Name of the Assignment: Statistics (Worksheet – 1)	
Submitted by	: Shalini Joshi
Designation	: Data Science Intern
Date of Submission	: 21st Dec,2022

Objective type Questions

1. Bernoulli random variables take (only) the values 1 and 0.
   a) True b) False

Ans: 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 b) Central Mean Theorem c) Centroid Limit Theorem d) All of the mentioned

Ans: a) Central Limit Theorem

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

Ans: b) Modeling bounded 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

Ans: d) All of the mentioned

5. _____ random variables are used to model rates.
   a) Empirical b) Binomial c) Poisson d) All of the mentioned

Ans: c) Poisson

6. Usually replacing the standard error by its estimated value does change the CLT.
   a) True b) False

Ans: b) False

7. Which of the following testing is concerned with making decisions using data?
   a) Probability b) Hypothesis c) Causal d) None of the mentioned

Ans: b) Hypothesis

8. Normalized data are centered at_____and have units equal to standard deviations of the original data.
   a) 0 b) 5 c) 1 d) 10

Ans: 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

Ans: c) Outliers cannot conform to the regression relationship

Subjective type Questions

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

Ans: Normal Distribution

- Normal Distribution is a continuous probability distribution where the data is symmetrical around a central value; mean with no left or right bias.
- The probabilities for values further away from the mean taper off equally in both directions.
- In this distribution, 50% values are less than the mean and 50% values are greater than the mean.
- Here, mean = median = mode. It is recognized by its 'bell shaped curve' in the statistical reports.
- It is the most important probability distribution as it accurately describes the distribution of values for many natural phenomena. For eg. Blood pressure, IQ Scores, heights etc. follow the normal distribution.

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

Ans: In order **to handle missing data**, we can use two methods:

- **Deletion of data** - The data related to the missing data points can be deleted to reduce bias. It may not be the best method if there are not enough observations.

- **Imputation of data** - This method develops reasonable guesses to substitute the missing data to retain most of the information of the dataset. It is most useful when the percentage of missing data is low.

  The following imputation techniques can be used in statistics to deal with missing data:

  - ✓ ***Imputation using Mean/Media/Mode***: This works by calculating the mean/median of the non-missing values in a column and then replacing the missing values within each column separately and independently from the others. This works with numerical data only.
  - ✓ ***Imputation using Most Frequent or Zero Constant values***: works with categorical data.

12. What is A/B testing?

Ans: A/B testing (also known as split testing or bucket testing) is a method of comparing two versions of a webpage or app against each other to determine which one performs better.

A/B testing is essentially an experiment where two or more variants of a page are shown to users at random, and statistical analysis (Hypothesis two tailed, p-value test) is used to determine which variation performs better for a given conversion goal.

13. Is mean imputation of missing data acceptable practice?

Ans: The process of replacing null values in a data collection with the data's mean is known as mean imputation.

Mean Imputation is not considered a good practice as:

- It ignores feature correlation
- Decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate and the confidence interval is narrower.

14. What is linear regression in statistics?

Ans: Linear regression is a type of predictive analysis.

The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = c + b*x, where y = estimated dependent variable

score, c = constant, b = regression coefficient, and x = score on the independent variable.

These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables.

The three major uses for regression analysis are:

(1) determining the strength of predictors,

(2) forecasting an effect, and

(3) trend forecasting.

15. What are the various branches of statistics?

Ans: There are two main branches of Statistics; Descriptive Statistics and Inferential Statistics.

- **Descriptive Statistics**: Descriptive statistics is considered as the first part of statistical analysis which deals with the collection and presentation of data. It can be defined as brief explanatory coefficients that are used by statisticians to summarize a given data set.
    - ***Measures of central tendency***: Measures of central tendency specifically help the statisticians to estimate the center of values distribution. These are: Mean, Median, Mode
    - ***Measures of variability*** : The measure of variability help statisticians to analyze the distribution spread out of a given set of data. Some of the examples of measures of variability include quartiles, range, variance and standard deviation.
- **Inferential Statistics**: Inferential statistics are techniques that enable statisticians to use the gathered information from a sample to make inferences, decisions or predictions about a given population. The different types of calculation of inferential statistics include:

- Regression analysis
- Analysis of variance (ANOVA)
- Analysis of covariance (ANCOVA)
- Statistical significance (t-test)
- Correlation analysis

********** **The End** **********