Introduction

What and Why ?

Statistics is the study of methods of collecting, analysing, presenting and interpreting data.

Collecting data : we have various methods for collecting data like random sampling.
Analysing : correlation and causation
Presenting : mean, mod, variance
Interpreting : getting insights, like people between age 20-30 are more likely to buy a particular smartphone

We know that data is new oil. So from data we can do many many things. For example we have data of clothing brand, which has customer age, billing amount, type of cloth purchased and time of purchase. We can get insight that people of an age group likes some kind of clothes, during January jackets are more sold, people having billing amount higher than 10000 are likely to purchase this kind of clothes, etc. So using statistics we can get insights and predict things.

Application :

1. Six Sigma : SS is a statistical method and set of principles and tools for minimise the error and maintain and improve the process performance. The term “Six Sigma” refers to a term when the process generates 3.4 defects in one million opportunities.

2. Business : In your intern you used various strategies to make portfolio, like company having m_cap growth higher than cutoff, company having stable DE, ROCE, etc. The cloth business example given above can be considered.

3. Weather forecast : in weather forecast we see the data and based on wind speed, humidity, temperature we forecast that it will rain today or not. Or the updates of any disaster.

4. In medicine trial : when new medicine or vaccine is developed it is tried on various samples of population and various insights are drawn. Lets say people with particular range of age, from particular region. Have some common effect, etc.

There are many applications of statistics. Remember data is NEW OIL.

Types of Data :

There are basically two types of data numerical data and categorical data.

1. Numerical Data : This data is represented by number. Like age, weight, etc. It has two types,
    a. Discrete : age, number of people it can be integer it can not be float
    b. Continuous : weight, height of people, it can be continuous.

2. Categorical Data : This data is represented by category. Like gender, Feedback, etc. It has two types,
    a. Nominal : Each category is different, they don't have relationship between them. Like male and female are different they don't have any relationship.
    b. Ordinal : Category has some relationship. Like feedback from worst to best, we know that worst < bad < not bad < good and so on.

Types of Statistics :

There are basically two types of statistics.
1. Descriptive Statistics
2. Inferential Statistics

Before jumping into types, lets understand two terms population and sample. Population is the whole data. And sample is subset of population. There are many ways to pick sample from population. If we want to know average happiness index of people in India, then population is all the people and sample is subset of it. Now lets understand types of statistics.

1. Descriptive Statistics : As its name suggests, DS uses data to represent population. Like to represent the heights of student in college, we can use mean or some other parameter. By looking at one value we get the insight of whole data, we don't need to see whole data. There are two categories of DS.

    a. Central Tendency : It is used to represent centre point of the data. There are three parameters in CT - mean, median and mod.
    b. Measure of Variability : It is used to measure the variability in dataset, like how far datapoints are from each other. The parameters used are - variance, standard deviation, range.

2. Inferential Statistics : As its name suggests, IS is about making inferences about population from the sample. It generalises to large dataset and applies probabilities to draw conclusion. It is mainly used to get interpretation and draw conclusion. IS is mainly related to and associated with hypothesis testing whose main target is to reject null hypothesis. IS is generally used to determine how strong relationship is within sample.

In practical scenario we use IS, like we can not ask 140 crores people about their happiness instead we pick samples and from those samples we generalise our conclusion to whole population.

Types of Statistical Studies :

There are three types of SS

1. Sample Study : In SS we study the sample which represents the population. i.e., we study the sample of 1 crore people's happiness to draw conclusion of 140 crore people's happiness.

2. Observation Study : In OS we study and analyse the data to get insights and interpretation. Consider the cloth business example. Here we don't manipulate or change the data.

3. Experimental Study : In ES after studying data, we do some experiments. Like after studying results of medicine1 and medicine2 on different samples, now we change the formula of medicine and give to samples again to get results and draw conclusions.

Types of Sampling Techniques :

Sampling techniques or sampling methods are used to pick sample(s) from population. There are two kind of sampling methods
1. Probability sampling : It includes random sampling. Each datapoint has equal probability of getting sampled. It allows to make strong inference about dataset.
2. Non probability sampling : Unlike PS, there are some rules or criteria to select samples.

Here we will discuss the probability sampling. The PS has four types of sampling methods.
1. Simple Random Sampling
2. Synthetic Sampling
3. Stratified Sampling
4. Cluster Sampling

Before jumping into the types, lets discuss sample frame and sample size. The sample frames represents the number of datapoints participate to be sampled where as the sample size is the size of sample. In our case we will consider example of employee data, where sample frame is 1000 and sample size is 100.

Simple Random Sampling :

SRS is the simple way of sampling. We randomly pick datapoints from sample frame. In our example we give number from 1 to 1000 to each employee. We generate 100 random numbers and pick the employees.

Pros :
1. No sample bias
2. Simple sampling
3. Requires no domain knowledge
4. Get balanced sample

Cons:
1. Population size should be high
2. Can't represent population well sometimes

Synthetic Sampling :

SS is very easy to conduct, instead of picking random datapoints, we generate a random starting point and pick employee by a regular interval. In our example lets say we randomly select number 5 then, we pick employee with number 5, 15, 25 and so on.

There is an issue with SS, let's say the numbers are given in ascending order of their rank then the sample will be skewed. So there a risk of skipping junior employees.

Pros :
1. Quick & easy
2. Less bias
3. Even distribution of data

Cons :
1. Data manipulation risk (someone manipulates data such at each datapoint at fixed interval is same)
2. Data randomness required
3. Population should not have pattern (ascending in order of seniority)

Stratified sampling :

In SS we divide the population into subgroups (called strata) based on the relevant characteristic (e.g., gender identity, age range, income bracket, job role).

Based on the overall proportions of the population, you calculate how many people should be sampled from each subgroup. Then you use random or systematic sampling to select a sample from each subgroup.

In our example we create strata based on rank, so from these strata we pick individuals. We must pick individuals based on strata size. i.e., if number of clerks are 200 and number of manager are 50, then they should have same proportion in final sample.

Pros :
1. Finds important characteristics in population
2. High precision can be obtained if difference in strata is high

Cons :
1. Can not be formed on population which can not grouped
2. Overlapping datapoints

Cluster sampling :

In CS we divide population into groups. Instead of picking individual from each group we randomly pick the group. This method is good for dealing with large and dispersed populations, but there is more risk of error in the sample, as there could be substantial differences between clusters. It's difficult to guarantee that the sampled clusters are really representative of the whole population.

In our example, lets say we have offices in 10 cities, so instead of travelling to 10 cities, we select all employees of randomly selected 3 cities.

Pros :
1. Requires fewer resources (travelling cost)
2. Reduce variability
3. Advantages of both random sampling and stratified sampling

Cons :
1. Can not be formed on populations without natural groups (there must be different offices)
2. Overlapping datapoints
3. Can not provide general insights of data

Most IMP :

This picture represents all sampling methods.