1. Bernoulli random variables take (only) the values 1 and 0.

ANS 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?

ANS a) Central Limit Theorem

3. Which of the following is incorrect with respect to use of Poisson distribution?

ANS b) Modelling bounded count data.

4. Point out the correct statement.

Ans c) The square of a standard normal random variable follows what is called chi-squared distribution

5. ______ random variables are used to model rates.

ANS c) Poisson

6. 10. Usually replacing the standard error by its estimated value does change the CLT.

ANS b) False

7. Which of the following testing is concerned with making decisions using data?

ANS b) Hypothesis

8. Normalized data are centered at_____and have units equal to standard deviations of the original data.

ANS a) 0

9. Which of the following statement is incorrect with respect to outliers?

ANS c) Outliers cannot conform to the regression relationship

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

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

$\mathbf{f(x) = (1/\sqrt{(2\pi\sigma^2)})\ (e^{[-(x-\mu)^{\wedge}2]/2\sigma^{\wedge}2})}$

x = value of the variable or data being examined and f(x) the probability function

μ = the mean

σ = the standard deviation

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

- Deleting Rows with missing values
- Impute missing values for continuous variable
- Impute missing values for categorical variable
- Other Imputation Methods
- Using Algorithms that support missing values
- Prediction of missing values
- Imputation using Deep Learning Library — Datawig

The choice of imputation technique depends on the characteristics of the data and the assumptions that can be reasonably made. It's often recommended to assess the impact of different imputation methods on the results and perform sensitivity analyses to understand the robustness of the conclusions drawn from the imputed data. Additionally, the reasons for missingness (missing completely at random, missing at random, or missing not at random) should be considered when selecting an imputation approach.

12. What is A/B testing?

A/B testing, also known as split testing, is a statistical method used in marketing, product development, and other fields to compare two versions of a product or strategy and determine which one performs better. It involves dividing a sample or audience into two groups (Group A and Group B) and exposing each group to a different version of a variable, such as a webpage, advertisement, or product feature. The purpose of A/B testing is to assess the impact of changes and make data-driven decisions.

formulate hypothesis: Define the objective of the test and formulate hypotheses about the expected impact of changes. For example, a hypothesis might be that changing the colour of a call-to-action button will increase the click-through rate.

Randomly assign participants or users to either Group A or Group B. This randomization helps ensure that the groups are comparable, and any differences in outcomes can be attributed to the changes being tested

Randomly assign participants or users to either Group A or Group B. This randomization helps ensure that the groups are comparable, and any differences in outcomes can be attributed to the changes being tested.

Apply the different versions (A and B) of the variable to the respective groups. Group A receives the original version (control), while Group B receives the modified version (treatment).

Collect relevant data on the outcomes of interest. This could include metrics such as click-through rates, conversion rates, sales, or other key performance indicators.

Use statistical methods to analyse the data and determine whether there is a statistically significant difference between the two groups. Common statistical techniques include t-tests or chi-square tests, depending on the nature of the data.

Based on the analysis, draw conclusions about the effectiveness of the changes. If there is a significant difference, it may be concluded that one version outperforms the other. This information can then inform decisions about which version to implement or further refine.

13. Is mean imputation of missing data acceptable practice?

Mean imputation, where missing values are replaced by the mean of the observed values for that variable, is a simple and commonly used method for handling missing data. However, its acceptability depends on the context and assumptions underlying the data

Mean imputation is straightforward and easy to implement. It retains all cases in the dataset, avoiding data loss. Mean imputation is unbiased when data are missing completely at random (MCAR), meaning the probability of missingness is unrelated to the observed or unobserved data

Disadvantages and considerations

If the missingness is related to the values of the variable being imputed (missing not at random, MNAR), mean imputation can introduce bias, leading to inaccurate parameter estimates.

Mean imputation tends to underestimate the variability of the imputed variable, potentially affecting statistical analyses.

Mean imputation does not consider relationships between variables, and it assumes that missing values are missing completely at random While mean imputation is a quick and easy method, researchers and analysts should be cautious about its use, especially when missingness is not completely at random. Alternative imputation methods, such as multiple imputation, k-nearest neighbors imputation, or predictive modeling, may be more appropriate in situations where the assumptions of mean imputation are violated. Multiple imputation, in particular, is considered a robust approach as it accounts for uncertainty related to missing data.

14. What is linear regression in statistics?

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data. The goal of linear regression is to find the best-fitting line (or hyperplane, in the case of multiple independent variables) that minimizes the sum of squared differences between the observed values and the values predicted by the model.

The general form of a simple linear regression equation with one independent variable is

$$Y=\beta_0+\beta_1X+\varepsilon$$

Linear regression is widely used for various purposes, including:

Predictive modelling: Predicting the value of the dependent variable based on the values of the independent variables.

Understanding relationships: Analysing and quantifying the relationships between variables. Hypothesis testing: Assessing the significance of the relationships and coefficients. Variable selection: Identifying the most important variables contributing to the model.

It's important to note that linear regression makes certain assumptions, such as linearity, independence of errors, homoscedasticity (constant variance of errors), and normality of errors. Violations of these assumptions can affect the validity of the results.

15. What are the various branches of statistics?

Statistics is a broad field with various branches, each focusing on specific aspects of data analysis, interpretation, and application. Some of the major branches of statistics include:

Descriptive Statistics: involves methods for summarizing and organizing data. Measures such as mean, median, mode, range, and standard deviation fall under descriptive statistics.

Inferential Statistics: Involves making inferences or predictions about a population based on a sample of data. This includes hypothesis testing, confidence intervals, and regression analysis.

Probability Statistic: Studies the likelihood of events occurring. It provides a theoretical foundation for statistical methods and is essential in understanding randomness and uncertainty.