

1. a) true
2. A) central limit theorem
3. B) modelling bounded count data
4. D) all of the mentioned
5. C) poisson
6. B) false
7. b) hypothesis
8. A) 0
9. C) outliers cannot conform to the regression relationship
10. Normal distribution/gaussian distribution: is a continuous probability distribution characterized by its bell – shaped curve. It is defined by two parameters: the mean(average) and standard deviation (measures the spread). In normal distribution ,about 68% of data falls within one standard deviation of the mean, 95% within two and 99.7% within three, a property known as the empirical rule. It is widely used in statistics and is the foundation for many statistical tests.
11. Handling the missing data can be approached through various techniques.
Common methods include:
 - a) mean/ median imputation: replacing missing values with the mean or median of the observed values. This is simple but can introduce bias.
 - b) K- nearest neighbors(knn) imputation: using the average of the nearest neighbours to fill in missing values, which considers the relationship between features.
 - c) multiple imputation: creating multiple datasets by imputing values several times to account for uncertainty and then pooling results.
 - d) regression imputation: predicting missing values based on other variables in the dataset.the right method depends on the nature of the data and the extent of missing values.
12. A/b testing (split testing): is a statistical method used to compare two versions of a variable to determine which one performs better. In this approach , a sample of users in randomly divided into two groups:
group a(control) experiences the original version , while group b(treatment) experiences the modified version.
key metrics, such as conversion rates, are then analyzed to assess the impact of the changes. A/b testing is widely used in marketing, web design and product development to make data driven decisions.

13. Mean imputation can be a quick and easy method for handling missing data, but it not always considered an acceptable practice. While it can preserve the overall sample size, it can lead to biased estimates and underestimation of variability. It assumes that data are missing completely at random and can distort relationship between variables. For better accuracy and integrity of data we can use more sophisticated technique like multiple imputation or knn.
14. LR in statistics: 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 observed data. The equation for this is $Y = a + bX + \epsilon$ (y=dependent variable, a=intercept, b=slope coefficient of independent variable X, ϵ =ERROR TERM.) LR widely used for prediction and forecasting.
15. Two main branches of statistics:
- a) Descriptive statistics: involves summarizing and describing the features of a dataset through measures such as mean, median, variance and standard deviation. It provides a way to present data in a meaningful manner.
 - b) Inferential statistics: this branch focuses on drawing conclusions and making predictions about a population based on sample of data. It involves hypothesis testing, confidence intervals and regression analysis. Other subfields include biostatistics, econometrics, psychometrics and quality control, among others. Each applies statistical methods to specific domains.

