1. b) False
2. a) Central Limit Theorem
3. b) Modeling 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:**

A symmetric probability distribution with a bell-shaped curve is called a normal distribution, sometimes referred to as a Gaussian distribution. The mean (μ) and standard deviation (σ) are the two parameters that define the curve's form. The standard deviation establishes the values' spread or dispersion, while the mean denotes the distribution's center.

The 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

In a distribution that is normal:

- Because of the curve's symmetry, values have an equal chance of occurring on either side of the mean.
- The distribution's center is home to the equal mean, median, and mode.
- One standard deviation of the mean is accounted for by roughly 68% of the data, two standard deviations by 95%, and three standard deviations by 99.7%.
- Even though the likelihood of extreme values rapidly diminishes as one moves away from the mean, the distribution's tails continue to extend indefinitely.

A basic idea in statistics, the normal distribution is frequently used to represent and examine real-world data in a variety of contexts. Height, IQ levels, measurement mistakes, and other natural phenomena all show signs of a normal distribution. The Central Limit Theorem also asserts that, independent of the initial distribution of the

variables, the distribution of the sum (or average) of a large number of independent, identically distributed random variables approaches a normal distribution.

****

11. In data analysis, handling missing data is essential since it can greatly affect the validity of the findings. There exist multiple methods to handle missing data, and the selection of a method is contingent upon the characteristics of the data and the underlying causes of the missingness.

The following are some typical methods for dealing with missing data:

i. **Removal/Deletion:**

Complete Case Analysis: Listwise Deletion: Eliminating any row that has missing values. This may result in biased findings and knowledge loss, particularly if the missingness is not entirely random.

ii. **Imputation:**

- Mode, Mean, or Median Imputation: Use the observed values' mean, median, or mode to fill up any missing values for that variable. Although this is a straightforward approach, the data's variability and distribution may be distorted.
- Forward Fill or Backward Fill: Propagate the most recent observed value forward or the subsequent observed value backward, respectively, in a forward or backward fill. This is appropriate for data in time series.
- Linear Interpolation: By linearly interpolating between observed values, one can estimate missing values. suitable for data that exhibit a distinct pattern.
- Multiple Imputation: Create several datasets with values imputed to them, then examine each dataset independently. This takes the imputation process's uncertainty into account.

iii. **Advanced Techniques:**

- K-Nearest Neighbors (KNN) Imputation: Impute missing values based on the values of their k-nearest neighbors in the feature space.
- Expectation-Maximization (EM) Algorithm: Iterative algorithm for estimating parameters in models with missing data.
- Matrix Factorization Techniques: Use methods like Singular Value Decomposition (SVD) or Principal Component Analysis (PCA) for imputation.

iv. **Domain-Specific Imputation:** Impute missing values based on domain knowledge or business rules.

The kind of data, the extent of missingness, and the presumptions made regarding the missing data mechanism all influence the imputation approach selection. It's critical to record the imputation process and give serious thought to any potential biases created by the selected method.

Sensitivity tests should be carried out to evaluate how well results hold up to various imputation techniques. When feasible, multiple imputation is advised since it takes the imputation process's uncertainty into account.

****

12. **A/B Testing**:
    A/B testing, sometimes referred to as split testing, is a statistical technique for contrasting two iterations of a variable. Usually, it involves comparing a subject's response to variant A and variant B to see which one performs better. It is frequently used to evaluate the effects of alterations to a webpage or campaign in the fields of web development and marketing.

    The process involves:
    a. Selection of Variants: Choose the elements or features to be tested and create two or more variants (A and B).
    b. Random Assignment: Assign users or subjects at random to groups that will be exposed to each variant. This aids in adjusting for outside influences that can skew the outcomes.
    c. Implementation: Give the intended audience access to the versions. This could entail displaying various versions to different users of a website.
    d. Data Collection: Depending on the objectives of the test, gather pertinent data such as user engagement, click-through rates, conversion rates, or any other key performance indicators (KPIs).
    e. Statistical Analysis: Use statistical techniques to examine the gathered data and ascertain whether the variants' performances differ significantly from one another.
    f. Decision-Making: Based on the analysis, decide whether to adopt one variant over the other, or to iterate and make further changes.

