Name: Adeola Odunewu
Internship N0: DS2311
Assessment: PFA Worksheet Set 1

STATISTICS WORKSHEET-1

Solutions

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

A) True

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

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

B) Modeling bounded count data

4. Point out the correct statement.

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

5. _____ random variables are used to model rates.

C) Poisson

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

B) False

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

B) Hypothesis

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

A) 0

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

C) Outliers cannot conform to the regression relationship

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

Normal Distribution: Also known as Gaussian distribution or bell curve, is a statistical concept that describes the symmetrical, bell-shaped probability distribution of a set of values. In a normal distribution:

Characteristic:
Symmetry: The distribution is symmetric around its mean (average). The mean, median, and mode are all equal and located at the center of the distribution.

Bell-shaped curve: The majority of the values cluster around the mean, and the farther a value is from the mean, the less likely it is to occur. The shape of the curve is characterized by its tails, which gradually approach but never touch the x-axis.

In conclusion, normal distribution is widely used in statistics and plays a crucial role in various fields, such as hypothesis testing, quality control, and modeling natural phenomena. Many statistical methods and tests assume that the data follow a normal distribution.

11. How do you handle missing data?

Handling missing data is a crucial aspect of data preprocessing in analysis or modeling. Various techniques exist for addressing missing values for example imputation methods (this include mean/median/mode, forward/backward fill, interpolation).

Advanced techniques: KNN imputation, matrix factorization, deep learning , and domain-specific approaches.

11b. What imputation techniques do you recommend?

Multiple imputation or a combination of methods may be preferred for robust handling, and documenting the chosen approach is crucial for result validity.

12. What is A/B testing?

A/B testing, also known as split testing, is a statistical method used in marketing and product development to compare two versions (A and B) of a variable. The process involves formulating a hypothesis, randomly assigning users to different groups, implementing the versions, collecting data on user interactions, performing statistical analysis, and making decisions based on the comparison.

A/B testing is widely employed to optimize web pages, emails, or other elements, allowing organizations to make data-driven decisions and improve user experiences by identifying which version performs better.

13. Is mean imputation of missing data acceptable practice?

Mean imputation of missing data is a common practice involving the replacement of missing values with the mean of observed values for a variable. It is simple and preserves sample size and variable distribution.

However, using mean imputation has downsides. It may change the variability in the data, and it relies on the assumption that missing data is completely random. It can also affect the relationships between variables, and it's not suitable for certain types of data like time series or categorical data. Additionally, it might introduce bias. While mean imputation is easy, it's important to be aware of these limitations. In some cases, it might be better to consider alternative methods like multiple imputation or model-based approaches for handling missing data more effectively.

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 linear relationship (line) that minimizes the sum of the squared differences between the observed values and the values predicted by the model.

15. What are the various branches of statistics?

Descriptive Statistics: This branch involves summarizing and presenting data in a meaningful way. Measures such as mean, median, mode, range, variance, and standard deviation fall under descriptive statistics.

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

Others:

Biostatistics: Biostatistics applies statistical methods to biological and health-related research. It is essential in medical research, clinical trials, epidemiology, and public health studies.

Econometrics: Econometrics is the application of statistical methods to economic data. It helps economists analyze economic relationships, test hypotheses, and make predictions.

Social Statistics: Social statistics involves the application of statistical methods to social science research, including sociology, psychology, political science, and education. It helps analyze social phenomena and trends.

Actuarial Science: Actuarial science applies statistical and mathematical methods to assess risk and uncertainty in the fields of insurance, finance, and pension planning.

Statistical Computing: Statistical computing involves the development and application of computational methods for statistical analysis. It includes programming and software development for statistical modeling and data analysis.

Reliability Engineering: Reliability engineering uses statistical methods to assess the reliability and maintainability of systems, products, and processes in engineering and manufacturing.