UNIT 1

Sampling Distributions, Re-Sampling, Statistical Inference, Prediction Error

Presented By :

Aditya Joshi

Assistant Professor

Graphic Era Deemed to be University, Dehradun

Sampling Distributions

- Sampling distribution of a statistic is a type of probability distribution created by drawing many random samples of a given size from the same population. These distributions help you understand how a sample statistic varies from sample to sample.

- Sampling distributions constitute the theoretical basis of statistical inference and are of considerable importance in business decision-making. Sampling distributions are important in statistics because they provide a major simplification on the route to statistical inference.

DEFINITION

- A sampling distribution is a theoretical probability distribution of a statistic obtained through a large number of samples drawn from a specific population.
- A sampling distribution is a graph of a statistics(i.e. mean, mean absolute value of the deviation from the mean, range, standard deviation of the sample, unbiased estimate of variance, variance of the sample) for sample data.
- Sampling distribution is a theoretical distribution of an infinite number of sample means of equal size taken from a population.

Re Sampling

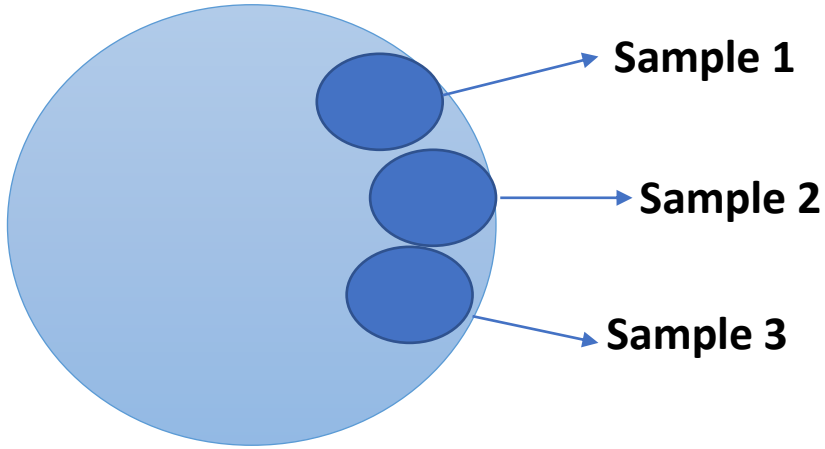
- Resampling techniques are a set of methods to either repeat sampling from a given sample or population, or a way to estimate the precision of a statistic.

- Informally, resample can mean something a little simpler: repeat any sampling method.

Two commonly used resampling methods that you may encounter are cross-validation and the bootstrap.

Bootstrap. Samples are drawn from the dataset with replacement (allowing the same sample to appear more than once in the sample), where those instances not drawn into the data sample may be used for the test set.

Draw more samples from training data and refit a model on each, to obtain additional information.



- A **bootstrap sample** is a smaller sample that is "bootstrapped" from a larger sample. Bootstrapping is a type of *re*sampling where large numbers of smaller samples of the same size are repeatedly drawn, with replacement, from a single original sample.

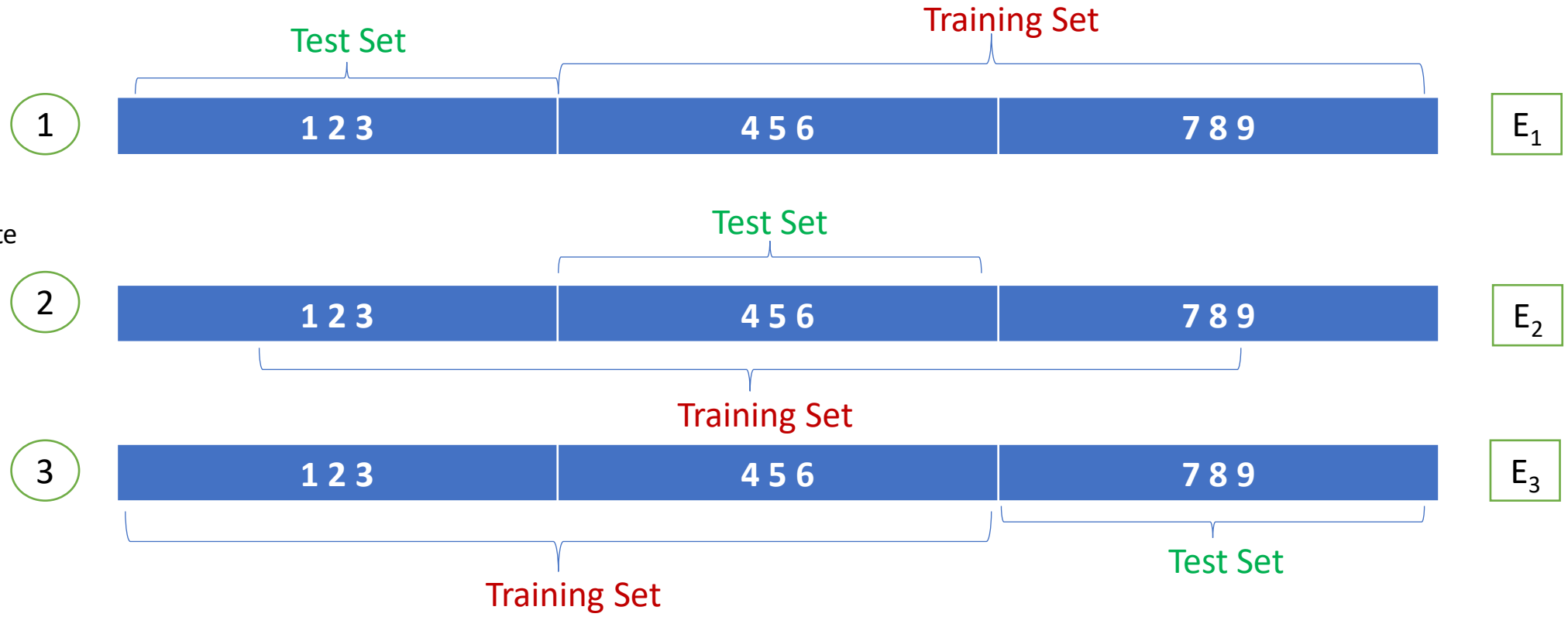
- For example, let’s say your sample was made up of ten numbers:

