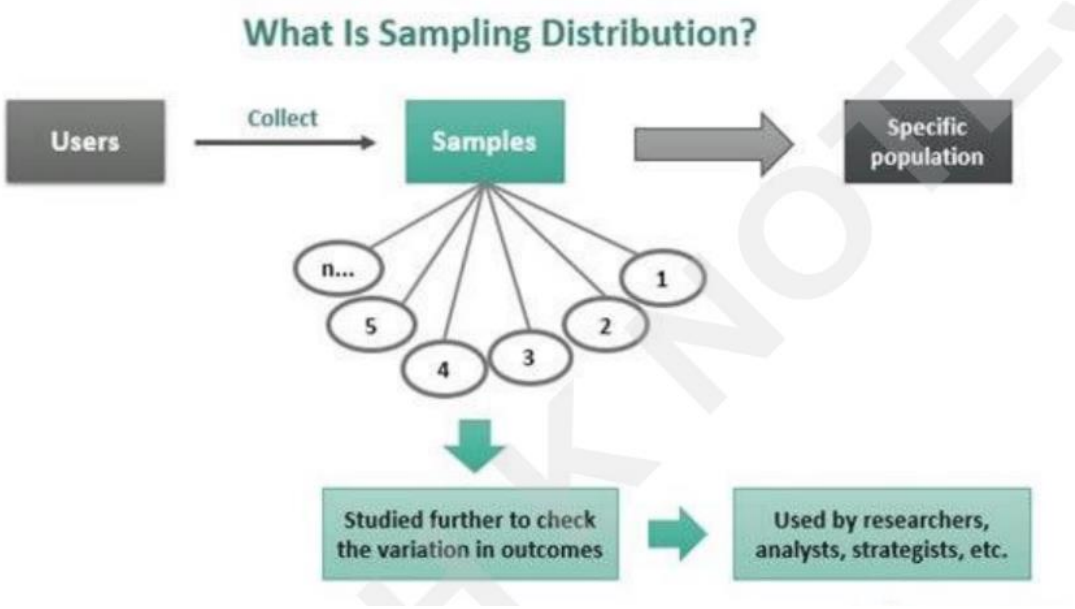
UNIT I INTRODUCTION TO BIG DATA

Introduction to Big Data Platform – Challenges of Conventional Systems - Intelligent data analysis –Nature of Data - Analytic Processes and Tools - Analysis Vs Reporting - Modern Data Analytic Tools- Statistical Concepts: Sampling Distributions - Re-Sampling - Statistical Inference - Prediction Error.

SAMPLING DISTRIBUTION DEFINITION

Sampling distribution in statistics refers to studying many random samples collected from a given population based on a specific attribute. The results obtained provide a clear picture of variations in the probability of the outcomes derived. As a result, the analysts remain aware of the results beforehand, and hence, they can make preparations to take action accordingly.



What are Sampling distributions?

A sampling distribution is a statistical idea that helps us understand data better. It shows the values of a statistic when we take lots of samples from a population. For example, if we want to know the average height of people in a city, we might take many random groups and find their average height. The sampling distribution helps us understand the potential variability in average heights. By analyzing this distribution, entities like governments and businesses can make more informed decisions based on their collected data.

Importance Sampling Distribution in Data Science

Sampling distributions allow data scientists to:

- Estimate Population Parameters: By analyzing the distribution of sample statistics, data scientists can make inferences about population parameters (e.g., population mean or proportion).

- Quantify Uncertainty: Sampling distributions provide a measure of the variability of a statistic, which is crucial for constructing confidence intervals and hypothesis tests.
- Model Performance Evaluation: They help in understanding the variability and performance of models, especially when dealing with small datasets or conducting resampling techniques like bootstrap.

Types of Sampling distributions

1. Sampling Distribution of the Sample Mean ($\underline{x}$)

If the population is normally distributed or the sample size is sufficiently large (according to the Central Limit Theorem), the sampling distribution of the sample mean is approximately normal with mean (μ) and standard error ($\frac{\sigma}{\sqrt{n}}$).

2. Sampling Distribution of the Sample Proportion (\(p\))

If the conditions for using the normal approximation to the binomial distribution are met (e.g., large sample size, $np \geq 10$,$n(1-p) \geq 10$), the sampling distribution of the sample proportion is approximately normal with mean(p) and standard error $\sqrt{\frac{p(1-p)}{n}}$.

