# STATISTICS WORKSHEET-1

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

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

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

A normal distribution, also known as Gaussian distribution, is a continuous probability distribution that is symmetric around its mean, forming a bell-shaped curve. This distribution is characterized by the following properties:

1. **Symmetry:** The curve is symmetric, with the mean, median, and mode all located at the center. The distribution is bell-shaped, and the tails extend infinitely in both directions.
2. **Mean, Median, and Mode:** All three measures of central tendency (mean, median, and mode) are equal and located at the center of the distribution.
3. **Standard Deviation:** The spread of the distribution is determined by the standard deviation. About 68% of the data falls within one standard deviation from the mean, 95% within two standard deviations, and 99.7% within three standard deviations.
4. **Probability Density Function (PDF):** The probability density function of a normal distribution is given by the famous bell-shaped curve described by the formula $(x)=\sigma 2\pi 1e-21(\sigma x-\mu)2$, where $\mu$ is the mean and $\sigma$ is the standard deviation.
5. **Z-Score:** The Z-score measures how many standard deviations a data point is from the mean. It is calculated as $Z=\sigma x-\mu$.

The normal distribution is fundamental in statistics and probability theory. Many natural phenomena, such as heights, IQ scores, and measurement errors, tend to follow a normal distribution. The Central Limit Theorem also states that the distribution of the sum (or average) of a large number of independent, identically distributed random variables approaches a normal distribution, even if the original variables are not normally distributed.



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

Handling missing data is a crucial step in the data preprocessing phase. The choice of imputation techniques depends on the nature of the data and the reason for missingness. Here are some common strategies for handling missing data:

1. **Deletion:**
    a. **Listwise Deletion:** Removing entire rows with missing values. This can lead to a loss of valuable information, especially if the missingness is not random.
    b. **Pairwise Deletion:** Analyzing only the available pairs of data for each variable in an analysis. This can lead to varying sample sizes for different analyses.
2. **Imputation:**
    a. **Mean, Median, or Mode Imputation:** Replace missing values with the mean, median, or mode of the observed values for that variable. This is a simple method but may not be suitable if the data has outliers.
    b. **Forward Fill/Backward Fill:** Propagate the last known value forward to fill missing values (forward fill) or use the next known value to fill missing values (backward fill). This is often used for time-series data.
    c. **Linear Interpolation:** Estimate missing values by linearly interpolating between neighboring observed values. This is suitable for ordered data.
    d. **Multiple Imputation:** Generate multiple datasets with imputed values to account for uncertainty in imputation. This method is more sophisticated but computationally intensive.

3. **Prediction Models:**
    a. **Regression Imputation:** Predict missing values using regression models based on other variables in the dataset.
    b. **K-Nearest Neighbors (KNN):** Impute missing values based on the values of their k-nearest neighbors in the feature space.
    c. **Machine Learning Models:** Train machine learning models to predict missing values based on other features in the dataset.
4. **Domain-Specific Imputation:**
    a. **Custom Imputation:** Use domain-specific knowledge to impute missing values. This might involve using external data sources or expert judgment.

12. What is A/B testing?

A/B testing, also known as split testing, is a method of comparing two versions of a webpage or app against each other to determine which one performs better. It is a controlled experiment where two variants, A and B, are compared by testing a subject's response to variant A against variant B and determining which of the two variants is more effective.

Here's a basic overview of how A/B testing works:

1. **Objective Definition:** Clearly define the objective of the test. This could be improving click-through rates, increasing conversion rates, or any other metric that aligns with your business goals.
2. **Variant Creation:** Create two versions (A and B) of the element you want to test. This could be a webpage, email, advertisement, or any other component.
3. **Random Assignment:** Users or participants are randomly assigned to either variant A or B. This helps ensure that any differences in the outcomes are due to the variations and not other external factors.
4. **Implementation:** Implement variants A and B simultaneously and collect relevant data on user interactions or conversions.
5. **Analysis:** Analyze the data to determine which variant performed better based on the defined objective. Common metrics for analysis include conversion rates, click-through rates, revenue, or other key performance indicators.
6. **Statistical Significance:** Ensure that the results are statistically significant, meaning that any observed differences are not likely due to random chance. Statistical significance helps ensure that the observed differences are likely to generalize to the entire population.
7. **Decision:** Based on the analysis, decide whether to adopt variant A or B. If one variant significantly outperforms the other, it can be implemented permanently.

A/B testing is widely used in marketing, web development, and product management to optimize user experience and achieve business objectives. It allows businesses to make data-driven decisions, iterate on design and content, and continuously improve performance.

