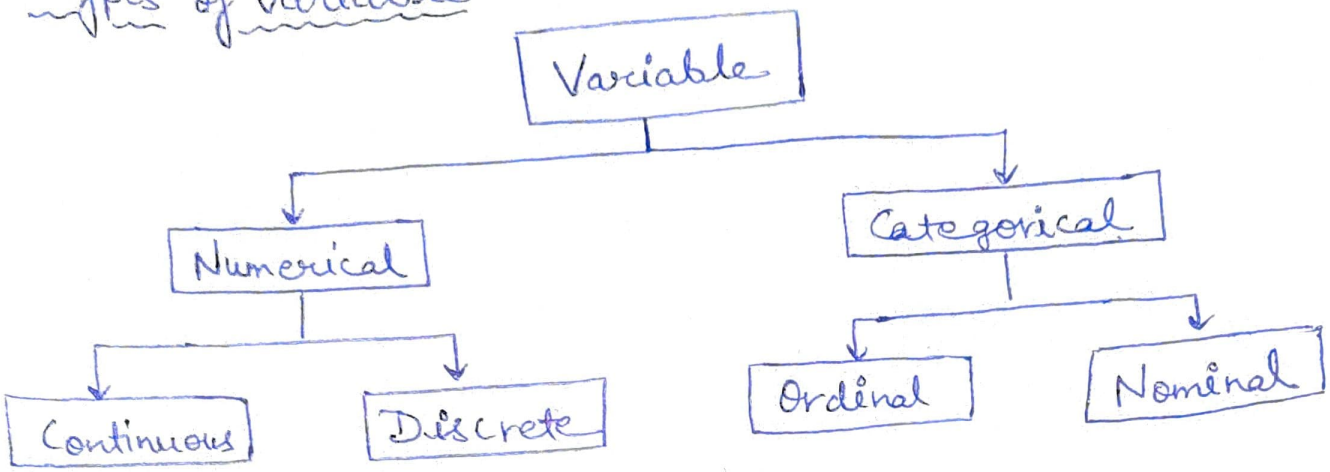
Statistics for Data Science

Day-1

Variables: A variable is any characteristic, number, or quantity that can be measured or counted. It can also be called as a data item.

The characteristics / number / quantity is called a variable because it can vary between data units in a population and may change in value over time.

Types of variable



→ Numerical variables have a value that describe a measurable quantity as a number, like 'how many' or 'how much'. Therefore, numerical variables are also called quantitative variables.

→ A continuous variable is a numeric variable, in thes observations can take any value between certain set of real numbers. Ex → Height, Weight, Temperature, Pressure, etc.

→ A discrete variable has observations that takes values based on a count from a set of whole values which are distinct. Ex → No. of collected coupons, No. of registered cars, etc.

→ Categorical variables have values that describe a 'quality' or 'characteristic' of a data unit, like 'what type' or 'which category'. Categorical variables fall into mutually exclusive groups (in one category or another) and exhaustive (include all possible options). They are qualitative in nature & are represented by non-numeric values.

→ A ordinal variable is a categorical variable that holds values that can be logically ordered or arranged. The categories associated with ordinal variables can be ranked higher or lower than other, but do not necessarily establish a difference between each category.

Ex→ Academic grades (A, B, C), clothing size (S, M, L, XL), etc.

