# STATISTICS WORKSHEET-1

**Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.**

1. Bernoulli random variables take (only) the values 1 and 0.
   a) True
   b) False

   Ans: a)True

2. Which of the following theorem states that the distribution of averages of aid variables, properly normalized, becomes that of a standard normal as the sample size increases?
   a) Central Limit Theorem
   b) Central Mean Theorem
   c) Centroid Limit Theorem
   d) All of the mentioned

   Ans: a) Central Limit Theorem

3. Which of the following is incorrect with respect to use of Poisson distribution?
   a) Modeling event/time data
   b) Modeling bounded count data
   c) Modeling contingency tables
   d) All of the mentioned

   Ans: a) Modelling bounded count data

4. Point out the correct statement.
   a) The exponent of a normally distributed random variables follows what is called the log- normal distribution
   b) Sums of normally distributed random variables are again normally distributed even if the variables are dependent
   c) The square of a standard normal random variable follows what is called chi-squared distribution
   d) All of the mentioned

   Ans:d) All of the mentioned



5. _____random variables are used to model rates.
   a) Empirical
   b) Binomial
   c) Poisson
   d) All of the mentioned

   Ans:c) Poisson

6. 10. Usually replacing the standard error by its estimated value does change the CLT.
   a) True
   b) False

   Ans:b) False

7. 1. Which of the following testing is concerned with making decisions using data?
   a) Probability
   b) Hypothesis
   c) Causal
   d) None of the mentioned

   Ans: b) Hypothesis

8. Normalized data are centered at_____and have units equal to standard deviations of the original data.
   a) 0
   b) 5
   c) 1
   d) 10

   Ans: a) 0

9. Which of the following statement is incorrect with respect to outliers?
   a) Outliers can have varying degrees of influence
   b) Outliers can be the result of spurious or real processes
   c) Outliers cannot conform to the regression relationship
   d) None of the mentioned

   Ans: c) Outliers cannot conform to the regression relationship



**Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.**

## 10.What do you understand by the term Normal Distribution?

The normal distribution, also known as the Gaussian of standard normal distribution, & the probability distribution is the most important probability statistics for independent, random variables. Most people recognize its familiar bell sloped curve, statistical report.

The normal distribution is a Continuous probability distribution that is symmetrical around its mean, most of observations cluster around the central the peak, and the probabilities for values further away from the mean taper off equally in both directions, Extreme values in both tails of the distribution are similarly unlikely, while the normal distribution is symmetrical, not all symmetrical distributions are normal.

As with any probability distribution the normal distribution describes how the values of a variable are distributed. It is the most important probability distribution in statistics because it accurately describes the distribution of values for many natural phenomena, Characteristics that are the sum of many normal distributions independent processes frequently follow normal distributions, Ex: heights, blood pressure.

## 11.How do you handle missing data? What imputation techniques do you recommend?

When dealing with missing data, data scientists can use two primary methods to solve the error: One is imputation or the removal of data (deletion)

In imputation method, if there is small number of missing observations, data scientist can calculate the mean or median of the existing observations.

There are many techniques for imputation of missing data's

Before deciding which approach to employ, data scientists must understand why the data is missing.

The following are the reasons for missing data's according to this. Data scientists decide the imputation techniques.

- Missing at Random
- Missing Completely at Random
- Missing Not at Random

**Imputation**: Instead of deletion, data scientists have multiple solutions to impute the value of missing data. Depending why the data are missing, imputation methods can deliver reasonably reliable results. These are examples of single imputation methods for replacing missing data.
The following are the imputation techniques:

- Mean, Median and Mode
- Time-Series Specific Methods
- Last Observation Carried Forward (LOCF) & Next Observation Carried Backward (NOCB)
- Linear Interpolation
- Seasonal Adjustment with Linear Interpolation
- Multiple Imputation

**Deletion**: There are two primary methods for deleting data when dealing with missing data: list wise and dropping variables.
The following are the Deletion techniques:

- List wise Deletion
- Pairwise Deletion
- Dropping Variables

## 12.What is A/B testing?

**A**/B testing is one of the most popular controlled experiments used to optimize web marketing strategies. It allows decision makers to choose the best design for a website by looking at the analytics results obtained with two possible alternatives A and B.

A/B test consist of randomized experiment with two variables, A and B, It includes application of statistical hypothesis testing or two hypothesis testing as used in the field of statistics, A/B testing

is a way to compare two versions of a single variable, typically by testing a subjects response to variant A against Variant B, and delimiting which of the two variants is more effective.

A/B testing, also known as split testing, is a marketing experiment wherein you split your audience to test a number of variations of a campaign and determine which performs better. In other words, you can show version A of a piece of marketing content to one half of your audience, and version B to another.

A/B testing let you change variables, such as your ad creative, audience, or placement to determine which strategy performs best and improve future campaigns. For example, you Might hypothesize that a custom audience strategy will outperform an interest-based Audience strategy for your business.

From the above, we'll see how different statistical methods can be used to make A/B testing successful.

To understand what A/B testing is about, let's consider two alternative designs: A and B. Visitors of a website are randomly served with one of the two. Then, data about their activity is collected by web analytics. Given this data, one can apply statistical tests to determine whether one of the two designs has better efficacy.

Now, different kinds of metrics can be used to measure a website efficacy. With discrete metrics, also called binomial metrics, only the two values 0 and 1 are possible. The following are examples of popular discrete metrics.

## 13.Is mean imputation of missing data acceptable practice?

The process of replacing null values in a data collection with the data's mean is known as mean imputation.

Mean imputation is typically considered terrible practice since it ignores feature correlation. Second, mean imputation decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate and the confidence interval is narrower.

- Bad practice in general
- If just estimating means: mean imputation preserves the mean of the observed data
- Leads to an underestimate of the standard deviation
- Distorts relationships between variables by "pulling" estimates of the correlation toward zero

## 14.What is linear regression in statistics?

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

- The independent variable is the cause. Its value is independent of other variables in your study.
- The dependent variable is the effect. Its value depends on changes in the independent variable.

## 15.What are the various branches of statistics?

Statistics is the study and manipulation of data, including ways to gather, review, analyse, and draw conclusions from data.

- The two major areas of statistics are descriptive and inferential statistics.
- Statistics can be used to make better-informed business and investing decisions.

DESCRIPTIVE STATISTICS

INFERENTIAL STATISTICS

Descriptive statistics

In this type of statistics, the data is summarised through the given observations. The summarisation is one from a sample of population using parameters such as the mean Standard deviation

Descriptive statistics is a way to organise, represent and describe a collection of data using tables graph and summary measures.
Descriptive statistics are also categorised into four different categories

* Measure of frequency

* Measure of dispersion

* Measure of Central tendency & Measure of position.

Inferential statistics

This type of statistics is used to interpret the meaning of Descriptive statistics. That mean once The data has been collected analysed and summarised there we use these stats to describe. The meaning of the collected data, or we can say it is used to draw conclusions from the data that depends on random Variations such as observational errors: sampling variations etc.

Inferential statistics is a method that allows us to use information collected for a simple to make decisions, Predictions to inferences from a population. It Jante us permission to give statements that goes beyond the available data or information For example, driving estimates from hypothetical research.