STATISTICS WORKSHEET-1

Harshal Pandit DS2406

# Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

1. Bernoulli random variables take (only) the values 1 and 0.
   1. True
   2. 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?
   1. Central Limit Theorem
   2. Central Mean Theorem
   3. Centroid Limit Theorem
   4. All of the mentioned
3. Which of the following is incorrect with respect to use of Poisson distribution?
   1. Modeling event/time data
   2. Modeling bounded count data
   3. Modeling contingency tables
   4. All of the mentioned
4. Point out the correct statement.
   1. The exponent of a normally distributed random variables follows what is called the log- normal distribution
   2. Sums of normally distributed random variables are again normally distributed even if the variables are dependent
   3. The square of a standard normal random variable follows what is called chi-squared distribution
   4. All of the mentioned
5. random variables are used to model rates.
   1. Empirical
   2. Binomial
   3. Poisson
   4. All of the mentioned
6. . Usually replacing the standard error by its estimated value does change the CLT.
   1. True
   2. False
7. 1. Which of the following testing is concerned with making decisions using data?
   1. Probability
   2. Hypothesis
   3. Causal
   4. None of the mentioned
8. 4. Normalized data are centered at and have units equal to standard deviations of the original data.
   1. 0
   2. 5
   3. 1
   4. 10
9. Which of the following statement is incorrect with respect to outliers?
   1. Outliers can have varying degrees of influence
   2. Outliers can be the result of spurious or real processes
   3. Outliers cannot conform to the regression relationship
   4. None of the mentioned

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

1. **What do you understand by the term Normal Distribution?**

he normal distribution, also known as the Gaussian distribution, is a continuous probability distribution characterized by its symmetric, bell-shaped curve. It is defined by two parameters: the mean (μ) and the standard deviation (σ). Here are some key features:

1. **Symmetry**: The distribution is symmetric around its mean, meaning the left and right sides of the curve are mirror images of each other.
2. **Mean and Standard Deviation**: The mean (μ) determines the center of the distribution, while the standard deviation (σ) measures the spread or dispersion of the data. A larger standard deviation results in a wider curve.
3. **68-95-99.7 Rule**: About 68% of the data falls within one standard deviation of the mean, 95% falls within two standard deviations, and 99.7% falls within three standard deviations. This is often referred to as the empirical rule.
4. **Probability Density Function**: The probability density function (PDF) of a normal distribution is given by:
5. **Asymptotic Nature**: The tails of the normal distribution curve approach the horizontal axis but never actually touch it, extending infinitely in both directions.

The normal distribution is widely used in statistics and natural sciences due to its properties and its role in the Central Limit Theorem, which states that the distribution of the sample mean of a large number of independent, identically distributed random variables approaches a normal distribution, regardless of the original distribution.

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

Handling missing data is crucial for ensuring the quality and validity of your analysis. Several imputation techniques can be used, depending on the nature of the data and the amount of missing information. Here are some common approaches:

### **1. Deletion Methods**

* **Listwise Deletion**: Remove entire rows with any missing values. This method is simple but can lead to loss of valuable data, especially if the proportion of missing data is high.
* **Pairwise Deletion**: Use all available data for each pair of variables. This can retain more data but might lead to inconsistencies in the analysis if the missing data patterns are not random.

### **2. Imputation Techniques**

* **Mean/Median/Mode Imputation**: Replace missing values with the mean (for continuous data), median, or mode (for categorical data) of the available data. This method is simple but can introduce bias, especially if the data is not missing at random.
* **K-Nearest Neighbors (KNN) Imputation**: Use the values of the nearest neighbors to estimate missing values. This technique considers the similarity between instances and can work well for numerical and categorical data.
* **Regression Imputation**: Use a regression model to predict the missing values based on other available variables. This method can be effective but assumes that the relationship between variables is linear.
* **Multiple Imputation**: Generate several different imputed datasets, analyze each dataset separately, and then combine the results. This approach accounts for the uncertainty of missing data and provides more robust estimates.
* **Interpolation**: For time-series data, interpolate missing values based on neighboring data points. This can be linear or more complex, depending on the nature of the data.
* **Expectation-Maximization (EM) Algorithm**: Use the EM algorithm to estimate missing values based on the likelihood of the data. This method is useful for data that follow a certain distribution.

### **3. Advanced Methods**

* **Machine Learning Algorithms**: Use algorithms like Random Forest or other advanced techniques to predict and impute missing values. These methods can handle complex relationships and interactions.
* **Data Augmentation**: Use techniques like Generative Adversarial Networks (GANs) to generate plausible data to fill in missing values, especially in complex datasets.