3. Sampling Distribution of the Sample Variance (\(S^2\))

Population is normally distributed, the sampling distribution of the sample variance follows a chi-square distribution with \(n-1\) degrees of freedom

RESAMPLING

Resampling Method is a statistical method that is used to generate new data points in the dataset by randomly picking data points from the existing dataset. It helps in creating new synthetic datasets for training machine learning models and to estimate the properties of a dataset when the dataset is unknown, difficult to estimate, or when the sample size of the dataset is small.

Two common methods of Resampling are

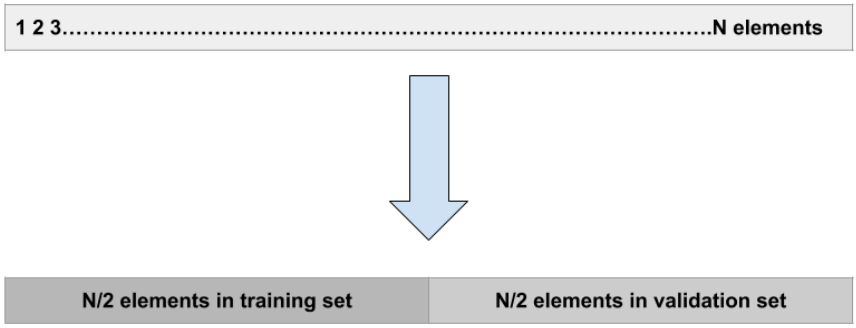
- Cross Validation
- Bootstrapping

Cross Validation

Cross-Validation is used to estimate the test error associated with a model to evaluate its performance.

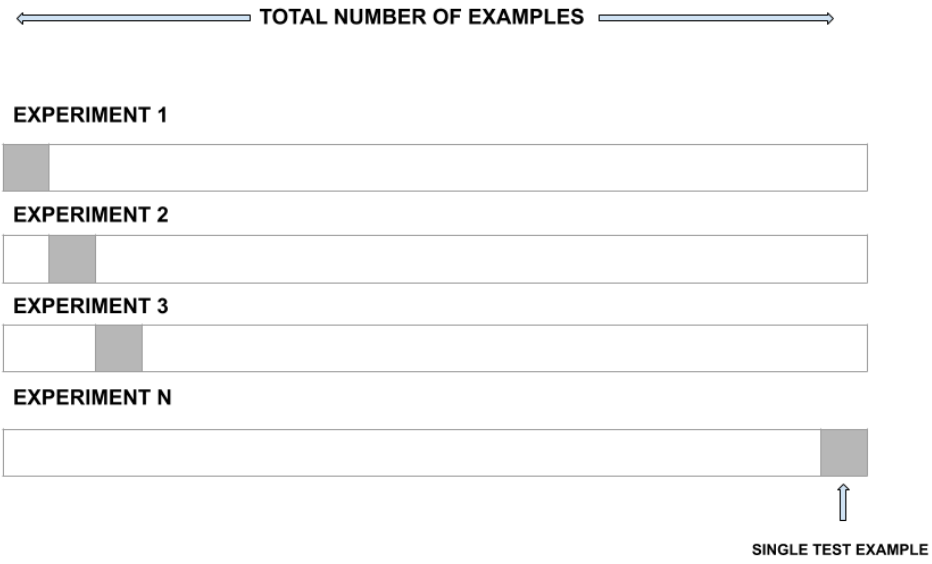
Validation set approach:

This is the most basic approach. It simply involves randomly dividing the dataset into two parts: first a training set and second a validation set or hold-out set. The model is fit on the training set and the fitted model is used to make predictions on the validation set.