13. Is mean imputation of missing data acceptable practice?

Mean imputation, which involves replacing missing values with the mean of the observed values for a variable, is a common and simple imputation method. However, its acceptability depends on the context and the assumptions underlying the missing data.

**Advantages:**

- **Simplicity:** Mean imputation is straightforward and easy to implement.

- **Preservation of Sample Size:** It preserves the sample size, which can be important in cases where data is limited.

**Concerns and Limitations:**

- **Bias:** Mean imputation can introduce bias, especially if the missing data is not missing completely at random (MCAR). If the missingness is related to the values of the variable itself or other variables, mean imputation can distort the distribution and relationships in the data.
- **Underestimation of Variability:** Mean imputation tends to underestimate the variability of the imputed variable, as all imputed values are the same. This can impact the accuracy of statistical analyses and hypothesis testing.
- **Impact on Correlations:** Mean imputation can affect correlations between variables, particularly when missing values are associated with specific groups or conditions.
- **Assumption of Normality:** Mean imputation assumes that the variable follows a normal distribution. If the variable is not normally distributed, mean imputation may not be appropriate.
- **Influence on Relationships:** In regression analysis, mean imputation can bias coefficients and standard errors, leading to incorrect inferences.

**Alternatives:**

- **Multiple Imputation:** Generating multiple datasets with imputed values to account for uncertainty.
- **Regression Imputation:** Predicting missing values using regression models based on other variables.
- **Domain-Specific Imputation:** Using domain knowledge to impute missing values.

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 observed data. The simplest form of linear regression is known as simple linear regression, which involves modeling the relationship between two variables: one independent variable (predictor) and one dependent variable (response). The equation of a simple linear regression model is:

$$\beta 0+\beta 1\cdot x+\varepsilon$$

- $y$ is the dependent variable.
- $x$ is the independent variable.
- $\beta 0$ is the intercept (the value of $y$ when $x$ is 0).
- $\beta 1$ is the slope (the change in $y$ for a one-unit change in $x$).
- $\varepsilon$ represents the error term, which accounts for the variability in $y$ that cannot be explained by the linear relationship with $x$.

The goal of linear regression is to estimate the coefficients ($\beta 0$ and $\beta 1$) that minimize the sum of squared differences between the observed values of the dependent variable and the values predicted by the linear model. This process is often done using the method of least squares.

Extensions of linear regression include multiple linear regression, where there are multiple independent variables, and polynomial regression, where the relationship between variables is modeled with higher-degree polynomials.

Linear regression is widely used in various fields for prediction, modeling, and understanding the relationships between variables. It is an essential tool in statistics and machine learning.

15. What are the various branches of statistics?

Statistics is a broad field with various branches, each focusing on different aspects of data analysis and interpretation. Some of the major branches of statistics include:

1. **Descriptive Statistics:** Involves methods for summarizing and describing the main features of a dataset. Measures such as mean, median, mode, standard deviation, and percentiles fall under descriptive statistics.
2. **Inferential Statistics:** Involves making inferences or predictions about a population based on a sample of data. Common techniques include hypothesis testing, confidence intervals, and regression analysis.
3. **Biostatistics:** Focuses on the application of statistical methods to biological and health-related fields. It plays a crucial role in medical research, clinical trials, and public health studies.
4. **Econometrics:** Applies statistical methods to economic data to test hypotheses, forecast economic trends, and estimate economic relationships.
5. **Social Statistics:** Deals with the statistical analysis of social phenomena, including demographics, surveys, and studies related to sociology.
6. **Business Statistics:** Applies statistical methods to analyze and interpret business data. It is used in market research, financial analysis, and quality control.
7. **Environmental Statistics:** Involves the use of statistical methods to analyze environmental data, including studies on climate, pollution, and ecology.
8. **Psychological Statistics:** Applies statistical techniques to psychological research and studies. It includes experimental design, hypothesis testing, and data analysis in psychology.
9. **Educational Statistics:** Involves the application of statistical methods to educational research and assessment. It includes the analysis of test scores, educational outcomes, and program evaluations.
10. **Statistical Computing:** Focuses on the development of computational techniques and algorithms for statistical analysis. It includes programming languages and software tools used for statistical modeling and data analysis.
11. **Spatial Statistics:** Deals with the analysis of spatial data, incorporating the spatial relationships and patterns in the analysis. It is used in fields such as geography, environmental science, and urban planning.
12. **Bayesian Statistics:** Based on Bayesian probability theory, it involves updating probabilities based on prior knowledge and new evidence. It is used in various fields for decision-making and inference.

These branches often overlap, and statisticians may specialize in one or more areas depending on their interests and expertise. The application of statistical methods is diverse, and statistics plays a crucial role in scientific research, business decision-making, public policy, and many other fields.