### **4. Considerations**

* **Analyze Missing Data Patterns**: Understand why data is missing and how it might impact your analysis. Techniques like Missingness Mechanism (MCAR, MAR, MNAR) can guide the choice of imputation method.
* **Evaluate Impact**: Assess how imputation affects your results by comparing analyses with and without imputation.

Choosing the right imputation technique depends on the data's characteristics, the missing data pattern, and the analysis goals. It's often useful to experiment with different methods and validate the results to ensure robustness.

1. **What is A/B testing?**

A/B testing, also known as split testing, is a method used to compare two versions of a variable to determine which one performs better. It is widely used in marketing, product design, and web development to make data-driven decisions. Here’s a breakdown of how A/B testing works:

### **1. Define Objectives**

* **Identify the Goal**: Determine what you want to improve or test, such as click-through rates, conversion rates, or user engagement.

### **2. Create Variants**

* **Version A (Control)**: This is the current version or the original variant.
* **Version B (Treatment)**: This is the modified version with changes made to test against the control.

### **3. Random Assignment**

* **Split the Audience**: Randomly assign your audience into two groups. One group is exposed to Version A, while the other is exposed to Version B. This helps ensure that the results are due to the changes made and not other factors.

### **4. Run the Test**

* **Collect Data**: Allow the test to run for a sufficient period to gather enough data for meaningful analysis.

### **5. Analyze Results**

* **Compare Performance**: Measure the performance of each variant based on the predefined metrics (e.g., conversion rates, click rates).
* **Statistical Analysis**: Use statistical methods to determine if the differences in performance are statistically significant.

### **6. Make Decisions**

* **Implement Findings**: Based on the results, decide whether to implement the changes from Version B, stick with Version A, or conduct further testing.

### **Key Considerations**

* **Sample Size**: Ensure the sample size is large enough to detect meaningful differences.
* **Duration**: Run the test long enough to account for variability and ensure the results are not due to random chance.
* **Bias and Variability**: Avoid biases in random assignment and control external factors that could affect the results.

### **Applications**

* **Web Design**: Testing different website layouts, colors, or call-to-action buttons to see which performs better.
* **Marketing**: Evaluating different email subject lines, ad copies, or promotional offers to maximize engagement and conversions.
* **Product Development**: Comparing new features or product designs to the existing ones to assess user preference and impact.

A/B testing is a powerful tool for optimizing performance and making informed decisions based on empirical evidence rather than intuition alone.

1. **Is mean imputation of missing data acceptable practice?**

Mean imputation, where missing values are replaced with the mean of the observed values, is a commonly used technique, but its appropriateness depends on the context. Here’s a closer look at its advantages and limitations:

### **Advantages:**

1. **Simplicity**: Mean imputation is easy to implement and understand.
2. **Preserves Sample Size**: It allows you to use all available data without removing cases with missing values.
3. **Maintains Mean**: It keeps the overall mean of the dataset unchanged, which can be useful for certain types of analyses.

### **Limitations:**

1. **Bias**: Mean imputation can introduce bias, especially if the data are not missing completely at random (MCAR). It assumes that the missing data mechanism is random and that the missing values are similar to the observed ones, which may not always be true.
2. **Underestimates Variability**: By replacing missing values with a constant mean, it reduces the variability in the data. This can lead to underestimation of the true variability and affect statistical analyses that rely on variance.
3. **Distorts Relationships**: Mean imputation may distort the relationships between variables and affect the results of regression or other modeling techniques.

### **When It Might Be Acceptable:**

* **Small Proportions of Missing Data**: If the amount of missing data is very small and randomly distributed, mean imputation may be acceptable as a quick fix.
* **Exploratory Analysis**: For preliminary analyses or if more sophisticated methods are not feasible, mean imputation can be used.

### **Alternative Techniques:**

* **Multiple Imputation**: Generates several imputed datasets and combines results, accounting for the uncertainty of missing data.
* **K-Nearest Neighbors (KNN)**: Imputes missing values based on the values of the nearest neighbors.
* **Regression Imputation**: Uses regression models to predict missing values based on other variables.
* **Interpolation**: For time-series data, interpolation methods can be used to estimate missing values.

In summary, while mean imputation can be a simple and practical approach in certain situations, it is often advisable to use more advanced imputation techniques, especially if a substantial amount of data is missing or if preserving the accuracy and variability of the data is crucial for your analysis.

1. **What is linear regression in statistics?**

Linear regression is a statistical method used to model and analyze the relationship between a dependent variable and one or more independent variables. The goal is to identify the linear relationship between these variables and make predictions based on this relationship. Here’s a breakdown of the key components and concepts:

### **1. Simple Linear Regression**

* **Model**: The simplest form of linear regression involves a single independent variable. The model is expressed as:

Y=β0+β1X+ϵY = \beta\_0 + \beta\_1 X + \epsilonY=β0​+β1​X+ϵ

where:

* + YYY is the dependent variable (the variable you want to predict or explain).
  + XXX is the independent variable (the predictor or explanatory variable).
  + β0\beta\_0β0​ is the y-intercept (the value of YYY when X=0X = 0X=0).
  + β1\beta\_1β1​ is the slope of the line (the change in YYY for a one-unit change in XXX).
  + ϵ\epsilonϵ is the error term (the difference between the observed and predicted values of YYY).

