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
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
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
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
5. _____random variables are used to model rates.
   a) Empirical
   b) Binomial
   c) Poisson
   d) All of the mentioned
6. 10. Usually replacing the standard error by its estimated value does change the CLT.
   a) True
   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
8. 4. Normalized data are centered at_____and have units equal to standard deviations of the original data.
   a) 0
   b) 5
   c) 1
   d) 10
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

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

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

Answer: -

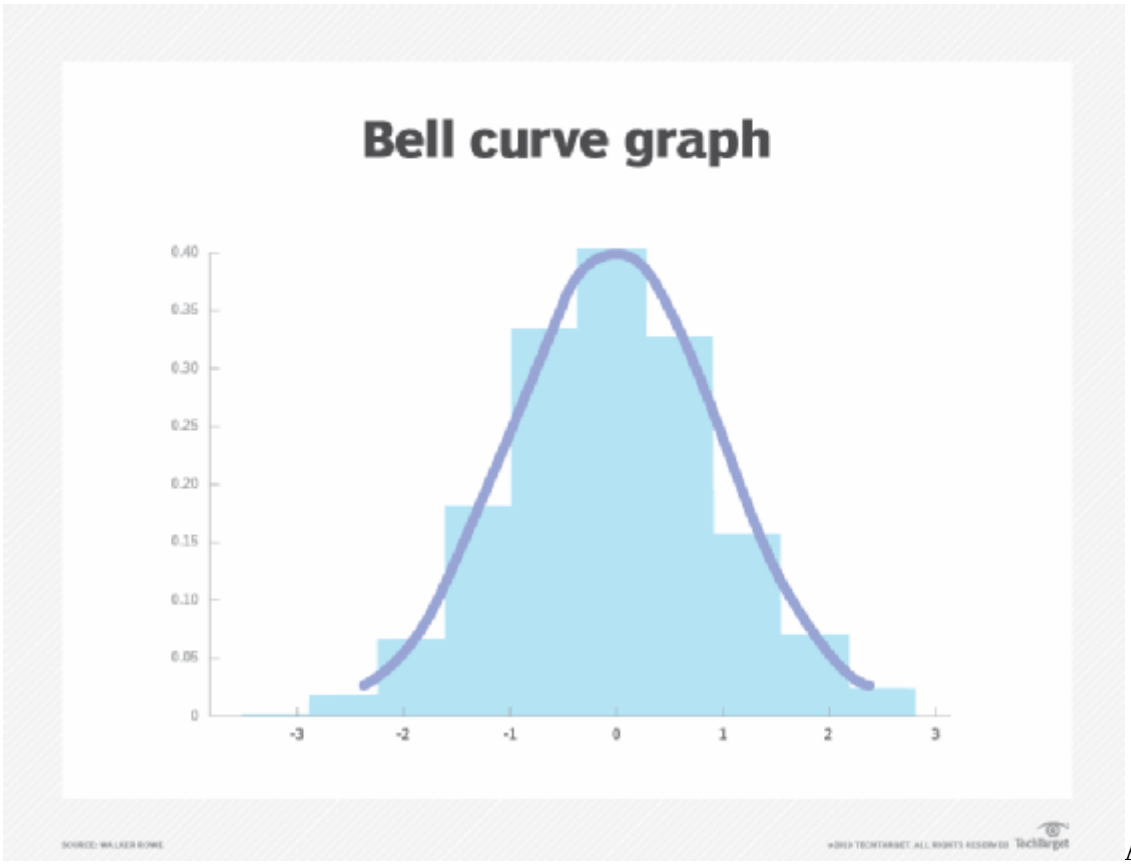
A normal distribution is a type of continuous probability distribution in which most data points cluster toward the middle of the range, while the rest taper off symmetrically toward either extreme. The middle of the range is also known as the *mean* of the distribution.

The normal distribution is also known as a *Gaussian distribution* or probability bell curve It is symmetric about the mean and indicates that values near the mean occur more frequently than the values that are farther away from the mean.

Normal distribution explained

Graphically, a normal distribution is a bell curve because of its flared shape. The precise shape can vary according to the distribution of the values within the population The population is the entire set of data points that are part of the distribution.

Regardless of its exact shape, 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 the population is less than the mean and half is greater. However, the reverse is not always true; that is, not all symmetrical distributions are normal. In the bell curve, the peak is always in the middle, and the mean, mode and median are all the same.



