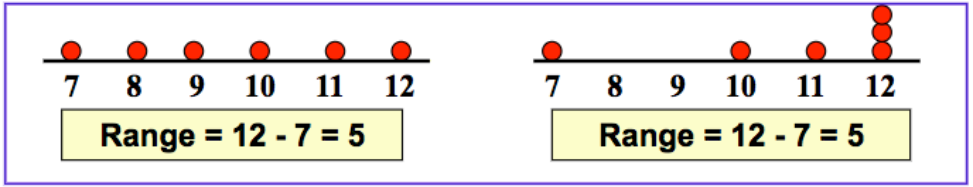
Range

- Simplest measure of variation
- Difference between the largest and smallest value



Range Disadvantages

- Does not represent actual data distribution
- Very sensitive to outliers
- Does not consider each value of dataset



1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,5

Range = 5 - 1 = 4

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,120

Range = 120 - 1 = 119

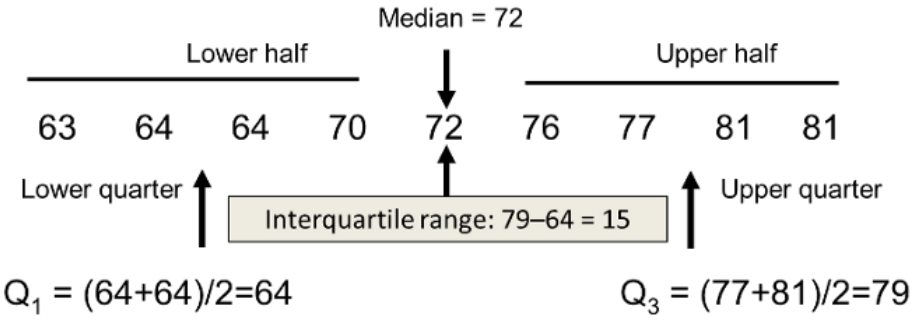
Interquartile Range (IQR)

Measure of Position

- Percentile
- Quartile
  - Divide sorted data in quarters
- Standard Score (z-score)

IQR

- IQR is a measure of where the majority of the values lies.
- 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.
- Interquartile range = 3rd quartile - 1st quartile



IQR

- Quartiles are a useful measure of spread because they are much less affected by outliers or a skewed data set than the equivalent measures
- Quartiles are often reported along with the median as the best choice of measure of spread and central tendency, when dealing with skewed and/or data with outliers

Measure of Spread

- Range
- IQR (Interquartile range)
- Mean Absolute Deviation
- Variance
- Standard deviation

Mean Absolute Deviation

- It represents the amount of variation that occurs around the mean value in data set.
- Taking average of sum of absolute difference between each value of data set and mean.

Variance

$$variance = \frac{\sum(X - \mu)^2}{N}$$

- The variance is a numerical value used to indicate how widely individuals in a group vary. It measures dispersion around the mean.
- If the values in data are spread out, the variance will be a large number.
- Conversely, if the values are spread closely around the mean, the variance will be a smaller number

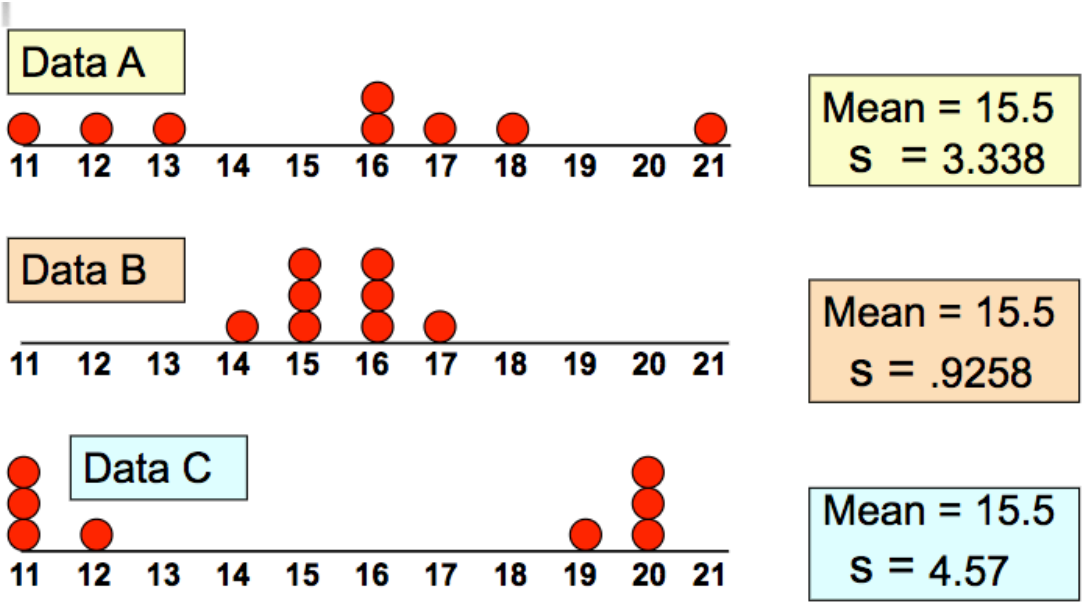
Variance

- Problems in using variance -
  - Gives more weight to extreme scores
  - Variance is not in the same units as the values in our data set, variance is measured in the units squared

Standard Deviation

- It is the measurement of average distance between each quantity and mean. That is, how data is spread out from mean
- It is square root of variance.
- A low standard deviation value indicates that the data points tend to be close to the mean, while a high value indicates that the data points are spread out over a wider range of values.

Standard Deviation



Measure of Position

Measure of Position

- Percentile
- Quartile
- Standard Score (z-score)

Standard Score (z-score)

- A standard score indicates how many standard deviations an element is away from the mean.
- Z-score is a measure how much the data differs from mean. We can determine whether it is different enough to be significant.

Standard Score (z-score)

- To calculate a z-score, we take the individual value and subtract the mean and then divide this difference by the standard deviation.
  - If z is negative then x is below average
  - If z is 0 then x is equal to the average
  - If z is positive then x is above the average

$$z_i = \frac{x_i - \bar{x}}{s}$$