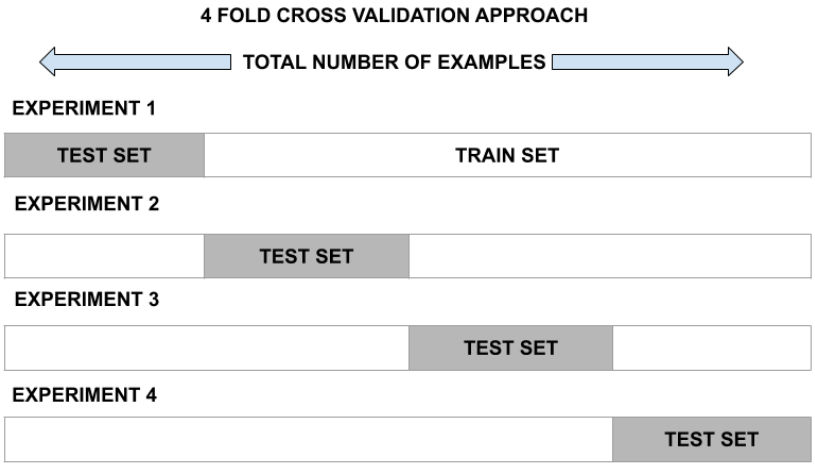
Leave-one-out-cross-validation:

LOOCV is a better option than the validation set approach. Instead of splitting the entire dataset into two halves, only one observation is used for validation and the rest is used to fit the model.



k-fold cross-validation

This approach involves randomly dividing the set of observations into k folds of nearly equal size. The first fold is treated as a validation set and the model is fit on the remaining folds. The procedure is then repeated k times, where a different group each time is treated as the validation set.



Bootstrapping

Bootstrap is a powerful statistical tool used to quantify the uncertainty of a given model. However, the real power of Bootstrap is that it could get applied to a wide range of models where the variability is hard to obtain or not output automatically.

Challenges:

Algorithms in Machine Learning tend to produce unsatisfactory classifiers when handled with unbalanced datasets. The main problem here is how to get a balanced dataset.

Challenges with standard ML algorithms:

Standard ML techniques such as Decision Tree and Logistic Regression have a bias towards the majority class, and they tend to ignore the minority class. They tend only to predict the majority class, hence, having major misclassification of the minority class in comparison with the majority class.

Evaluation of the classification algorithm as measured by a confusion matrix.

		ACTUAL VALUES
PREDICTED VALUES	TRUE POSITIVE	FALSE POSITIVE
	FALSE NEGATIVE	TRUE NEGATIVE

Confusion Matrix

A confusion matrix is a very useful tool for evaluating the performance of a classification model. The diagonal values of the confusion matrix represent the number of correct predictions, and therefore, higher diagonal values indicate better predictive accuracy. The off-diagonal values of the matrix represent incorrect predictions, which can provide insights into the types of errors the model is making. Overall, the confusion matrix is a valuable tool for understanding the strengths and weaknesses of a classification model and identifying areas for improvement.

Handling Approach:

1. Random Over-sampling:

It aims to balance class distribution by randomly increasing minority class examples by replicating them.

2. SMOTE (Synthetic Minority Oversampling Technique)

It synthesizes new minority instances between existing minority instances. It randomly picks up the minority class and calculates the K-nearest neighbor for that particular point. Finally, the synthetic points are added between the neighbors and the chosen spot.

3. Random Under-Sampling:

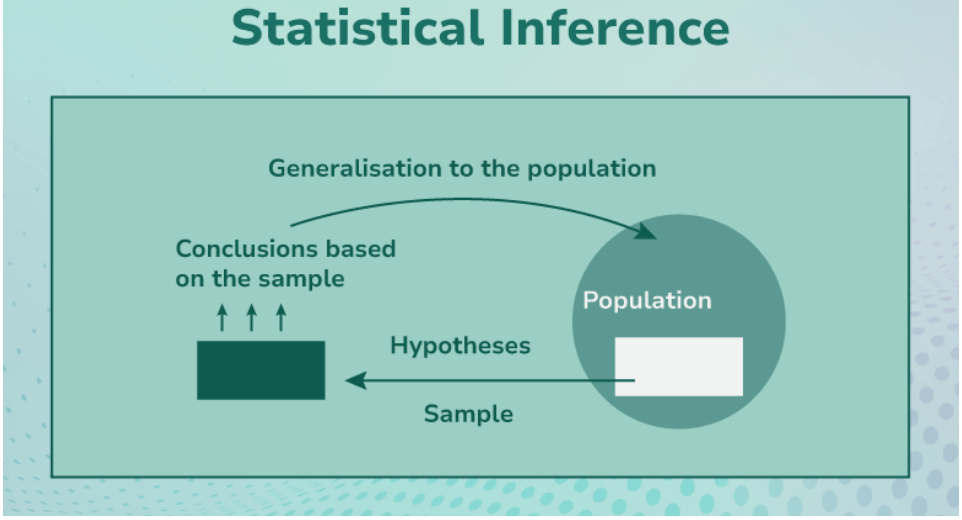
It aims to balance class distribution by randomly eliminating majority class examples. When instances of two different classes are very close to each other, we remove the instances of the majority class to increase the spaces between the two classes. This helps in the classification process.

