

Fundamentals of Data Analytics

Rishikeshavan *

April 2021

1 Introduction to Basic Numeric Descriptive Measures

1.1 Collection of Data: Population and Sampling

- **Population:** The universe of individuals or objects that are required to be analysed is known as population. The method of information collection from the entire population is called *census*.
- **Sample:** However, most of the times it is difficult and sometimes impossible to measure every unit in the population. In such cases, a subset of the population is considered for analysis, wherein, the subset is highly representative of the overall population. This subset is called the sample and the method of data collection is called *sampling*.

A **parameter** is a value that is usually unknown and is associated with the Population. It is usually represented with Greek Letters (e.g. σ for Population Standard Deviation, μ for Population Mean). A Parameter requires being estimated with minimum error using **Statistics**, which are associated with the Sample. Statistics are represented by an English alphabet like 'S' signifying standard deviation.

*LEAPS Analyttica - <https://leapsapp.analyttica.com/courses/Fundamentals-of-Data-Analytics>

1.2 Types of Data

Data can be broadly classified as **qualitative** and **quantitative**.

1. **Qualitative:** Also known as **Categorical data** in data analysis.
 - **Nominal data:** Qualitative data without order. (*e.g.*) Types of schools could be Public, Private, Vocational etc. They do not hold any order.
 - **Ordinal data:** Qualitative information with order, meaning that the measurement classifications are different and can be ranked. It is also known as **ordered categorical data**. (*e.g.*) The grading system of A, B, C, D where A is ranked higher than B and so on.
2. **Quantitative:**
 - **Interval data:** Measures with order and gives numerically equal distances on scale.
 - **Quasi-Interval data:** It is a combination of Ordinal and Interval data. (*e.g.*) A poll with options from Strongly Agree to Strongly Disagree.
 - **Ratio data:** They have equal intervals and a ‘true’ zero point. It has all the properties of Interval data with a clear definition of a true zero point.

1.3 Representation of Data

Representation of data in the most relevant manner is a key step towards analysis. Data can be represented in various ways as follows:

- **Tabular**
- **Textual**
- **Semi-Tabular**
- **Graphical**

1.4 Measures of Location

Whenever any metric of interest is measured, a fairly large number of observations in the sample tend to center around a single value. This value can be considered as a single representative value of the desired metric for the sample. As these values give the ‘location’ of a ‘central’ value of the sample, they are called **Measures of Location** or more commonly **Measures of Central Tendency** of the sample.

Some of the most important Measures of Central Tendency are: Mean, Median, Mode, Percentile and Quartile.

1.4.1 Mean

It is the average value of any data series and can be defined for all ratio-scale and interval-scale data. Mean can be given mathematically as,

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

Note that the total deviation around the mean is always zero (*i.e.*) $\sum_{i=1}^n (X_i - \bar{X}) = 0$.

One major disadvantage of mean is that it is particularly susceptible to outliers or extreme values in the data. So, you need to be careful when summarizing data using mean as outliers present in the data can distort this measure. Outliers need to be treated or data needs to be normalized in presence of extreme values.

1.4.2 Median

It is the middle number of any data series which has been sorted in ascending or descending order. Median is defined for ordinal data too, along with interval-scale or ratio-scale data.

However, the median is based on only the relative positioning of the observations, not their actual value. Hence, often it may not reflect the slow shift in the sample or population.

1.4.3 Mode

It is the most frequent number in the array. Mode is defined for all kinds of data, viz. nominal, ordinal, interval- scale, and ratio-scale.

If data is normally distributed or symmetric, then mean, median and mode are the same. However, for non-symmetric/skewed data, median is preferred, as that remains unchanged and is not affected by skewness in the data. Also, mode is mostly used for categorical (Nominal) data, where we need to find out which category is the most frequent. A general rule of thumb for finding the best measure of central tendency is listed below:

- For Symmetric interval or ratio data, mean is the central tendency measure to be used.
- For Skewed interval or ratio data, median is used.
- For Ordinal data, median is used.
- For Nominal data, mode is used.

1.4.4 Percentile and Quartile

The p^{th} percentile is a value such that at most $p\%$ of the observations are less than this value and that at most $(1 - p)\%$ are greater, when the data is sorted. The terms associated with percentiles are:

- 25th percentile of the data is the **First Quartile (Q1)**
- 50th percentile is the Median or the **Second Quartile (Q2)**
- 75th percentile is the **Third Quartile (Q3)**

1.5 Measures of Variability

Variability measures how much the data is spread or scattered. A variable with less variability indicates more confidence in making any conclusion from data based on *location parameter*. Hence, variability of a data indicates how much we still do not know about the data. The measures used to describe variability in a data are known as **Measures of Dispersion**.

