# 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 iid 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-b)Modeling 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-c) The square of a standard normal random variable follows what is called chi-squareddistribution**

5. _______random variables are used to model rates.
   a) Empirical
   b) Binomial
   c) Poisson
   d) All of the mentioned
   **Ans-c) Poisson**
6. Usually replacing the standard error by its estimated value does change the CLT.
   a) True
   b) False
   **Ans-b) False**
7. 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**

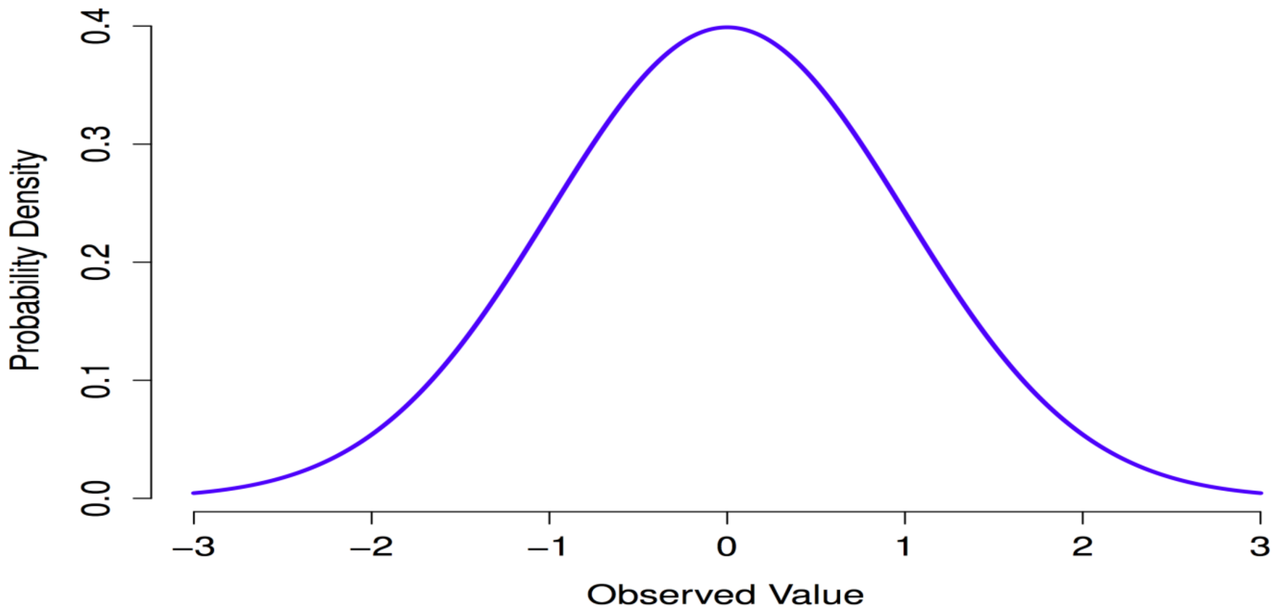
**Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.**

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

**Ans->**

## Normal Distribution

In probability theory and statistics, the **Normal Distribution**, also called the **Gaussian Distribution**,is the most significant continuous probability distribution. Sometimes it is also called a bell curve. A large number of random variables are either nearly or exactly represented by the normal distribution, in every physical science and economics. Furthermore, it can be used to approximate other probability distributions, therefore supporting the usage of the word ‘normal ‘as in about the one, mostly used.



### Normal Distribution Definition

The Normal Distribution is defined by the probability density function for a continuous random variable in a system. Let us say, f(x) is the probability density function and X is the random variable. Hence, it defines a function which is integrated between the range or interval (x to x + dx), giving the probability of random variable X, by considering the values between x and x+dx.

$f(x) \geq 0 \ \forall \ x \in (-\infty,+\infty)$

And $\int_{-\infty}^{+\infty} f(x) = 1$

### Normal Distribution Formula

The probability density function of normal or gaussian distribution is given by;

$$f(x,\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Where,

- x is the variable
- μ is the mean
- σ is the standard deviation

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

**Ans->** Handling missing data is an important part of data analysis, as missing data can introduce bias and reduce the power of statistical analyses. There are several ways to handle missing data, including:

**Complete case analysis**: This method involves excluding all observations that have missing data on any variable included in the analysis. This method is simple but can lead to biased results if missing data are not missing completely at random.