### **2. Multiple Linear Regression**

* **Model**: When there are two or more independent variables, the model is extended to:

### **3. Assumptions of Linear Regression**

* **Linearity**: The relationship between the dependent and independent variables is linear.
* **Independence**: The residuals (errors) are independent of each other.
* **Homoscedasticity**: The residuals have constant variance across all levels of the independent variables.
* **Normality**: The residuals are normally distributed.

### **4. Estimation and Interpretation**

* **Estimation**: The coefficients (β\betaβ) are estimated using methods like Ordinary Least Squares (OLS), which minimizes the sum of the squared residuals to find the best-fitting line.
* **Interpretation**: The coefficients indicate the strength and direction of the relationship between each independent variable and the dependent variable. For example, a positive coefficient means that as the independent variable increases, the dependent variable is expected to increase as well.

### **5. Applications**

* **Prediction**: Linear regression can be used to predict the value of the dependent variable based on new values of the independent variables.
* **Trend Analysis**: It helps identify trends and relationships in data.
* **Data Exploration**: Useful for understanding the impact of one or more independent variables on the dependent variable.

Linear regression is a foundational technique in statistics and machine learning, providing a straightforward approach to understanding relationships between variables and making predictions.

1. **What are the various branches of statistics?**

Statistics is a broad field with various branches that cater to different types of data analysis and inference. Here are the major branches:

### **1. Descriptive Statistics**

* **Purpose**: Summarize and describe the main features of a dataset.
* **Techniques**: Measures of central tendency (mean, median, mode), measures of dispersion (range, variance, standard deviation), and graphical representations (histograms, box plots).

### **2. Inferential Statistics**

* **Purpose**: Make inferences and draw conclusions about a population based on a sample.
* **Techniques**: Hypothesis testing, confidence intervals, and estimation. Includes various tests such as t-tests, chi-squared tests, and ANOVA.

### **3. Probability Theory**

* **Purpose**: Study the mathematical framework of randomness and uncertainty.
* **Concepts**: Probability distributions, random variables, expected value, variance, and probability laws.

### **4. Regression Analysis**

* **Purpose**: Explore and model relationships between a dependent variable and one or more independent variables.
* **Types**: Linear regression, multiple regression, logistic regression, and non-linear regression.

### **5. Experimental Design**

* **Purpose**: Plan and conduct experiments to ensure valid and reliable results.
* **Concepts**: Randomization, control groups, factorial designs, and the principles of causation and confounding.

### **6. Bayesian Statistics**

* **Purpose**: Use Bayesian inference to update the probability of a hypothesis as more evidence or information becomes available.
* **Concepts**: Prior distribution, likelihood, posterior distribution, and Bayesian inference methods.

### **7. Multivariate Statistics**

* **Purpose**: Analyze data involving multiple variables simultaneously to understand relationships and patterns.
* **Techniques**: Principal Component Analysis (PCA), Factor Analysis, Cluster Analysis, and Multivariate Analysis of Variance (MANOVA).

### **8. Non-Parametric Statistics**

* **Purpose**: Analyze data without assuming a specific distribution shape.
* **Techniques**: Rank-based tests such as the Wilcoxon signed-rank test, Kruskal-Wallis test, and Spearman's rank correlation.

### **9. Time Series Analysis**

* **Purpose**: Analyze data points collected or recorded at specific time intervals.
* **Techniques**: Autoregressive models (AR), Moving Average models (MA), ARIMA models, and Seasonal decomposition.

### **10. Survival Analysis**

* **Purpose**: Analyze the time until an event of interest occurs, often used in clinical trials and reliability engineering.
* **Techniques**: Kaplan-Meier estimator, Cox proportional hazards model, and survival curves.

### **11. Statistical Learning**

* **Purpose**: Use statistical methods for predictive modeling and machine learning.
* **Techniques**: Classification and regression trees (CART), Support Vector Machines (SVM), Neural Networks, and Ensemble methods.

### **12. Statistical Computing**

* **Purpose**: Develop and use computational methods and software for statistical analysis.
* **Tools**: Programming languages like R and Python, and software packages like SAS, SPSS, and MATLAB.

Each of these branches serves different purposes and is applied in various fields such as economics, engineering, social sciences, health sciences, and more. Understanding these branches helps in selecting the appropriate statistical methods for analyzing data and making informed decisions.

