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
   Answer: Option A – True
   - **The Bernoulli distribution arises as the result of a binary outcome.**

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
   Answer: Option A – Central Limit Theorem
   - **The Central Limit Theorem (CLT) is one of the most important theorems in statistics.**

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
   Answer: Option B – Modeling bounded count data
   - **Poisson distribution is used for modeling unbounded count data.**

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
   Answer: Option D – All the mentioned
   - **Many random variables, properly normalized, limit to a normal distribution.**

5. _____random variables are used to model rates.
   a) Empirical
   b) Binomial
   c) Poisson
   d) All of the mentioned
   Answer: Option C – Poisson
   - **Poisson distribution is used to model counts.**

6. 10. Usually replacing the standard error by its estimated value does change the CLT.
   a) True
   b) False
   Answer: Option B – False
   - **Usually replacing the standard error by its estimated value doesn't change the CLT.**

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

   Answer: Option B – Hypothesis

   ➢ **The null hypothesis is assumed true and statistical evidence is required to reject it in favor of a research or alternative hypothesis.**

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

   Answer: Option A – 0

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

   Answer: Option C - Outliers cannot conform to the regression relationship

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

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

- The normal distribution is described by the mean (μ) and the standard deviation (σ).
- The normal distribution is often referred to as a 'bell curve' because of it's shape:
  - Most of the values are around the center (μ)
  - The median and mean are equal
  - It has only one mode
  - It is symmetric, meaning it decreases the same amount on the left and the right of the center
- The area under the curve of the normal distribution represents probabilities for the data.
- The area under the whole curve is equal to 1, or 100%

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

- Deleting Rows with missing values.
- Impute missing values for continuous variable.
- Impute missing values for categorical variable.
- Other Imputation Methods.
- Using Algorithms that support missing values.
- Prediction of missing values.

- **Techniques for Handling the Missing Data**
  - Listwise or case deletion.
  - Pairwise deletion.
  - Mean substitution.
  - Regression imputation.
  - Last observation carried forward.
  - Maximum likelihood.
  - Expectation-Maximization.
  - Multiple imputation.

12. What is A/B testing?

- A/B testing, also known as split testing, refers to a randomized experimentation process wherein two or more versions of a variable are shown to different segments of website visitors at the same time to determine which version leaves the maximum impact and drive business metrics.
  Essentially, A/B testing eliminates all the guesswork out of website optimization and enables experience optimizers to make data-backed decisions. In A/B testing, A refers to ‘control’ or the original testing variable. Whereas B refers to ‘variation’ or a new version of the original testing variable.

13. Is mean imputation of missing data acceptable practice?

- **True**, imputing the mean preserves the mean of the observed data. So if the data are missing completely at random, the estimate of the mean remains unbiased. Since most research studies are interested in the relationship among variables, mean imputation is not a good solution.

14. What is linear regression in statistics?

- In statistics, linear regression is a linear approach for modeling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables). The case of one explanatory variable is called simple linear regression.
  Linear regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications. This is because models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters and because the statistical properties of the resulting estimators are easier to determine.

15. What are the various branches of statistics?

- The two branches of statistics are

  1. descriptive statistics
  2. inferential statistics