**Mean imputation**: This method involves replacing missing values with the mean value of the variable. This method is simple but can lead to biased results if there is a systematic difference between missing and non-missing values.

**Multiple imputation**: This method involves creating multiple imputed datasets, each with a different estimate of the missing values based on the observed data and some randomness. These imputed datasets are then analyzed separately, and the results are combined to produce a single estimate of the parameter of interest. This method is more complex but can produce less biased results than other methods.

**Maximum likelihood estimation**: This method involves using all available data, including incomplete data, to estimate the parameters of interest. This method requires a model for the missing data mechanism and can be computationally intensive.

The choice of imputation technique depends on the nature of the missing data and the specific research question. However, multiple imputation is generally considered to be a good approach as it can produce less biased results than other methods while accounting for the uncertainty introduced by missing data. It is important to note that imputation cannot completely eliminate the bias introduced by missing data, but it can reduce it if done appropriately.

## 12. What is A/B testing?

**Ans->** A/B testing is a statistical method used to evaluate the effectiveness of a new intervention (treatment) by comparing it to an existing intervention or a control group. A/B testing involves randomly assigning participants to one of two groups: the treatment group or the control group. The treatment group receives the new intervention, while the control group receives the existing intervention or no intervention at all.

The outcomes of the two groups are then compared to see if there is a statistically significant difference between them. If there is a significant difference, it can be inferred that the new intervention is effective. A/B testing is commonly used in marketing and website design to evaluate the effectiveness of new designs, content, or features.

A/B testing can be conducted using a variety of statistical tests, including t-tests and chi-squared tests, depending on the type of outcome variable and the research question. Careful consideration of sample size, randomization, and blinding is important to ensure the validity and reliability of the results.

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

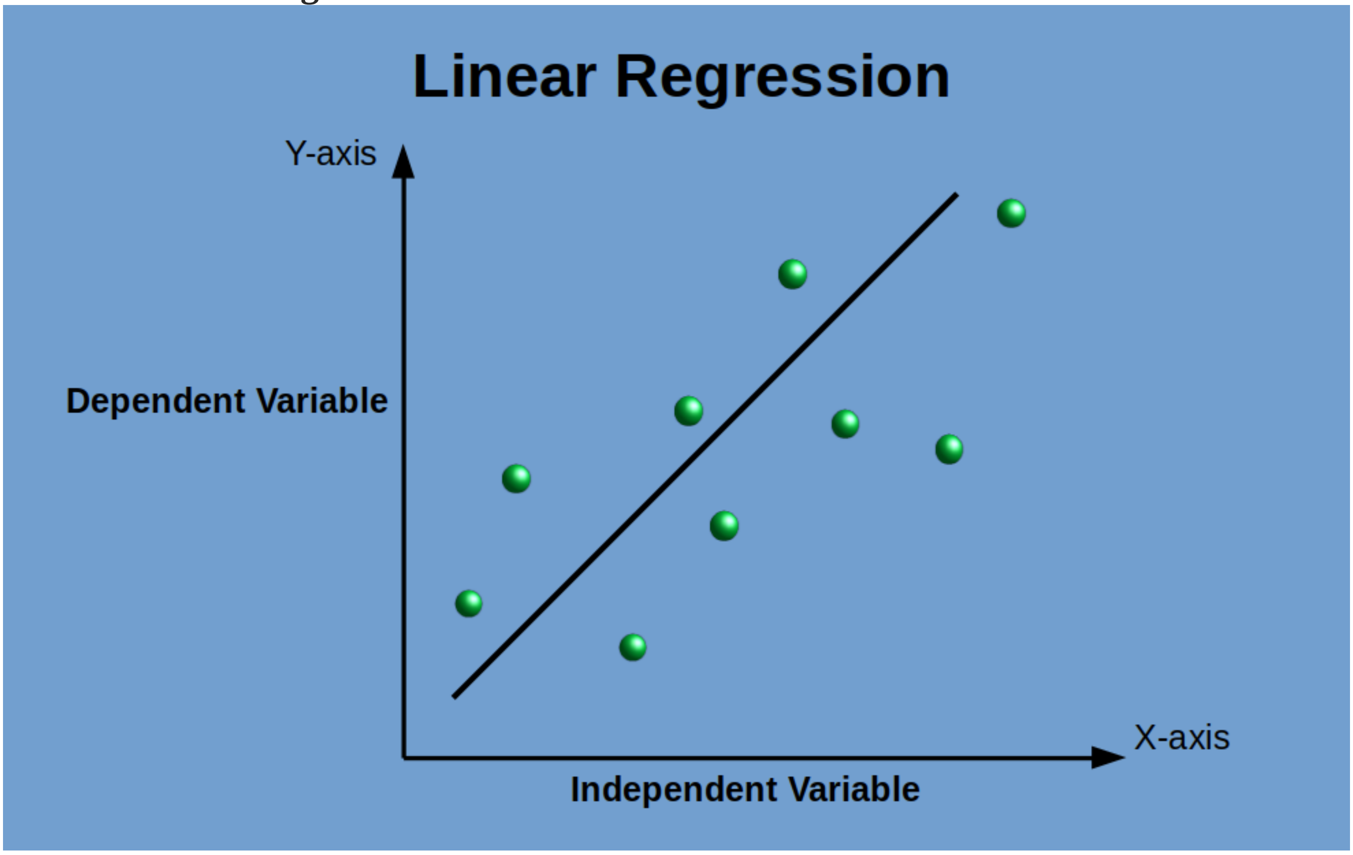
**Ans->**Mean imputation is a simple method for handling missing data, which involves replacing missing values with the mean value of the variable. While mean imputation is a common practice due to its simplicity, it has several drawbacks that can lead to biased estimates and reduce the accuracy of statistical analyses.