  49, 34, 21, 18, 10, 8, 6, 5, 2, 1.
- You randomly draw three numbers 5, 1, and 49.
- You then replace those numbers into the sample and draw three numbers again.
- Repeat the process of drawing x numbers B times.
- Usually, original samples are much larger than this simple example, and B can reach into the thousands.
- After a large number of iterations, the bootstrap statistics are compiled into a bootstrap distribution. You’re replacing your numbers back into the pot, so your resamples can have the same item repeated several times (e.g. 49 could appear a dozen times in a dozen resamples).

- **Cross Validation: Cross Validation**(also called rotation estimation, or out-of-sample testing) is one way to ensure your model is robust. A portion of your data (called a holdout sample) is held back; The bulk of the data is trained and the holdout sample is used to test the model. This is different from the "classical" method of model testing, which uses all of the data to test the model.
- It is used to estimate test error or to select appropriate level of flexibility.

k-fold Cross-Validation. A dataset is partitioned into k groups, where each group is given the opportunity of being used as a held out test set leaving the remaining groups as the training set.

3 fold cross validation: k=3

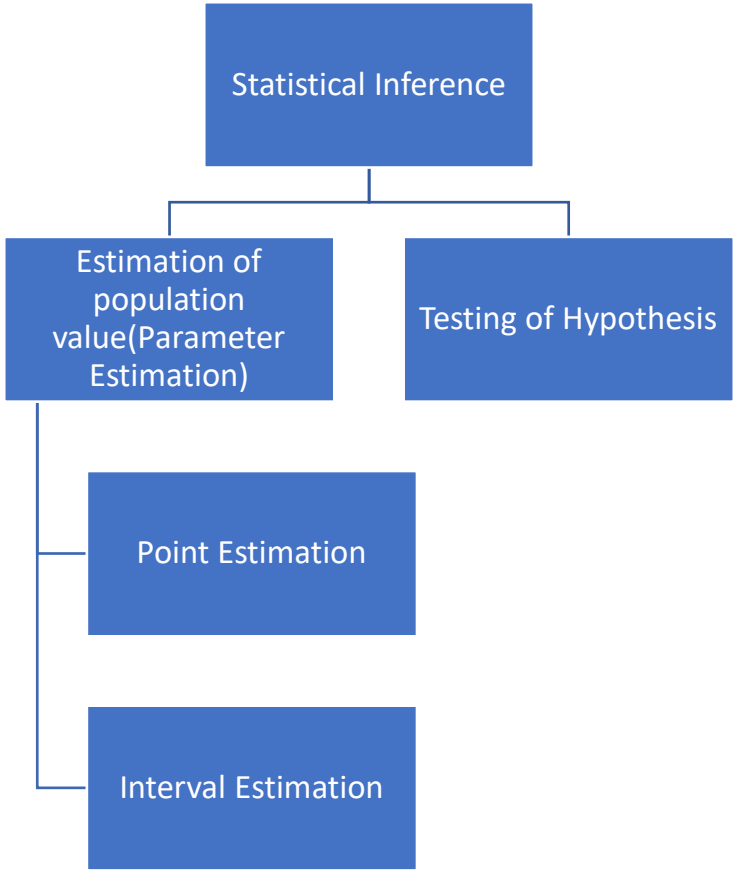


Overall error estimate

$$E = \frac{1}{k}\sum_{i=1}^{k} Ei$$

Statistical inference

- It is a branch of statistics which is concern with using probability concept to deal with uncertainty in decision making.
- It refers to the process of selecting & using a sample to draw inference about population from which sample is drawn.



- In **point estimation**, we estimate an unknown parameter using a single number that is calculated from the sample data.

Example: Based on sample results, we estimate that p, the proportion of all Indian adults who are in favor of social media , is 0.6.

- In **interval estimation**, we estimate an unknown parameter using an interval of values that is likely to contain the true value of that parameter (and state how confident we are that this interval indeed captures the true value of the parameter).

Example: Based on sample results, we are 95% confident that p, the proportion of all Indian adults who are in favor of social media, is between 0.57 and 0.63.

- In **hypothesis testing**, we begin with a claim about the population (we will call the null hypothesis), and we check whether or not the data obtained from the sample provide evidence AGAINST this claim.

Importance of Statistical Inference

Inferential Statistics is important to examine the data properly. To make an accurate conclusion, proper data analysis is important to interpret the research results. It is majorly used in the future prediction for various observations in different fields. It helps us to make inference about the data. The statistical inference has a wide range of application in different fields, such as:

- Business Analysis
- Artificial Intelligence
- Financial Analysis
- Fraud Detection
- Machine Learning
- Share Market
- Pharmaceutical Sector

Prediction Error

Prediction error quantifies one of two things:

- In regression analysis, it’s a measure of how well the model predicts the response variable.
- In classification (machine learning), it’s a measure of how well samples are classified to the correct category.