A normal distribution bell curve is always symmetrical about the mean.
Basic examples of normal distribution: Height and weight

Height is one simple example of values that follow a normal distribution pattern. Most people are of average height -- whatever that may be for a given population. If the heights of these people are represented in graphical format along with the heights of people who are taller and shorter than the average, the distribution will always be a normal distribution. This is because the people of average height will be clustered near the middle, while those who are taller and shorter will be farther away.

Further, these latter groups will consist of very small numbers of people. The number of people who are extremely tall or extremely short will be even smaller, so they will be the farthest away from the mean.

Similarly, weight can also follow a normal distribution if the average weight of the population under consideration is known. Like height, the weight outliers will be those who weigh more or less than the average. The bigger the deviation from the average, the farther away those data points will be on the distribution graph.

Importance of normal distribution

The normal distribution is one of the most important probability distributions for independent random variables for three main reasons.

First, normal distribution describes the distribution of values for many natural phenomena in a wide range of areas, including biology, physical science, mathematics, finance and economics. It can also represent these random variables accurately.

In addition to height and weight, normal distributions are also used to represent many other values, including the following:

- measurement error
- blood pressure
- IQ scores
- asset prices
- price action

Second, the normal distribution is important because it can be used to approximate other types of probability distribution, such as binomial, hypergeometric, inverse (or negative) hypergeometric, negative binomial and Poisson distribution.

Third, normal distribution is the key idea behind the central limit theorem, or CLT, which states that averages calculated from independent, identically distributed random variables have approximately normal distributions. This is true regardless of the type of distribution from which the variables are sampled, as long as it has finite variance.

Normal distribution formula and empirical rule

The formula for the normal distribution is expressed below.

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

The formula for the normal distribution.

Here, x is value of the variable; $f(x)$ represents the probability density function; μ *(mu)* is the mean; and σ *(sigma)* is the standard deviation.

The empirical rule for normal distributions describes where most of the data in a normal distribution will appear, and it states the following:

- 68.2% of the observations will appear within +/-1 standard deviation of the mean;
- 95.4% of the observations will fall within +/-2 standard deviations; and
- 99.7% of the observations will fall within +/-3 standard deviations.

All data points falling outside of three standard deviations (3σ) indicate rare occurrences.

Parameters of normal distribution

Since the mean, mode and median are the same in a normal distribution, there's no need to calculate them separately. These values represent the distribution's highest point, or the peak. All other values in the distribution then fall symmetrically around the mean. The width of the mean is defined by the standard deviation.

In fact, only two parameters are required to describe a normal distribution: the mean and the standard deviation.

1. The mean

The mean is the central highest value of the bell curve. All other values in the distribution either cluster around it or are at some distance away from it. Changing the mean on a graph will shift the entire curve along the x-axis, either toward the left or toward the right. However, its symmetricity will still be maintained.

2. The standard deviation

In general, standard deviation is a measure of variability in a distribution. In a bell curve, it defines the width of the distribution and shows how far away from the mean the other values fall. In addition, it represents the typical distance between the average and the observations.

Changing the standard deviation will change the distribution of values around the mean. A smaller deviation will reduce the spread -- tightening the distribution -- while a larger deviation will increase the spread and produce a wider distribution. As the distribution gets wider, it becomes more likely that values will be farther away from the mean.

Skewness and kurtosis in a normal distribution

Skewness represents a distribution's degree of symmetry. Since the normal distribution is perfectly symmetric, it has a skewness of zero. In other distributions with a skewness less than or greater than zero, the left tail (left skewness) or the right tail (right skewness) will be longer, respectively.

Kurtosis measures the thickness of each tail end of a distribution vis-à-vis the tails of a normal distribution. For a normal distribution, kurtosis is always equal to 3. In a distribution with kurtosis greater than 3, the tail data will exceed the tails of the normal distribution, resulting in a phenomenon called *fat tails*. In financial markets, fat tails describe tail risk -- the chance of a loss due to some rare event. Distributions with kurtosis less than 3 show tails that are skinnier than the tails of a normal distribution.

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

Answer: -

Handling missing data is an important task in data analysis and requires careful consideration. Here are some common techniques for handling missing data:

1. Deletion: In this approach, the rows or columns with missing data are simply removed from the dataset. This can be done using either list-wise deletion (removing entire rows with missing data) or pairwise deletion (retaining available data for analysis on a case-by-case basis). Deletion can be an option if the missing data is minimal and random. However, it can lead to a loss of information and potential bias if the missing data is not random.

2. Mean/Mode Imputation: In this technique, missing values are replaced with the mean (for numerical data) or mode (for categorical data) of the available values for that variable. Imputation with mean or mode assumes that the missing values are missing completely at random (MCAR) and does not take into account the relationships between variables. This method is simple but may introduce bias and underestimate the variability in the data.