One of the main drawbacks of mean imputation is that it assumes that missing values are missing completely at random (MCAR), meaning that the probability of missingness is unrelated to the value of the variable or any other variables in the dataset. If the missing values are not MCAR, mean imputation can lead to biased estimates and invalid statistical inferences.

Another issue with mean imputation is that it does not take into account the uncertainty introduced by missing data. The imputed values are treated as if they are true values, which can lead to an underestimation of the standard errors and confidence intervals and an overestimation of statistical significance.

Therefore, while mean imputation can be a convenient approach, it is generally not recommended as a first choice for handling missing data. Instead, multiple imputation, maximum likelihood estimation, or other advanced imputation methods that account for the uncertainty introduced by missing data should be used.

### 14. What is linear regression in statistics?



**Ans->**Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. The goal of linear regression is to find a linear function that best describes the relationship between the variables, such that the sum of the squared differences between the predicted values and the actual values is minimized.

The simplest form of linear regression, called simple linear regression, involves one independent variable and one dependent variable. The linear function can be represented by the equation y = mx + b, where y is the dependent variable, x is the independent variable, m is the slope of the line, and b is the y-intercept.

In multiple linear regression, there are two or more independent variables, and the linear function can be represented by the equation y = b0 + b1x1 + b2x2 + ... + bnxn, where y is the dependent variable, x1, x2, ..., xn are the independent variables, and b0, b1, b2, ..., bn are the coefficients that represent the slope of each independent variable.

Linear regression can be used for a variety of purposes, such as predicting future values of the dependent variable based on the independent variables, estimating the strength and direction of the relationship between variables, and identifying which independent variables have the greatest impact on the dependent variable. Linear regression can be performed using statistical software, such as R or Python, and the quality of the model can be assessed using measures such as the coefficient of determination (R-squared) and residual plots.

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

**Ans->**Statistics is a broad field of study that encompasses many different branches, including:

**Descriptive statistics:** This branch of statistics involves summarizing and describing the main features of a dataset, such as measures of central tendency, variability, and correlation.

**Inferential statistics**: This branch of statistics involves making predictions and drawing conclusions about a larger population based on data from a smaller sample.

**Probability theory:** This branch of mathematics deals with the study of random events and their likelihood of occurrence.

**Statistical modeling**: This branch of statistics involves building mathematical models to describe and predict relationships between variables in a dataset.

**Biostatistics:** This branch of statistics is concerned with the analysis and interpretation of data in the field of health sciences.

**Econometrics:** This branch of statistics is used to analyze economic data and build models to explain and predict economic phenomena.

**Psychometrics:** This branch of statistics is used to measure and analyze psychological and behavioral data.

**Statistical computing:** This branch of statistics is focused on developing algorithms and software for analyzing and visualizing data.

**Bayesian statistics**: This branch of statistics is based on the principles of Bayes' theorem and involves updating probabilities as new data becomes available.

**Time series analysis:** This branch of statistics is used to analyze and model time-dependent data, such as stock prices or weather patterns.