# STATISTICS-WORKSHEET 1

## Q1 to Q9

1. a) True
2. a) Central Limit Theorem
3. c) Modeling contingency tables
4. d) All of the mentioned
5. c) Poisson
6. b) False
7. b) Hypothesis
8. a) 0
9. d) none of the mentioned

## Q10 to Q15

**10 Ans)** Normal distribution is also known as Gaussain distribution. It is the bell- shaped frequency distribution curve of continuous random variable. In these values are equally distributed on the left and right side of the central tendency. Thus, a bell- shaped curve is formed.

It is one of the most important probability distribution in statistics because it accurately describes the distribution of values for many natural phenomena.

Normal distribution have the following features:

- Symmetric bell shape
- Mean and median are equal; both located at the centre of the distribution.
- ~68% of the data falls within 1 standard deviation off the mean
- ~95% of the data falls within 2 standard deviations of the mean
- ~99.7% of the data falls within 3 standard deviations of the mean

**11 Ans**) It is important to handle the missing data value approximately

- Many machine learning algorithms fails if the dataset contains missing values. However algorithms like K-nearest and need base support data with missing values.
- You may end up building a biased missing learning model which will lead to incorrect result if the missing value are not handled properly.

- For the kind of missing data we have we need to choose the appropriate imputation techniques to handle a data.
  The most commonly used imputation are
  1. Mean Imputation
  2. Knn imputer - It will try to find the relation with other columns and impute the data according the relation with other columns.
  3. Iterative Imputer- This method treat other columns (which doesnot have nulls as feature and train on them and treat Null column as label. Finally it will predict the NaN data and impute. Its just like regression problem. Here Null column is label.

**12 Ans**) A/B testing, also known as split testing, refers to a randomized experimentation process wherein two or more versions of a variable (web page, page element, etc.) are shown to different segments of website visitors at the same time to determine which version leaves the maximum impact and drives business metrics.

Essentially, A/B testing eliminates all the guesswork out of website optimization and enables experience optimizers to make data-backed decisions. In A/B testing, A refers to 'control' or the original testing variable. Whereas B refers to 'variation' or a new version of the original testing variable.

A/B testing is one of the components of the overarching process of Conversion Rate Optimization (CRO), using which you can gather both qualitative and quantitative user insights. You can further use this collected data to understand user behaviour, engagement rate, pain points, and even satisfaction with website features, including new features, revamped page sections, etc. If you're not A/B testing your website, you're surely losing out on a lot of potential business revenue.

**13 Ans**) It is acceptable when the missing value proportion is not large enough. But, when the missing values are large enough and you impute them with the mean, the standard errors will be lesser than what they actually would have been.

The missing value is replaced for the mean of all data formed within a specific cell or class. This technique isn't a good idea because the mean is sensitive to data noise like outliers.

**14 Ans)** Linear regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable? (2) Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable? These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = c + b*x, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

Naming the Variables. There are many names for a regression's dependent variable. It may be called an outcome variable, criterion variable, endogenous variable, or regressand. The independent variables can be called exogenous variables, predictor variables, or regressors.

Three major uses for regression analysis are (1) determining the strength of predictors, (2) forecasting an effect, and (3) trend forecasting.

**15 Ans)** The various branches of statistics are descriptive statistics and inferential statistics both of these are used in scientific data analysis.

**Descriptive Statistics**: The first aspect of statistic is descriptive statistics, which deals with the presentation and collection of data. It is not so simple as it appears and the statisticians must be aware of how to design and experiment, select the appropriate focus group, and prevent biases that are all too easy to introduce into the experiment.

Generally, descriptive statistics can be categorised into

- Measures of central tendency
- Measure of variability

To understand both measure of tendency and measure of variability easily use graphs, tables, and general discussions.

**Inferential Statistics**: Inferential statistics are statistical technique that allows statisticians to utilise data from a sample to conclude, predict the behaviour of a given population, and make judgment or decisions.

Using descriptive statistics, inferential statistics frequently talk in terms of probability. Furthermore, as statistician uses these techniques mainly for data analysis, writing and drawing conclusion from the limited data. This is accomplished by taking samples and determine their reliability.

Most future predictions and generalization based on a population study of a smaller specimen are covered by inference statistics. Furthermore, the majority of social science experiments involve the investigation of a small sample population and that helps in determining community behaviour.

The researchers can bring the study related conclusions buy designing a practical experiment. When drawing conclusions, it is important to avoid drawing incorrect and biased conclusions.

And there are some of the different types of inferential statistics which includes the following which are shown below

- Regression analysis
- Analysis of variance (ANOVA)
- Analysis of covariance (ANCOVA)
- Statistical significance (t-test)
- Correlation analysis