Some of the most important Measures of Dispersion are: Range, Interquartile Range (IQR), Variance, Standard Deviation and Coefficient of Variation.

1.5.1 Range

It is the difference between the minimum value and the maximum value in the dataset. It offers a crude insight into the spread of the data, but is very susceptible to outliers. This measure does not make any assumption about the distribution of the data.

1.5.2 Interquartile Range (IQR)

It is the difference between the third quartile(75th) and the first quartile(25th) of the data (i.e.) it measures the spread of the middle half of the data. So, by definition it accounts for only 50% of the data. Like range, IQR also does not make any assumption regarding distribution of data (i.e. it is non-parametric).

As a rule of thumb, data points are more spread out as the IQR goes up and if IQR is small they are assumed to be uniformly spread around the mean. For normally distributed data, the IQR is closely related to the standard deviation and is about 35% greater than standard deviation. IQR can also help to determine the outlier in the data series. Data values that deviate from twice the IQR are often defined as outlier, where as those deviating more than 3.5 times of IQR are far outliers. *Note, data needs to be sorted in ascending order to calculate IQR.*

1.5.3 Population and Sample Variance

To get the ‘variability’ of the data from the ‘central’ value, we use the concept of **deviation**. Deviation of a data point is usually measured from mean. But we have seen before that the total deviation around mean is always zero. Hence, we measure the “average squared deviation” from mean μ . This measure is called **Variance** and is represented by σ^2 .

$$\sigma^2 = \frac{\sum_{i=1}^n (X_i - \mu)^2}{N}$$

In the above formula, the assumption is that the population mean μ is known and thus the variance is called as **Population variance**.

However, if μ is unknown, then the variance can be calculated using \bar{X} , the sample mean, and will known as **Sample variance**.

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}$$

1.5.4 Standard Deviation

Mathematically, Standard deviation is the square root of Variance and it is the parametric equivalent of Inter Quartile range. This together with variance represents the **variability** in the data around mean. This is also often termed as **volatility**.

$$StandardDeviation = \sqrt{Variance}$$

Standard Deviation and Variance together are important measures of variability.

1.5.5 Coefficient of Variation

The CV is a measure of relative variation which is used to compare the variability in two or more data series with different units or very disparate averages. It is the ratio of the standard deviation to the mean. It is a unit free of measure and is measured in percentage.

$$Coefficient of Variation = (\sigma / \bar{X}) * 100$$

Lower Coefficient of Variation indicates lower dispersion around the mean, indicating higher stability.

1.6 Basic Bivariate Analysis

Analysis of single variable is called *Univariate Analysis*. The analysis using two variables is called *Bivariate Analysis*. When more than two variables are analysed together, that is referred to as *Multivariate Analysis*.

1.6.1 Pearson's Correlation and Covariance

Correlation is a measure of the strength of association of two variables. A positive value indicates that if one increases the other also increases, where as a negative value indicates the opposite.

The most common measure of correlation for numerical data is the **Pearson's Correlation**. This measures the degree of linear relationship between 2 numeric variables and lies between -1 to +1 and is represented by r .

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

- $r = 1$ indicates a perfect positive correlation
- $r = -1$ indicates a perfect negative correlation
- $r = 0$ indicates no *linear* correlation (does not mean no correlation)

The numerator in the above formula represents another measure called **Covariance** and is calculated by

$$Cov(x, y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n(n-1)}$$

Pearson's correlation is a scaled measure and does not have any unit. Hence, it is often preferred over covariance to understand the association between the variables.

1.6.2 Rank Correlation

Spearman Correlation is a rank correlation that works on ordered data. Rather than looking at the absolute value of an observation, it looks at the order of the observation in the entire data. Unlike Pearson's correlation coefficient, Spearman Rank Correlation measures the degree of monotonic (move in the same direction but not at a constant rate) relationship between two variables.

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where d_i = difference in paired ranks and n = number of observations. Like Pearson's Correlation, Spearman Correlation also lies between -1 and +1.

- $r_s = 1$ indicates a perfect positive correlation
- $r_s = -1$ indicates a perfect negative correlation
- $r_s = 0$ indicates no *monotonous* correlation (does not mean no correlation)

1.6.3 Correlation does not Signify Causation

In two correlated variables, change in magnitude of one variable does not indicate that it will cause change in the other variable. *Correlation does not signify causation*. Rather, correlation merely suggests of related movement in the same direction of these two variables.

Whereas, if there is a causal relationship between two random variables, it will automatically imply that if one changes that will cause change in the other one. It is very important to remember the difference between association and causation. It is also interesting to note that causation is an asymmetric relationship whereas correlation is symmetric relationship.