3. Median Imputation: Similar to mean imputation, missing values are replaced with the median of the available values. Median imputation is less sensitive to extreme values compared to mean imputation and can be useful for handling skewed distributions.

4. Regression Imputation: In this technique, missing values in a variable are predicted based on the relationship with other variables using regression analysis. A regression model is built using the available data, and the model is then used to estimate the missing values. This method considers the relationships between variables but assumes that the relationships remain stable in the presence of missing data.

5. Multiple Imputation: Multiple imputation involves creating multiple plausible imputed datasets based on the observed data and the estimated relationships between variables. Statistical models are used to impute missing values, taking into account the uncertainty associated with the imputation process. Multiple imputation accounts for both within-variable and between-variable relationships and provides more accurate estimates compared to single imputation methods.

The choice of imputation technique depends on various factors, including the amount and pattern of missing data, the nature of the data, and the analysis goals. It is recommended to carefully assess the missing data mechanism and consider multiple imputation techniques, which provide more reliable and robust results compared to single imputation methods.

It is important to note that imputation does not guarantee accurate results, and it is crucial to interpret the analysis outcomes with caution and consider the potential impact of missing data on the validity of the results.

Q.12 What is A/B testing?

Answer: -

A/B testing, also known as split testing, is a statistical experiment used in marketing and web analytics to compare two or more versions of a webpage or marketing campaign and determine which one produces better results or performs better in terms of user behavior or conversion rates.

In A/B testing, two or more variants, typically referred to as A and B, are created and randomly assigned to different groups of users. The groups are exposed to the different variants, and their responses or behaviors are measured and compared to determine which variant is more effective. The goal is to identify the variant that leads to higher engagement, conversions, click-through rates, or any other desired outcome.

Here's a general process of conducting an A/B test:

1. Identify the objective: Clearly define the goal of the A/B test, such as increasing click-through rates, improving conversion rates, or enhancing user engagement.

2. Formulate a hypothesis: Develop a hypothesis about how the changes in the variants may impact the desired outcome. For example, you might hypothesize that changing the color of a call-to-action button will lead to a higher conversion rate.

3. Create variants: Develop multiple versions of the webpage, email, advertisement, or any other element that you want to test. These versions should differ based on the specific elements or features you want to compare (e.g., different headlines, layouts, colors, or content).

4. Random assignment: Randomly assign the different variants to separate groups of users or visitors. This ensures that any observed differences in behavior or outcomes can be attributed to the specific variant and not to other factors.

5. Collect data: Monitor and collect relevant data or metrics from each variant, such as click-through rates, conversion rates, bounce rates, or any other key performance indicators (KPIs). It is important to ensure proper tracking and measurement of the desired outcomes.

6. Analyze the results: Use statistical analysis to compare the performance of the different variants. Common statistical techniques include hypothesis testing, confidence intervals, and p-values. Analyzing the results helps determine if there is a statistically significant difference between the variants and which variant performs better.

7. Draw conclusions and make decisions: Based on the analysis of the results, draw conclusions about the performance of the variants and make data-driven decisions. If one variant outperforms the others, it can be implemented as the preferred version.

A/B testing allows businesses and marketers to make informed decisions about their website design, marketing campaigns, user experiences, and more. It provides valuable insights into customer behavior and preferences, allowing organizations to optimize their strategies and improve their overall performance.

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

Mean imputation of missing data is a commonly used technique due to its simplicity and ease of implementation. However, it is important to understand that mean imputation has limitations and potential drawbacks. Here are some considerations:

Advantages:
1. Simple and straightforward: Mean imputation is easy to understand and implement. It involves replacing missing values with the mean of the available data for that variable.

2. Preserves sample size: Mean imputation allows you to retain the same number of observations in your dataset, which can be important for certain analyses or modeling techniques.

Disadvantages:
1. Distorts the distribution: Mean imputation can distort the distribution of the variable by artificially inflating the number of observations with the mean value. This can lead to biased estimates and incorrect inferences.

2. Underestimates the variability: Mean imputation tends to underestimate the variability of the variable because it artificially reduces the spread of the data. This can lead to an underestimation of standard errors and confidence intervals, affecting the reliability of statistical analysis.

3. Ignores relationships with other variables: Mean imputation does not take into account the relationships between variables. It assumes that missing values are missing completely at random (MCAR), which might not be a valid assumption in many cases. This can introduce bias and affect the accuracy of subsequent analyses.

4. Masking effects: Mean imputation can mask true associations between variables. By replacing missing values with the mean, any existing relationship between the missing variable and other variables may be diminished or hidden.

