1. Populations and Samples:
    - A population refers to the entire group or collection of individuals, objects, or events of interest to a researcher.
    - A sample is a subset of the population selected for study.
    - The selection of a representative sample is crucial for generalizing the findings from the sample to the entire population.
    - Different sampling techniques include simple random sampling, stratified sampling, cluster sampling, and systematic sampling.
    - Sampling distributions help in making inferences about population parameters based on sample statistics.
2. Statistical Modelling:
    - Statistical models are mathematical representations that describe the relationships between variables.
    - Descriptive statistics summarize and describe the main features of data, while inferential statistics make inferences and predictions about populations based on sample data.
    - Types of statistical models include linear regression, logistic regression, time series models, and ANOVA models.
    - Assumptions and limitations of statistical models vary depending on the specific model and analysis technique.
3. Probability Distributions:
    - Probability distributions describe the likelihood of different outcomes in a random experiment or event.
    - The normal distribution is one of the most commonly encountered probability distributions and is characterized by a bell-shaped curve.
    - Other important distributions include the binomial distribution (for discrete outcomes), Poisson distribution (for rare events), and exponential distribution (for continuous random variables).
    - Probability distributions have parameters that describe their shape, such as mean, variance, skewness, and kurtosis.
4. Fitting a Model:
    - Model fitting involves estimating the parameters of a statistical model using observed data.
    - Maximum likelihood estimation (MLE) is a common technique for fitting models, where the parameters are chosen to maximize the likelihood of the observed data.
    - Least squares is another widely used method, particularly for linear regression models, where the objective is to minimize the sum of squared differences between the observed and predicted values.
    - Goodness of fit measures assess how well the model fits the data, such as R-squared (proportion of variance explained), p-values (significance of predictors), and information criteria (e.g., AIC, BIC).
5. Statistical Methods for Evaluation:
    - Hypothesis testing is used to make inferences about population parameters based on sample data.

- Significance testing involves comparing the observed data with an assumed null hypothesis to determine if the observed results are statistically significant.
- Type I error occurs when a true null hypothesis is rejected, while Type II error occurs when a false null hypothesis is not rejected.
- Confidence intervals provide a range of values within which the true population parameter is likely to fall.
- Different statistical methods have different assumptions and properties, such as parametric tests (e.g., t-test, ANOVA) assuming specific distributional forms, and non-parametric tests (e.g., Mann-Whitney U-test, Wilcoxon signed-rank test) making fewer assumptions.

6. Exploratory Data Analysis:
    - Exploratory Data Analysis (EDA) is the initial step in data analysis to understand and summarize the main characteristics of the data.
    - EDA involves calculating summary statistics (mean, median, mode, range, etc.) and visualizing data using graphs and plots.
    - Graphical techniques include scatter plots, box plots, histograms, and bar plots.
    - EDA helps identify patterns, trends, outliers, missing data, and potential relationships between variables.
    - Data preprocessing steps in EDA include data cleaning (handling missing values, outliers), data transformation (e.g., log transformation), and feature engineering (creating new variables or aggregating existing ones).
7. Getting Started with R:
    - R is a programming language and software environment for statistical computing and graphics.
    - To get started with R, you need to download and install the R programming language from the official R website (https://www.r-project.org/) and an integrated development environment (IDE) like RStudio (https://www.rstudio.com/).
    - R uses a command-line interface where you can execute code line by line or write scripts.
    - R has various data structures, including vectors (single-dimensional arrays), matrices (multi-dimensional arrays), data frames (tabular data), and lists (collections of objects).
8. Manipulating and Processing Data in R:
    - R provides functions and packages for importing and exporting data in different formats, such as CSV, Excel, and databases.
    - Data manipulation techniques in R include subsetting (selecting specific rows or columns), merging (combining datasets based on common variables), and reshaping (converting data between wide and long formats).
    - Handling missing values in R involves identifying missing data, imputing missing values, or excluding missing cases based on the analysis requirements.
    - Transforming variables in R involves scaling, standardizing, or applying mathematical operations to variables. Creating new variables can be done by combining existing variables or extracting information from text or dates.
9. Working with Functions in R:

- Functions are reusable blocks of code that perform specific tasks in R.
- R provides built-in functions for various purposes, such as mathematical calculations, data manipulation, and statistical analysis.
- You can create user-defined functions in R using the `function()` keyword, specifying input arguments and return values.
- Functions can be customized and extended to suit specific requirements by using conditional statements, loops, and other programming constructs.

10. Working with Descriptive Statistics:

- Descriptive statistics summarize and describe the main features of a dataset.
- Measures of central tendency include the mean (average), median (middle value), and mode (most frequent value).
- Measures of dispersion quantify the spread or variability of the data, such as variance, standard deviation, and range.
- Skewness measures the asymmetry of the data distribution, while kurtosis measures the peakedness or flatness of the distribution.
- Descriptive statistics help in understanding the data, identifying outliers, and comparing different groups or variables.

