Data Analysis Note: Funnel, AARRR, and Cohort Analysis Those with knowledge and experience in the field

Statistics Note

Chapter 1: Introduction to Statistics

1. Basically, the statistical analysis is meant to collect and study the information available in large quantities.
   1. For example, we want to analyze numerical data such as Name, number, age.
2. Statistics is a branch of mathematics, where computation is done over a bulk of data using charts, tables, graphs, etc.
3. The data collected for analysis here is called measurement, now if we have to measure the data based on a scenario, a sample is taken out of a population.
4. Then the analysis or calculation is done for the following measurement.

Types of DATA and statistical analysis:

1. Types of statistics
2. Descriptive Statistics
3. Inferential Statistics
4. Types of data

**Descriptive Statistics**

* Descriptive statistics summarizes or describes the characteristics of a data set.
* Descriptive statistics consists of three basic categories of measures: measures of central tendency, measures of variability (or spread), and frequency distribution.
* Measures of variability describe the dispersion of the data set (variance, standard deviation).
* Measures of frequency distribution describe the occurrence of data within the data set (count).

**Inferential Statistics**

A set of method that is used to draw a conclusion about the characteristics of population based on the sample of the data

Used to find the population parameter when you have no initial number to start with

The two main areas of inferential statistics

1. Estimating parameter
2. Hypothesis testing

Chapter-2:

**Basics Types of Data:**

Qualitative:

1. Nominal data
2. Ordinal data

Quantitative:

1. Discrete data
2. Continuous data

Qualitative or Categorical Data

Qualitative data, also known as the categorical data describes the data that fits into the categories.

* Qualitative data are not numerical.

The categorical information involves categorical variables that describes the features such as a person’s gender, hometown etc.

Sometimes categorical data can hold numerical values (quantitative value), but those values do not have a mathematical sense.

Examples of the categorical data are birthdate, favorite sport, school postcode. Here, the birthdate and school postcode hold the quantitative value, but it does not give numerical meaning.

Nominal data

Nominal data is one of the types of qualitative information which helps to label the variables without providing the numerical value.

Nominal data is also called the nominal scale. It cannot be ordered and measured.

But sometimes, the data can be qualitative and quantitative. Examples of nominal data are letters, symbols, words, gender, etc.

The nominal data are examined using the grouping method. In this method, the data are grouped into categories, and then the frequency or the percentage of the data can be calculated.

These data are visually represented using the pie charts.

Quantitative Data

Quantitative data is also known as numerical data which represents the numerical value (i.e, how much, how often, how many).

Numerical data gives information about the quantities of a specific thing. Some examples of numerical data are height, length size, weight and so on.

The quantitative data can be classified into two different types based on the data sets.

The two different classifications of numerical data are discrete data and continuous data.

Discrete Data: can take only discrete values. Discrete information contains only a finite number of possible values. Those values cannot be subdivided meaningfully. Here, things can be counted in whole numbers.

Continuous data: is data that can be calculated. It has an infinite number of probable values that can be selected within a given specific range.

Chapter 3: Sampling Techniques

Agenda:

1. Population and samples
2. Type of sampling technique - 3
3. Random sampling
4. Non-probability sampling
5. Population sampling

Population and sample

Population is an entire group that is required to be analyzed is known as population.

Sample is the subset of population, the specific group of individuals that you will collect data from the population is known as sample.

Chapter 3:

Why sampling is important?

* Gathering data from entire population is not possible. Sampling is applicable in such situation.
* Using sampling one can make information faster.
* Surveying and measuring everyone is not cost effective.
* We can easily analyze the data when using sample of the data.
* A smaller set of individuals often results in lesser data collection error.

Types of sampling:

Probability Sampling and Non-Probability Sampling

Probability: Every member of the population has an equal chance of being selected.