5. Fails to account for uncertainty: Mean imputation does not account for the uncertainty associated with the imputed values. It assumes that the imputed values are known with certainty, which is not the case. Ignoring the uncertainty can lead to incorrect conclusions and invalid statistical inferences.

Given these limitations, mean imputation should be used with caution, and its appropriateness depends on the specific context and the nature of the missing data. It is important to consider alternative imputation methods, such as multiple imputation, which provide more reliable estimates by accounting for the uncertainty associated with the imputed values and the relationships between variables. Multiple imputation is generally recommended when dealing with missing data, especially in situations where missingness is not completely random or when a high percentage of values are missing.

Q.14 What is linear regression in statistics?

Answer: -

Linear regression is a statistical technique used to model the relationship between a dependent variable and one or more independent variables. It aims to find a linear equation that best describes the relationship between the variables, allowing for prediction or inference about the dependent variable based on the values of the independent variables.

In linear regression, the dependent variable is the variable of interest that we want to predict or explain, while the independent variables (also called predictor variables or features) are the variables used to predict or explain the dependent variable.

The basic assumption in linear regression is that there is a linear relationship between the independent variables and the dependent variable. This relationship is expressed by a linear equation of the form:

$$Y = \beta0 + \beta1*X1 + \beta2*X2 + ... + \beta n*Xn + \varepsilon$$

where:
- Y is the dependent variable.
- X1, X2, ..., Xn are the independent variables.
- β0, β1, β2, ..., βn are the regression coefficients or parameters that represent the effect of each independent variable on the dependent variable.
- ε is the error term, representing the unexplained variability in the dependent variable that is not accounted for by the independent variables.

The goal of linear regression is to estimate the regression coefficients that minimize the sum of squared differences between the observed values of the dependent variable and the predicted values from the linear equation. This is typically done using a method called ordinary least squares (OLS) estimation.

Once the regression coefficients are estimated, the linear equation can be used to make predictions or infer the relationship between the variables. The coefficients provide information about the direction and magnitude of the effect of each independent variable on the dependent variable.

Linear regression can be used for various purposes, such as predicting sales based on advertising expenditure, studying the impact of variables on an outcome, assessing the strength of relationships between variables, and identifying influential factors in a given phenomenon.

Q.15 What are the various branches of statistics?

Answer: -

Statistics, as a field of study, encompasses various branches that focus on different aspects of data analysis, inference, and interpretation. Some of the main branches of statistics include:

1. Descriptive Statistics: Descriptive statistics involves summarizing and describing data using measures such as measures of central tendency (mean, median, mode) and measures of dispersion (variance, standard deviation, range). It provides a way to organize, summarize, and present data in a meaningful and interpretable manner.

2. Inferential Statistics: Inferential statistics involves making inferences and drawing conclusions

about a population based on a sample. It includes techniques such as hypothesis testing, confidence intervals, and estimation. Inferential statistics helps us make predictions, test hypotheses, and generalize findings from a sample to a larger population.

3. Probability Theory: Probability theory is the foundation of statistical analysis. It deals with the mathematical concepts and principles used to quantify uncertainty and randomness. Probability theory enables us to understand and model the likelihood of different outcomes and events occurring.

4. Biostatistics: Biostatistics is the branch of statistics that focuses on the design and analysis of data related to biological, medical, and health sciences. It involves the application of statistical methods to study and interpret data in areas such as clinical trials, epidemiology, genetics, and public health.

5. Econometrics: Econometrics applies statistical methods to economic data to analyze and model economic relationships and phenomena. It involves the development and application of statistical techniques to estimate economic parameters, test economic theories, and forecast economic variables.

6. Social Statistics: Social statistics deals with the analysis and interpretation of data related to social phenomena and human behavior. It involves studying data on topics such as demographics, education, crime, social inequality, and survey research. Social statistics helps in understanding social trends, making policy decisions, and studying societal issues.

7. Statistical Modeling: Statistical modeling involves building mathematical models to represent and analyze complex relationships and patterns in data. It includes techniques such as linear regression, logistic regression, time series analysis, and multivariate analysis. Statistical modeling helps in understanding the underlying structure and dynamics of data.

8. Data Mining and Machine Learning: Data mining and machine learning involve the application of statistical techniques to extract knowledge and patterns from large datasets. These branches focus on developing algorithms and models to analyze and predict outcomes based on patterns and relationships found in the data.

These are just a few examples of the branches of statistics, and the field continues to evolve with new developments in data science, computational statistics, and interdisciplinary applications.