4. Cluster-based Over Sampling:

K means clustering algorithm is independently applied to both the class instances such as to identify clusters in the datasets. All clusters are oversampled such that clusters of the same class have the same size.

STATISTICAL INFERENCE

Statistical inference is the process of using data analysis to infer properties of an underlying distribution of a population. It is a branch of statistics that deals with making inferences about a population based on data from a sample.



Statistical inference is based on probability theory and probability distributions. It involves making assumptions about the population and the sample, and using statistical models to analyze the data. Statistical inference is the process of drawing conclusions or making predictions about a population based on data collected from a sample of that population. It involves using statistical methods to analyze sample data and make inferences or predictions about parameters or characteristics of the entire population from which the sample was drawn.

Consider a scenario where you are presented with a bag which is too big to effectively count each bean by individual shape and colours. The bag is filled with differently shaped beans and different colors of the same. The task entails determining the proportion of red-coloured beans without spending much effort and time. This is how statistical inference works in this context. You simply pick a random small sample using a handful and then calculate the proportion of the red beans. In this case, you would have picked a small subset, your handful of beans to create an inference on a much larger population, that is the entire bag of beans.

Branches of Statistical Inference

There are two main branches of statistical inference:

- Parameter Estimation
- Hypothesis Testing

1. Parameter Estimation

Parameter estimation is another primary goal of statistical inference. Parameters are capable of being deduced; they are quantified traits or properties related to the population you are studying. Some instances comprise the population mean, population variance, and so on-the-list. Imagine measuring each person in a town to realize the mean. This is a daunting if not an impossible task. Thus, most of the time, we use estimates.

There are two broad methods of parameter estimation:

- Point Estimation
- Interval Estimation

2. Hypothesis Testing

Hypothesis testing is used to make decisions or draw conclusions about a population based on sample data. It involves formulating a hypothesis about the population parameter, collecting sample data, and then using statistical methods to determine whether the data provide enough evidence to reject or fail to reject the hypothesis.

Statistical Inference Methods

There are various methods of statistical inference, some of these methods are:

- Parametric Methods
- Non-parametric Methods
- Bayesian Methods

1. Parametric Methods

In this scenario, the parametric statistical methods will assume that the data is drawn from a population characterized by a probability distribution. It is mainly believed that they follow a normal distribution thus can allow one to make guesses about the populace in question . For example, the t-tests and ANOVA are parametric tests that give accurate results with the assumption that the data ought to be

Example: A psychologist may ask himself if there is a measurable difference, on average, between the IQ scores of women and men. To test his theory, he draws samples from each group and assumes they are both normally distributed. He can opt for a parametric test such as t-test and assess if the mean disparity is statistically significant.

2. Non-Parametric Methods

These are less assumptive and more flexible analysis methods when dealing with data out of normal distribution. They are also used to conduct data analysis when one is uncertain about meeting the assumption for parametric methods and when one have less or inadequate data . Some of the non-parametric tests include Wilcoxon signed-rank test and Kruskal-Wallis test among others.

Example: A biologist has collected data on plant health in an ordinal variable but since it is only a small sample and normal assumption is not met, the biologist can use Kruskal-Wallis testing.

3. Bayesian Methods

Bayesian statistics is distinct from conventional methods in that it includes prior knowledge and beliefs. It determines the various potential probabilities of a hypothesis being genuine in the light of current and previous knowledge. Thus, it allows updating the likelihood of beliefs with new data.

Example: consider a situation where a doctor is investigating a new treatment and has the prior belief about the success rate of the treatment. Upon conducting a new clinical trial, the doctor uses Bayesian method to update his “prior belief” with the data from the new trials to estimate the true success rate of the treatment.