11. Working with Graph Plot in R:

- R provides several packages and functions for data visualization, including base graphics and popular packages like ggplot2.
- Graphical plots in R include scatter plots, line plots, bar plots, histograms, box plots, and pie charts.
- Customizing plots involves adding labels, titles, legends, colors, and different plot aesthetics.
- Multiple plots can be created in R using functions like `par()` or by combining individual plots using functions like `grid.arrange()` or facetting in ggplot2.

Statistical Modeling

INTRODUCTION

When presented with a set of data for any purpose, it’s important to interpret it correctly, draw the right conclusions, and make accurate approximations based on the information at hand. Statistics offers an organized and mathematical approach to ensure this in the form of statistical models.

In this article let us look at:

1. WHAT IS STATISTICAL MODELING?

An introduction to statistical modeling is pivotal for any data analyst to make sense of the data and make scientific predictions. In its essence, statistical modeling is a process using statistical models to analyze a set of data. Statistical models are mathematical representations of the observed data.

Statistical modeling methods are a powerful tool in understanding the consolidated data and making generalized predictions using this data. A statistical model could be in the form of a mathematical equation or a visual representation of the information.

2. TECHNIQUES IN STATISTICAL MODELING

There are several statistical modeling techniques used during data exploration. Here are some of the common techniques:

A) Linear Regression

Linear regression uses a linear equation to model the relationship between two variables, where one variable is dependent and the other is independent. If one independent variable is utilized to predict a dependent variable, it is called simple linear regression. If more than one independent variable is used to predict a dependent variable, it's called a multiple linear regression.

B) Classification

Classifications groups the data into different categories to allow for a more accurate prediction and analysis. This technique can enable effective analysis of very large data sets. There are two major techniques under classification:

- **Logistic Regression**

When the dependent variable is binary, the logistic regression technique is used to model and predict the relationship between the binary variable and one or more independent variables.

- **Discriminative Analysis**

Here, two or more groups are known as prior and new observations are grouped into known clusters based on the measured features. The distribution of the predictor variable X is modeled separately into each of the response classes, Bayes' theorem is then used to calculate the probability of each response class, based on the value of X.

C) Resampling

In this technique, repeated samples are drawn from the original set of data, creating a unique sampling distribution based on actual data. It uses experimental methods as opposed to analytical methods to create a unique sampling distribution. Since the samples drawn are unbiased, the estimates obtained are also unbiased.

Knowledge of two main concepts are essential to understand the concept of resampling in its entirety:

- **Bootstrapping**

This takes into account the data samples that weren't selected in the initial sample as a replacement. The process is repeated several times and the average score is calculated for the estimation of the model performance.

- **Cross-Validation**

The training data is divided into k number of parts. Here, k – 1 parts are considered training sets, and the one remaining set is used as the test set. This is repeated k number of times and the average of the k scores are calculated as the performance estimation.

D) Non-linear Models

Here the data under observation is modeled using a non-linear combination of model parameters and this is dependent on one or more independent variables. The data is then fitted using a method of successive approximations.

E) Tree-Based Methods

In a tree-based method, the predictor space is segmented into different simple regions. The set of splitting rules can be summarized in a tree, giving it the name decision-tree method. This can be used for both, regression and classification problems. Bagging, boosting, and random forest algorithm are some of the approaches used in this method.

F) Unsupervised Learning

Unsupervised learning relies on the algorithm to identify a pattern in the data. Here the categories of data are not known. For example, in clustering, closely related items are grouped, making it a method of unsupervised learning.

G) Time Series

This forecasting model can be used to predict future values based on historical values. It is used to identify the phenomenon represented by the data and then integrated with other data to draw predictions for the future.

H) Neural Networks

Modeled loosely on the human brain, these are algorithms designed to identify patterns in the data. Neural networks have non-linear elements that process information, called neurons. These are arranged in layers and normally executed in parallel. Neural networks are being increasingly used to make predictions and classifications as they have minimal demands on assumptions and model structure and can approximate a wide range of models.

3. TYPES OF STATISTICAL MODELS

The different types of statistical models are essentially the statistical methods used for computation. These are the mathematical equations and visual representations that make statistical modeling possible. Some of them are:

- Linear regression
- Logistic regression
- Cluster analysis
- Factor analysis
- Analysis of variation (ANOVA)
- Chi-squared test
- Correlation
- Decision trees
- Time series
- Experimental design
- Bayesian theory – Naïve Bayes classifier
- Pearson's r
- Sampling
- Association rules
- Matrix operations
- K-nearest neighbor algorithm (k-NN)

Statistical Modeling in Pharma, R, and Excel

Statistical modeling holds an important place in all types of data analysis, making it relevant to various fields of science and industry. This especially holds in the data analytics field, where analysts rely heavily on statistical methods and techniques to interpret and draw conclusions from any given dataset.

- **Statistical modelling in pharmaceutical research and development**

Statistical models are being introduced into the pharmaceutical industry to determine the efficacy of drugs for particular individuals, ensuring that individuals are given the right drugs for optimal response. Statistical techniques are used to filter biomarkers from the data, using which models are developed to predict the groups in which the drugs are most effective.

