STATISTICS WORKSHEET-1

Q.1] A [TRUE]

Q.2] A

Q.3] D

Q.4] C

Q.5] C

Q.6] A [TRUE]

Q.7] B

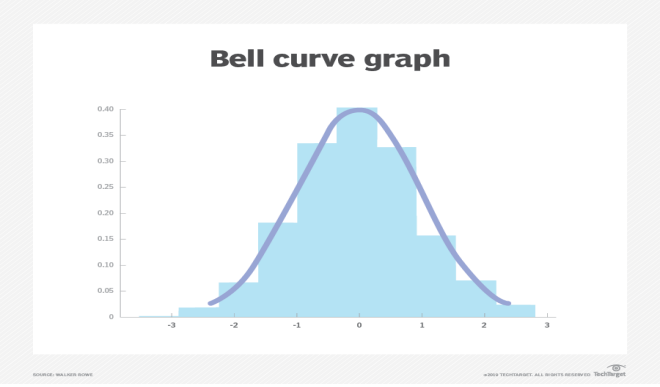
Q.8] A

Q.9] C

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

ANS. = The normal distribution, also known as the Gaussian distribution or bell curve, is a probability distribution that is symmetric about the mean and describes the distribution of many types of data. It is defined by its mean (μ) and standard deviation (σ). It is a continuous probability distribution that has a bell-shaped probability density function and is defined by two parameters, the mean (μ) and the standard deviation (σ). A normal distribution bell curve is always symmetrical about the mean. A symmetrical distribution means that a vertical dividing line drawn through the maximum/mean value will produce two mirror images on either side of the line, in which half of the left side is less than the mean and half to the right side is greater. The right side of the line is know as positive side and the left side of the line is know as negative side. In the bell curve, the peak is always in the middle, and the mean, mode and median are all the same.

The normal distribution is widely used in statistics, physics, engineering and economics for modeling data and making predictions. It is often used as a model for random errors or disturbances in measurements or other processes, and is the foundation for many statistical tests and procedures.



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

ANS. = Handling missing data can be a complex task, and the best approach will depend on the specific circumstances of the data and the research question. There are several imputation techniques that can be used to handle missing data, including:

1] Mean/mode imputation: This method replaces missing values with the mean or mode of the non-missing values in the same variable.

2] Median imputation: This method replaces missing values with the median of the non-missing values in the same variable.

3] Regression imputation: This method uses a regression model to predict missing values based on the values of other variables in the dataset.

4] Hot-Deck imputation: This method replaces missing data with a randomly selected non-missing value from the same variable.

5] Multiple imputation: This method generates multiple plausible values for each missing data point using a model that accounts for uncertainty in the imputations.

6] Cold deck imputation: A value picked deliberately from an individual with similar values on other variables.

It's important to note that imputation techniques should be chosen based on the underlying assumptions of the data and the research question. It's also important to report the amount of missing data and the method used for handling missing data in the final analysis.

Q.12 What is A/B testing?

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. A/B testing is a statistical method used to compare two versions of a product, website, or marketing campaign to determine which one performs better. The "A" and "B" versions are typically identical, except for one variation, such as a different color, layout, or messaging. The goal of A/B testing is to determine if the change in the "B" version results in a statistically significant difference in a specific metric, such as click-through rate, conversion rate, or revenue. It can be used to optimize a website, landing page, email campaign, or any other digital product or service.

The process of A/B testing generally involves the following steps:

1. Define the problem or question to be answered.

2. Create two or more versions of the product or website, with one or more variations.

3. Randomly assign users to the different versions.

4. Track the specific metric(s) of interest for each version.

5. Analyze the data to determine if there is a statistically significant difference between the versions.

It's important to note that A/B testing is a type of experiment design, so it's important to follow experimental design principles such as randomization, control and independence to ensure that the results are valid and reliable.

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

ANS. = Mean imputation is a commonly used method for handling missing data, but it may not always be an appropriate technique. The suitability of mean imputation depends on the underlying assumptions of the data and the research question. Mean imputation is appropriate if the data are missing completely at random (MCAR) and the variable has a normal distribution. In this case, the mean can be considered as a good estimator for the missing values. However, if the data is not MCAR, or the variable does not have a normal distribution, mean imputation can introduce bias and lead to inaccurate conclusions.

Mean imputation also assumes that the missing values are similar to the non-missing values, which may not always be the case. For example, if the missing values are from a different population than the non-missing values, the mean of the non-missing values may not be representative of the missing values. In general, mean imputation is a simple and easy-to-use method, but it can introduce bias and lead to inaccurate conclusions if the assumptions are not met. Therefore, it is important to consider the assumptions and limitations of the method and to be aware of the potential effects of imputation on the results. In cases where the assumptions are not met, other imputation methods such as multiple imputation may be more appropriate.

Q.14 What is linear regression in statistics?

ANS. = Linear regression is a statistical method used to model the relationship between a dependent variable (also known as the response variable) and one or more independent variables (also known as predictor variables). It is used to make predictions about the value of the dependent variable based on the values of the independent variables. The basic idea behind linear regression is that there is a linear relationship between the independent and dependent variables, which can be represented by a straight line. This line is defined by the equation of a straight line, which is given by $y = mx + b$, where y is the dependent variable, x is the independent variable, m is the slope of the line, and b is the y-intercept.

Linear regression can be used for both simple linear regression, which has one independent variable and multiple linear regression which has multiple independent variables. The goal of the linear regression is to find the best-fitting straight line through the data, which is done by finding the line that minimizes the sum of the squared differences between the predicted and actual values of the dependent variable.

Linear regression assumes that the relationship between the independent and dependent variables is linear, that there is no multicollinearity among the independent variables, that the errors are normally distributed with constant variance and that the errors are independent of the independent variables. If these assumptions are not met, the results of the linear regression may be biased or unreliable.

Q.15 What are the various branches of statistics?

ANS. = Statistics is a study of presentation, analysis, collection, interpretation and organization of data. There are two main branches of statistics:

- Inferential Statistic.

- Descriptive Statistic.

1] Inferential statistics: This branch of statistics deals with making inferences about a population based on a sample of data. It includes techniques for estimating population parameters and making statistical decisions, such as hypothesis testing and confidence intervals.

2] Descriptive statistics: This branch of statistics is concerned with the collection, organization, summarization, and presentation of data. It includes techniques for describing and summarizing data, such as measures of central tendency (mean, median, mode) and measures of dispersion (range, variance, standard deviation).