Statistical Inference Techniques

Some of the common techniques for statistical inference are:

- Hypothesis Testing
- Confidence Intervals
- Regression Analysis

1. Hypothesis Testing

One of the central parts of statistical analysis is hypothesis testing which assumes an inference or withstand any conclusions concerning the element from the sample data. Hypothesis testing may be defined as a structured technique that includes formulating two opposing hypotheses, an alpha level, test statistic computation, and a decision based on the obtained outcomes. Two types of hypotheses can be distinguished: a null hypothesis to signify no significant difference and an alternative hypothesis H1 or Ha to express a significant effect or difference.

Example: If a car manufacturing company makes a claim that their new car model gives a mileage of not less than 25miles/gallon. Then an independent agency collects data for a sample of these cars and performs a hypothesis test. The null hypothesis would be that the car does give a mileage of not less than 25miles/gallon and they would test against the alternative hypothesis that it doesn't. The sample data would then be used to either fail to reject or reject the null hypothesis.

2. Confidence Intervals (CI)

Another statistical concept that involves confidence intervals is determining a range of possible values where the population parameter can be, given a certain confidence percentage – usually 95%. In simpler terms, CI's provide an estimate of the population value and the level of uncertainty that comes with it.

Example: A study on health records could show that 95% CI for average blood pressure is 120-130 . In other words, there is a 95% chance that the average blood pressure of all population is in the values between 120 and 130.

3. Regression Analysis

Multiple regression refers to the relationship between more than two variables. Linear regression, at its most basic level, examines how a dependent variable Y varies with an independent variable X. The regression equation, $Y = a + bX + e$, $a + bX + e$, which is the best fit line through the data points quantifies this variation.

Example: Consider a situation in which one is curious about one's advertisement on sales and is presented with it. Ultimately, it may influence questionnaire allocation as well as lead staff to feel disgruntled or upset and dissatisfied. In several regression conditions, regression analysis allows for the quantification of these two effects as well. Specifically, Y is the predicted outcome factor while X1, X2, and X3 are the observed variables used to anticipate it.

PREDICTION ERROR

In statistics, prediction error refers to the difference between the predicted values made by some model and the actual values.

Prediction error is often used in two settings:

1. Linear regression:

Used to predict the value of some continuous response variable. We typically measure the prediction error of a linear regression model with a metric known as RMSE, which stands for root mean squared error.

It is calculated as:

RMSE = $\sqrt{\Sigma(\hat{y}i - yi)2 / n}$

where:

Σ is a symbol that means "sum"

$\hat{y}i$ is the predicted value for the ith observation

yi is the observed value for the ith observation

n is the sample size

2. Logistic Regression:

Used to predict the value of some binary response variable. One common way to measure the prediction error of a logistic regression model is with a metric known as the total misclassification rate.

It is calculated as:

Total misclassification rate = (# incorrect predictions / # total predictions)

The lower the value for the misclassification rate, the better the model is able to predict the outcomes of the response variable.

The following examples show how to calculate prediction error for both a linear regression model and a logistic regression model in practice.

Example 1: Calculating Prediction Error in Linear Regression

Suppose we use a regression model to predict the number of points that 10 players will score in a basketball game. The following table shows the predicted points from the model vs. the actual points the players scored:

Predicted Points ($\hat{y}_i$)	Actual points (y_i)
14	12
15	15
18	20
19	16
25	20
18	19
12	16
12	20
15	16
22	16

We would calculate the root mean squared error (RMSE) as:

RMSE = $\sqrt{\Sigma}(\hat{y}i - yi)2 / n$

$$RMSE = \sqrt{(((14-12)2+(15-15)2+(18-20)2+(19-16)2+(25-20)2+(18-19)2+(12-16)2+(12-20)2+(15-16)2+(22-16)2) / 10)}$$

RMSE = 4

The root mean squared error is 4. This tells us that the average deviation between the predicted points scored and the actual points scored is 4.

Related: What is Considered a Good RMSE Value?

Example 2: Calculating Prediction Error in Logistic Regression

Suppose we use a logistic regression model to predict whether or not 10 college basketball players will get drafted into the NBA.

The following table shows the predicted outcome for each player vs. the actual outcome (1 = Drafted, 0 = Not Drafted):

Prediction	Actual
1	0
1	1
0	0
1	1
1	1
0	0
0	1
1	1
1	0
0	1

We would calculate the total misclassification rate as:
Total misclassification rate = (# incorrect predictions / # total predictions)
Total misclassification rate = 4/10
Total misclassification rate = 40%
The total misclassification rate is 40%.
