Statistics worksheet

1: true

2: Central Limit Theorem

3: Modeling bounded count data

4: The square of a standard normal random variable follows what is called the chi-squared distribution

5: Poisson

6: False

7: Hypothesis

8: 0

9: Outliers cannot confirm to the regression relationship

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

10: The Normal Distribution, also known as the Gaussian Distribution, is a symmetric probability distribution that is characterized by a bell-shaped curve.

In a normal distribution:

1. Symmetry: The distribution is symmetric around its mean, with the mean, median, mode all being equal and positioned at the center of the curve.
2. Bell-shaped curve: The curve is bell-shaped and continuous, tapering off indefinitely in both directions from the mean.
3. 68-95-99.7 Rule: Around 68% of the data falls within one standard deviation from the mean, about 95% within two standard deviations, and approximately 99.7% within three standard deviations.
4. Standardization: Normal distributions can be standardized, allowing conversion of data to Z-scores to facilitate comparisons across different normal distributions

11: Handling missing data is crucial in maintaining the integrity and reliability of analyses.

There are several approaches to address missing data:

1. Deletion: This involves removing cases lead to a loss of valuable information, especially if the missing data isn't random (MCAR-Missing Completely at Random)
2. Imputation: Imputation involves estimating missing values based on the available information.

Some common imputation techniques include:

1: Mean/Median/Mode Imputation: Replace missing values with the mean, median or mode of the observed data for that variable. It's simple but doesn’t consider the relationship between variables.

2. Multiple Imputation: Generate multiple complete datasets by imputing missing values multiple times with different plausible values based on the observed data's distribution.

3. K-Nearest Neighbors (KNN) Imputation: Estimate missing values by considering the values of similar cases based on other variables.

4. Regression Imputation: Predict missing values using regression models based on other variables that are correlated with the variables containing missing data.

5. Machine Learning-based Imputation: Techniques like Random Forest or deep learning can be used to predict missing values based on the relationship with other variables in datasets.

12: A/B testing is a controlled experiment used in various fields, particularly in marketing, web/app development, and other areas where decisions are data-drive

Here’s how it generally works:

1. Setup
2. Randomization
3. Measurement
4. Statistical Analysis
5. Decision Making

A/B testing allows business and organisation to make data-driven decision by testing changes and innovation with real users, thereby minimizing risks and improving the likelihood of implementing changes that positively impact key metrics

It's important to design A/B tests carefully, considering factors like sample size, duration and ensuring that the differences observed arre statistically significant and not due to random variation.

13: Mean imputation, where missing values are replaced by the mean of the observed data for that variable, is a simple and quick method to handle missing data.

However, its use comes with some caveats and limitations:

1: Impact on Variance

2: Bias Introduction

3: Distortion of Relationships

4: Applicability

While mean imputation is straightforward, it's considered a relatively crude method, and more sophisticated imputation techniques often provide more accurate estimates and better account for the complexity of missing data patterns.

In summary, mean imputation can be used when missingness is minimal and missing data are completely at random but it's essential to consider its limitations and potential impact on the data analysis before applying it, especially in scenarios where the assumptions might not hold true.

14: Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the variables, aiming to find the best-fitting linear equation that predicts the dependent variable based on the independent variables.

The basic form of a linear regression model with one independent variable can be expressed as:

$Y=\beta 0+\beta 1X+\varepsilon$

Where:

Y is the dependent variable.

X is the independent variable.

β0 is the y-intercept (constant term).

β1 is the slope of the line (coefficient for the independent variable X).

ε represents the error term, capturing the difference between the observed and predicted values.

The goal of linear regression is to estimate the coefficients β0 and β1 that minimize the sum of squared differences between the observed and predicted values.

15: Statistics is a broad field that encompasses various branches, each focusing on different aspects of data analysis, inference, and application.

Some major branches of statistics include:

1. Descriptive Statistics: Involves methods to summarize and describe the features of a dataset, such as measures of central tendency (mean, median, mode), dispersion (variance, standard deviation), and graphical representations (histograms, box plots).
2. Inferential Statistics: Concerned with making inferences or predictions about a population based on a sample. It includes hypothesis testing, estimation, and determining the reliability of those inferences.

3. Probability Theory: The foundation of statistics, dealing with the likelihood of events occurring. It includes concepts like random variables, probability distributions, joint and conditional probabilities, and the laws of probability.
4. Biostatistics: Focuses on the application of statistical methods in biology, medicine, and health sciences. It involves clinical trials, epidemiology, genetics, and analysis of health-related data.
5. Econometrics: Applies statistical methods to economic data, studying relationships within economic systems, forecasting economic trends, and evaluating economic policies.
6. Spatial Statistics: Concerned with the analysis of spatial and geographical data, exploring patterns, distributions, and relationships in geographic space.
7. Time Series Analysis: Deals with analyzing and modeling data that are collected and recorded over time, studying trends, seasonality, and forecasting future values.
8. Multivariate Statistics: Involves the analysis of datasets with multiple variables simultaneously, examining relationships and patterns among these variables.
9. Statistical Learning/Machine Learning: Focuses on developing and applying algorithms that allow computers to learn from and make predictions or decisions based on data without explicit programming.

10. Bayesian Statistics: Utilizes Bayesian methods for statistical inference, involving the use of prior probabilities and updating those probabilities as new information becomes available.

These branches often overlap and intersect, and the application of statistical techniques from multiple branches is common in solving complex problems across various fields.