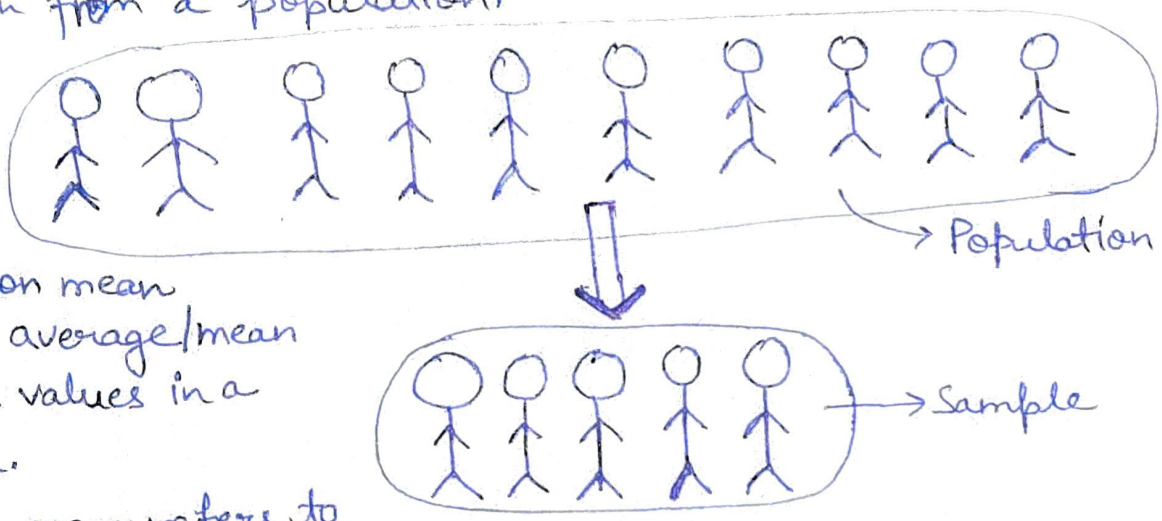
→ A nominal variable is a categorical variable that can take values which cannot be organised in a logical order.

Ex→ Sex of an individual, Business type, Eye colour, Religion, etc

(II) **Random Variable:** A variable whose values depends on outcomes of a random experiment are called random variable.

(III) **Population:** Population refers to the total set of observations that can be made.

(IV) **Sample:** A sample refers to a set of observations drawn from a population.



* Population mean refers to average/mean of all the values in a population.
* Sample mean refers to the mean of all the values present in a sample.

(V) **Sampling:** The process of selecting few observations from a large population is called sampling.

It is of two types —

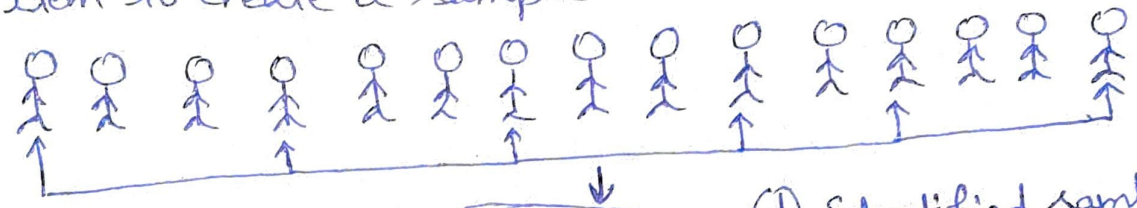
(i) Probabalistic Sampling (ii) Non-probabalistic Sampling

Objective of sampling: The goal of sampling is to come up with a sample space that should reflect all the attributes of a large population. It should reflect the attributes very closely because in case if its not, the results derived on sample population will not work on original population and hence resources will be lost.

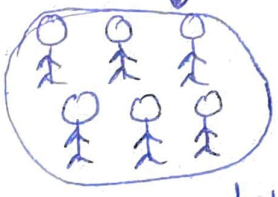
(i) Probalistic Sampling: It is based on the principle of randomization which ensures all the items present in a population has equal chances of getting selected in the sample. It generally allows us to make strong statistical inference. It can be sub-divided into several types —

(a) Simple random sampling: In this sampling strategy, has item has equal chance of getting selected in the sample. It is generally used when we do not have any proper information about target.

(b) Systematic sampling: From a population, we select every n^{th} item to create a sample.



(C) Cluster-based sampling: In this first we cluster items based on their similarity and then we select random items from these clusters.



(d) Stratified sampling: A stratified sampling involves dividing the entire population into homogenous groups called strata (plural for stratum). Then random samples are selected from this sample keeping in mind to retain the distribution of people as in the stratum, so that it can best represent original population. We need to have prior information of data.

(ii) Non-probabalistic Sampling: It does not follow the principle of random selection, rather it is more dependent on user's selection approach. The outcome of such sampling can be biased and it makes difficulty for all the items in the population to be a part of the sample.

It can be further sub-divided into—

(a) Convenience Sampling: In this methodology, items are selected from the population, as per user convenience and availability.

(b) Purposive Sampling: In this sample is selected based on some specific purpose.

(c) Quota sampling: It is selection based on some preset criteria like, the sample should have same/exact proportion as it is in the population.

(d) Referral/ Snowball sampling: It is used when population is completely unknown. Here, we take the help of first element to select second, and then second helps to select more and in this manner sample is selected.