****

13. A straightforward and popular technique for dealing with missing data is mean imputation, in which the observed values of a variable are substituted for the missing values. Mean imputation has benefits and drawbacks, and its acceptability varies depending on the assumptions and context, despite being simple to use and maintaining the same sample size.

    Advantages:
    - Simple: Mean imputation is easy to understand and implement.
    - Preservation of Sample Size: It keeps the sample size unchanged, which is desirable in some situations.
    - Works Well for Missing Completely at Random (MCAR) Data: If the missing data is completely at random, mean imputation can provide unbiased estimates.

Drawbacks:

- May induce Bias: When missing data is not entirely absent at random (MCAR), mean imputation may induce bias. It may not be accurate to infer that the mean of the missing data is the same as the observed values.
- Distorts Variability: Because mean imputation ignores the uncertainty arising from the imputation process, it tends to underestimate the underlying variability in the data.
- Assumes Normality: The distribution of the data is assumed to be normal. The use of mean imputation might not be appropriate if the data are not roughly regularly distributed.
- Connections may be Affected: Mean imputation has the potential to skew connections between variables, particularly when there are consistent patterns in the missing data.
- Does Not Reflect the True Distribution: It does not preserve the true distribution of the variable, potentially leading to inaccurate statistical inferences.

Given these factors, mean imputation is frequently seen as appropriate in some circumstances, particularly when the assumptions match the data distribution and the missing data is absent entirely at random. Other imputation strategies, such as multiple imputation or model-based imputation, may be better applicable in more complex scenarios or when working with non-random missing data. It's critical to thoroughly assess the presumptions and potential biases related to mean imputation and take into account alternate strategies depending on the particulars of the data.

****

14. Linear Regression:

By fitting a linear equation to the observed data, linear regression is a statistical technique used to model the connection between a dependent variable and one or more independent variables. Finding the best-fitting straight line across the data points that minimizes the sum of the squared differences between the observed and predicted values is the aim of linear regression.

A linear regression line equation is written in the form of:
Y = a + bX

where,
X is the independent variable and plotted along the x-axis
Y is the dependent variable and plotted along the y-axis

The slope of the line is b, and a is the intercept (the value of y when x = 0).

Many different fields employ linear regression to model relationships between variables, make predictions, and determine the direction and strength of such

correlations. To make sure the model is reliable, it's critical to evaluate the linear regression assumptions of linearity, independence, and residual normality.

****

15. **<u>Types of Statistics</u>:**

A wide range of areas within the general science of statistics are dedicated to different facets of data analysis and interpretation. Among the principal areas of statistics are the following:

- **Descriptive Statistics:** Includes techniques for enumerating and characterizing a dataset's primary characteristics. Descriptive statistics includes metrics like mean, median, mode, range, and standard deviation.
- **Inferential Statistics:** Focuses on using a sample of data to draw conclusions or forecasts about the population. Regression analysis, confidence intervals, and hypothesis testing are all included.
- **Probability:** The investigation of chance and uncertainty. Inferential statistics is theoretically supported by probability theory, which is a fundamental aspect of statistics.
- **Biostatistics:** Uses statistical techniques to examine and understand data in the public health, medical, and biological sciences. Epidemiology, clinical trials, and health research are all included.
- **Spatial Statistics:** Deals with the analysis of spatial data, including the study of patterns, relationships, and processes that vary across geographic space. Geostatistics is a subset of spatial statistics.
- **Bayesian Statistics:** Based on Bayesian probability theory, it involves updating beliefs or probabilities based on new evidence. Bayesian methods are used in various fields, including machine learning and data science.

****