# **Session1- Intro to Statistics**

What is data?? Types of data:

1. Quantitative vs Qualitative 🡪 1-10( numbers) , good bad ( qualitative)
2. Discrete and continuous:
   1. Ex: runs in cricket 🡪 always a whole number ( like 100, 110 and not 100.5)
   2. Continuous data can be any number within lower – upper boundary
      1. Ex: height and weight

## **What is Statistics**

Its data science 🡪 discipline that allows us to gather , analyze , depict and make sense of data to extract useful information from raw data.

Data 🡪 information

Why do we need statistics??

1. **Data access widens and easier to access by anyone**
   1. We are living in an information age 🡪 news passes very easily all over the world.
2. **Data overload**
   1. Due to too much of data moving around , at times leads to information overload , sometime the line between facts and opinion gets blurred that affects decision making.
3. Leading to simple rules of thumb
   1. When too much data , we as human beings always tend to fall back on simple rules of managing and categorizing data based on some simple thumb rules.
   2. These rules gives a perception that we are in control of data rather than data controlling our emotions and decision making
4. **Develop shortcuts**
   1. Sometime when things work in the past , so we tend to follow the same when we are in the same situation.
   2. Not cos of rationale but cos we see similar parameters as we have seen earlier and the decision that we have made earlier.

## **Things to be taken care of when handling data**

1. Agenda driver data 🡪 people sample data based on some bias when they have some hidden motive. And the same is used for presenting to other party. This affects decision making as well
2. Social media 🡪 once bad data is in news 🡪 media magnifies it 100 folds and people use it for personal motives and at times leads to violent clashes.

Understanding statistics can help in this situation:

1. Look for red flags when you detect that data is sampled incorrectly
2. Asking right questions to separate fact from fiction

## **Big data and data analytics**

What’s big data?? 🡪 not only refers to quantity of data but also scale of data which wouldn’t have been possible few years back ( may be due to tech limitations). Ex: location data on earth, now we have google earth , GPS

Similarly so many apps in phone is having access to our location right.

Data analytics 🡪 analyzing big data and finding patterns

## **Data Collection and sampling**

Data collection and storage -> manually collect data from public -> old way of collecting data . as tech evolved , we have devices which can do it on our behalf.

Population vs sampling

As population has gone up , at times its impossible to collect data for entire population , so we sample the data based on certain conditions . 2 things which we must keep in mind when sampling:

1. Bias 🡪 sample must repr entire population , no personal bias else you cant extrapolate to the population
2. Noise 🡪 even if sample is good, there will be some noise , spread when compared to entire population

## **Data descriptives**

Pointers that describe the data . helps in categorizing the data and summarizing it. Used for analyses

Ex: measures of location , measures of dispersion / divergence 🡪 how much variance is there in data , skewness 🡪 whether data is symmetrical or asymmetrical

## **Data distribution**

Visual description of the data 🡪 tables, histogram, grouping, pivoting

Once we have data distributed , then we can convert it into one of the distribution ( normal distribution)

## **Data relationships**

Once we have these different variables representing the data, we can start looking of any potential relationships between these variables.

Finding dependencies, linkages between these variables and also looking to make a prediction using these relationships. Forecasting , prediction ,

Directly proportional or inversely. Also, what’s the convexity between them. Like how much one variable should change to bring about a change in another variable.

## **Probability**

This can be a discrete or a continuous event.

Discrete 🡪 either this or that. Heads or tails. You have finite | limited number of outcomes from an event happening.

Continuous 🡪 there are infinite number of outcomes of an event.

**Probabilistic tools**

1. Probit / logit 🡪 measures that tell us the likelihood of an event to happen based on observable variables.
   1. Ex: an asteroid coming to earth, what’s the probability of it entering earth atmosphere, provided its speed, mass etc.…
2. Decision trees 🡪 where one event decides the fate of other. Dependent variables / outcomes.
3. Monte Caro Simulations 🡪 multiple inputs are provided and a distribution is calculated for each input variable. :
   1. Ex: to value a company, you take inputs based on cashflow, growth, risk etc..

# **Session-2- Sampling and population**

Population 🡪 universe of all instances of object that you are trying to study

Sample 🡪 subset of the entire population.

## **Types of sampling**

1. timeseries sampling 🡪 stock market data (dates back to 1860) , but if we need to analyze , we will probably sample to few decades only ( ~1960 – 2020).. this is called sampling based on timeseries.

Similarly, sampling can be done based on major parameters of the system.

1. Cross section 🡪 looking for some parameter for the object to be part of sample. Ie. Revenue > 10mil

**Why do we need sampling**

Practicality 🡪 if population is too large, then its almost impossible to take entire population into account. Instead, we take sample of data only.

Costs 🡪 even if we do collect the data , there is too much ( time and money) required to analyze and produce working results from that.

Time trade off 🡪 we also need to factor that population keeps changing, and maintaining the updates for the entire population is almost infeasible. Instead sampling update is much more feasible

Ex: census of a country. Population is increasing as the census happens..Instead we sample the data for census in such as way that it represents the entire population of data.

## **Sampling approaches**

Probability vs non-probability sampling

In probability sampling, the observations are picked at random with no bias, however for non-probability sampling , we choose certain characteristics from the population data and then based on that we select a sample.

**Variants of random sampling**

1. Simple Random 🡪 pick random from the population
2. Stratified random 🡪 break your population into groups, then select randomly from those groups
3. Cluster random 🡪 break the population into groups and then randomly select some of those groups and collect data from those chosen groups.

**Sampling bias**

A biased sample is the one that diverges from the population in its characteristics. And this bias can arise from many reasons:

1. Exclusion 🡪 data that’s missed when collecting the samples. For ex: in a country, remote places which doesn’t covered in census
2. Self-selection 🡪 some parts of population are easily accessible from others , based on the way we collect data.
3. Non-response 🡪 parts of population which doesn’t respond to requests of data. Ex: for voting cast , there are people who can vote but don’t respond and hence gets excluded based on their response
4. Survivorship 🡪 if success then part of population , else not