Examples:

1. Simple random sampling
2. Stratified random sampling
3. Cluster sampling
4. Systematic sampling

Non-probability sampling

* Samples are selected on the basis of judgement or the convenience of accessing data
* Largely depends on a researcher sample selection skill:
  + Examples:
    - Convenience sampling
    - Purposive sampling
    - Voluntary response sampling
    - Snowball sampling

Probability Sampling

1. Random: Randomly choose a member from the population
2. Systematic: Similar to simple random sampling.
3. Stratified: First divide the population into groups
4. Cluster: Divide the population into groups, and then randomly select the group from all the groups.

Non-Probability Sampling

1. Convenience Sampling: Include the respondents/members who are easy to reach for researcher.
2. Purposive sampling:
   1. Select a sample based on the purpose of the research.
   2. Researcher select the sample by using their expertise and knowledge.
3. Voluntary Response sampling:
   1. Based on the ease of access.
   2. Members volunteer themselves instead of researcher selecting the participants and directly contacting them.
4. Snowball Sampling:
   1. Recruit the participants via research participants for test or study
   2. Used where it’s hard to find the potential population for research.

Population Sampling

Analyzing or testing entire population is impossible and also a cost and time taking.

Population sampling is the process of selecting a subset of the objects that is representative of the entire population. The sample must have sufficient size of the objects to warrant statistical analysis.

Must be perform correctly since errors can lead to inaccurate and misleading result.

Why n-1 and why not n?

This is actually called as Bessel’s correction.

This method corrects the bias in the estimation of the population variance.

It also partially corrects the bias in the estimation of the population standard deviation

The idea behind this is that this is a more unbiased measure of variance than the usual definition.

Chapter 4: Descriptive Analytics

Agenda:

1. Measure of central tendency
   1. Mean, median and mode
2. Measure of dispersion
   1. Range,
   2. Inter quantile range
   3. Variance
   4. Standard variance
   5. Median variance

Measures of central tendency: or measures of central location

1. Measures the center value of the dataset.
2. Give us idea about the concentration of the value in the central part of the distribution.

Mean often called as average.

Mean: an average set of observation of the data. It computes the sum of all observation present in the datasets divided by total number of observations.

Add/Sum each number/observation present in the dataset.

Calculate the total number present in the dataset.

Divide sum of observation to the total number of observations.

¯¯X = ∑x / N

Median: arrange the observation is ascending order. The middle number and found by ordering all data points and picking out the one in the middle. If there are two middle numbers taking the mean of those two numbers.

Chapter 4: Mode

The value which occurs most frequently in the set of the observation.

It can have more than one mode as an uni-modal, bi-modal and multi-modal. Mostly useful for categorical data.

**Measure of dispersion:**

Measure of dispersion indicates that how the data is dispersed from the measure of central tendency.

Range, inter quartile range, variance, standard deviation and mean deviation.

Range:

- Simplest measure of dispersion

- Measures the difference between highest value and lowest value present in the dataset

- Used to construct control chart in quality assurance

- Useful when you want to focus on extremes values of the dataset.

Interquartile range:

Measure the middle 50% of the data

Indicates how the data is dispersed around the mean.

Difference between the third quartile and first quartile value of the dataset.

Helpful to detect the outlier present in the dataset.

Formula: IQR =Q3-Q1

Variance:

Measure the dispersion of the data around the mean of the data

Indicate how the data is dispersed from its mean.

If the value of variance is closer to mean then it’s a low variance.

If there is significant difference in the value from the mean then it is a high variance.

Standard deviation:

Most important and frequently method in measure of dispersion

Simply the square root of the variance

Indicate how far away the datapoints is dispersed from the mean.

Mean Deviation:

Average sum of the absolute values of the deviation from any arbitrary values, eg. Mean, median, mode, etc.

Suggest to calculate from median because its give least value when measured from the median.

The deviation of an observation xi from the assumed mean

* + - A is defined as (xi-A)
* Therefore, the mean deviation can be defined as:

Mean deviation = ∑ |x-µ| / N