- **Statistical modelling in R**

Owing to the extensive usage of statistical modeling in data science, convenient tools embedded within the R programming language. R allows analysts to run various statistical models and is built specifically for statistical analysis and data mining. It can also enable the analyst to create software and applications that allow for reliable statistical analysis. Its graphical interface is also beneficial for data clustering, time-series, lineal modeling, etc.

- **Statistical modelling in Excel**

Excel can be used conveniently for statistical analysis of basic data. It may not be ideal for huge sets of data, where R and Python work seamlessly. Microsoft Excel provides several add-in tools under the Data tab. Enabling the Data Analysis tool on Excel opens a wide range of convenient statistical analysis options, including descriptive analysis, ANOVA, moving average, regression, and sampling.

CONCLUSION

It is safe to say that statistical modeling is an essential part of data analysis and is used across industries. Statistical models and techniques can present large datasets as mathematical representations, enabling approximations and accurate predictions.

What is a Statistical Model?

“Modeling is an art, as well as a science and, is directed toward finding a good approximating model … as the basis for statistical inference” – Burnham & Anderson

A statistical model is a **type of mathematical model** that comprises of the **assumptions** undertaken to describe the data generation process.

Let us focus on the two highlighted terms above:

1. Type of mathematical model? Statistical model is non-deterministic unlike other mathematical models where variables have specific values. Variables in statistical models are stochastic i.e. they have probability distributions.

2. Assumptions? But how do those assumptions help us understand the properties or characteristics of the true data? Simply put, these assumptions make it easy to calculate the probability of an event.

Quoting an example to better understand the role of statistical assumptions in data modeling:

Why do we need Statistical Modeling?

The statistical model plays a fundamental role in carrying out statistical inference which helps in making propositions about the unknown properties and characteristics of the population as below:

1) Estimation:

It is the central idea behind Machine Learning i.e. finding out the number which can estimate the parameters of distribution.

Note that the estimator is a random variable in itself, whereas an estimate is a single number which gives us an idea of the distribution of the data generation process. For example, the mean and sigma of Gaussian distribution

***2)** Confidence Interval:*

It gives an error bar around the single estimate number i.e. a range of values to signify the confidence in the estimate arrived on the basis of a number of samples. For example, estimate A is calculated from 100 samples and has a wider confidence interval, whereas estimate B is calculated from 10000 samples and thus has a narrower confidence interval

3) Hypothesis Testing

It is a statement of finding statistical evidence. Let's further understand the need to perform statistical modeling with the help of an example below.

We have a discrete random variable with 8 (9-1) parameters to learn i.e., probability of 0,1,2.. research papers. As the number of parameters to be estimated increase, so is the need to have those many observations, but this is not the purpose of data modeling.

So, we can reduce the number of unknowns from 8 parameters to only 1 parameter lambda, simply by assuming that the data is following Poisson distribution.

Our assumption that the data follows Poisson distribution might be a simplification as compared to the real data generation process, but it is a good approximation.

Types of modeling assumptions:

Now that we understand the significance of statistical modeling, **let's understand the types of modeling assumptions:**

1) **Parametric:** It assumes a finite set of parameters which capture everything about the data. If we know the parameter θ which very well embodies the data generation process, then predictions (x) are independent of the observed data (D)

2) **Non-parametric:** It assumes that no finite set of parameters can define the data distribution. The complexity of the model is unbounded and grows with the amount of data

3) **Semi-parametric:** It's a hybrid model whose assumptions lies between parametric and non-parametric approaches. It consists of two components – structural (parametric) and random variation (non-parametric). Cox proportional hazard model is a popular example of semi-parametric assumptions.

Definition of a statistical model: (S,P)

S: Assume that we have a collection of N i.i.d copies such as X1, X2, X3…Xn through a statistical experiment (it is the process of generating or collecting data). All these **random variables are measurable over some sample space which is denoted by S**.

P: It is the **set of probability distributions on S** that contains the distribution which is an approximate representation of our actual distribution.

Let's internalize the concept of **sample space** before understanding how a statistical model for these distributions could be represented.

1) Bernoulli : {0,1}

2) Gaussian : (-∞, +∞)

So now we have seen a few examples of sample space of some of the distribution's family, now let's see how a statistical model is defined:

1) Bernoulli : ({0,1},(Ber(p))p∈(0,1))

2) Gaussian: ((-∞, +∞),(N(μ,0.3))μ∈R)

Model specification consists of selecting an appropriate functional form for the model. For example, given “personal income” (y) together with “years of schooling” (s) and “on-the-job experience” (x), we might specify a functional relationship y=f(s,x)} as follows:

$$\ln y = \ln y_0 + \rho s + \beta_1 x + \beta_2 x^2 + \varepsilon$$

Model Misspecification: Has it ever happened with you that the model is converging properly on simulated data, but the moment real data comes, its robustness degrades, and it is no more converging? Well, this could typically happen if the model you developed does not match the data which is generally known as Model Misspecification. It could be because the class of distribution assumed for modeling does not contain the unknown probability distribution p from where the sample is drawn i.e. the true data generation process.