Sampling Distribution

Sampling distribution is a statistic that determines the probability of an event based on data from a small group with in a large population. Since the population is too large to analyze, you can select a smaller group and repeatedly sample or analyze them. the gathered data, or statistic is used to calculate the likely Occurance or Probability of an event

Each random sample selected may have a diffrent value assigned to the statistic being studied.

For ex:- if you randomly sample data three times & determine mean or the average of each sample, all three means are likely to be different & fall somewhere along the graph That's variability.

bell curve →

→ The number Observed in a population (N)

→ The no. Observed fn in the sample (n)

→ The method of choosing the sample

Types of distribution

① Sampling distribution of mean

② The most common type of sampling distribution is the mean. It focuses on calculating the mean of every sample group chosen from the population & plotting the data points. The graph shows a normal distribution where the center is the mean of the sampling distribution, which represents the mean of the entire population.

② Sampling distribution of proportion.

This sampling distribution focuses on proportions in a population. you select samples & calculate their proportions. The means of the sample proportions from each group represent the proportion of the entire population.

③ T-distribution.

A T-distribution is a sampling distribution that involves a small population. It is used to estimate the mean of the Population & Other statistics such as Intervals, statistical differences & lineae regression.

Sampling & Resampling.

Sampling

Sampling is a process of selecting group of observations from the population to study the characteristics of the data to make conclusion about the population

Probability Non-Probability

Probability Sampling (Random Sampling)

In this type, data is randomly selected so that every observation of population gets the equal chance to be selected for sampling.
- Simple Random Sampling
- Cluster Sampling
- stratified "
- Systematic "

Non probability Sampling

In this type, data is not randomly selected. It mainly depends upon how the statistician wants to select the data
- convenience
- judgmental/purposive
- Snowball /Referral
- Quota

Sampling error :- Errors which occur during sampling process are known as SE(s)

Difference b/w Obs value of a sample statistics & the actual value of a population parameters

Advantage of Sampling

Reduce Cost & Time
Accuracy of Data
Less resource needed

Resampling

→ Resampling is the method that consist of drawing repeatedly drawing samples from the population.

- k-fold cross-validation
- Bootstrapping

In this method population data is divided into k equal sets in which one set is considered as the test set for the experiment while all other set will be used to train the model

Bootstrapping :- In bootstrapping samples are drawn with replacement (i.e One Observation can be repeated in more than One group & the remaining data which are not used in samples are used to test the model.

Statistical Inferences (or) Inferential Statistics

Statistical inference is a method of making decisions about the parameters of a population, based on random sampling. It helps to ~~asses~~ assess the relationship b/w. dependent & indepen Var

The purpose of SI to estimate the uncertainty (or) sample to sample variation

Components

- Sample size
- Variability in the sample
- Size of the Observed differences

Prediction Error

In prediction error refers to the diff b/w the Predicted Values made by some model & the actual Values

→ LINEAR REGRESSION

→ LOGISTIC "

Linear regression – used to predict the ~~V~~ Value of some continuous response variable

We typically measure the prediction error of a linear regression model with a metric known as RMSE

$$\hat{y}_i \quad y \qquad RMSE = \sqrt{\Sigma(\hat{y}_i - y_i)^2/n}$$ — sample size

14 12

Predicted value

observed value

Logistic Regression - used to predict the Value of some binary response variable.

Prediction error of Logistic Reg model is with a metric known as total misclassification rate

$$\text{Total Misclassification rate} = \frac{\#\ \text{incorrect Predictions}}{\#\ \text{total Pre}}$$