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

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

Ans. 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

Ans. b) 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

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

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.
10. What do you understand by the term Normal Distribution?

In Normal Distribution, we have a graph like normal curve or bell shaped graph. In this all the data is symmetry and the data is in decimal numbers. We have an empirical rule in normal distribution, this empirical rule is also known as three sigma rule / 68%-95%-99.7% rule. According to this rule, the data that is lying between one standard deviation is 68%, data lies under 2 standard deviation is 95%, and the data lies between 3 standard deviation is 99.7%. So, the majority of data lies between 1 standard deviation.

The data which lies under 3rd standard deviation comes under the normal curve. Normal distribution is Determined by its mean, median, mode, and standard deviation. In this, the mean = 0, median = 0, and mode = 0, and standard deviation = 1.

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

Ans. Certain times, there may be some situations when we have a dataset with missing values. It is important to deal with these missing values as it can cause bias in the dataset and it can also lead to errors or faulty analysis. To check the missing values inside the datasets, we use .isnull() and to check total number of data missing in each column, we use .isnull().sum()

There are two strategies which can be used to handle those missing values:
1. If we have low number of rows, that is less than 5% rows with missing values, then we can drop it.
In this case, We can use dropna(inplace = True) method to drop the rows with missing values.

2. If we have high number of rows with missing values, then we can replace it with mean values.
Here, we can use replace(np.NaN, column_name.mean, inplace = True) method, this will first calculate the mean value of a particular column and replace all the missing values of that column with the mean value.

Imputation Techniques:

There is a one of the popular approach for imputation is to calculate the mean value of each column and replace all the missing values of that column with the computed mean value. This approach is the easiest and popular as it is easy to calculate using training dataset and it often results in good performance.

12. What is A/B testing?

Ans. A/B test is a form of hypothesis testing. It is the way of comparing two versions of variable to check which performs better. For example, if a company wants to increase the sales of its products, it will either use random experiment or a statistical method. A/B is the tool which is widely used as statistical tools.
The company will divide the products into two parts that is A and B. There will be no changes made in A Product and significant changes would be made in product B's packaging. The company will analyse which product has performed better on the basis of customer's response over the products A and B. So, A/B test is a Hypothetical test which is used by the researchers to find out evidence and make an effective decisions based on that evidence.
When we work with hypothesis testing, there are two things, we need to take into consideration.

1. Null Hypothesis:
    - What we assume is true to begin with
    - There is no difference/effect/relationship etc.
2. Alternative Hypothesis
    - What we aim to gather an evidence of
    - There is a difference/effect/relationship etc.

   Alternative Hypothesis is a contradictory statement of Null Hypothesis.

There is an important term associated with Hypothesis Testing that is P-Value.

- If the P-Value > .05, then Null Hypothesis is True and Accepted and Alternative Hypothesis is rejected.
- If the P-Value < .05, then Alternative Hypothesis is True and Accepted and Null Hypothesis is rejected.

13. Is mean imputation of missing data acceptable practice?

Ans. Mean imputation can be an acceptable practise where the missing values are replaced with the mean value in the entire column but it can only done with numerical data. Whereas, mode imputation can be done with numerical and categorical data as well.

14. What is linear regression in statistics?

In linear regression, we have input variables that is also known as x variables, explanatory variables, predictors, and independent variables. And output variables that is also known as y variables, label, predicted, outcome, and dependent variables.

In linear regression analysis, we use independent variables to predict the value of dependent variables. So, the Value of one variable is used to predict the value of another variable.

Linear Regression uses an equation: $y = a + bx$..

Here, y is dependent variable/output variable, a is intercept, b is coefficient and x is independent variable.

The role of coefficient is whenever there is an increase of 1 in x variable, the corresponding coefficient value b That is .5 is added to y variable.

Linear regression will make the best fit line/straight line where maximum data points can meet.

If we have more x variables in the dataset, then the linear regression equation would be:

$y = a + b1x1 + b2x2 + b3x3 + bnxn$

15. What are the various branches of statistics?

Ans. There are two branches of statistics:

Statistics:

- Descriptive Statistics:
    - Central of Tendency/Measures of centres
        - Mean
        - Median
        - Mode
    - Spread of Data/Measures of dispersion
        - Variance
        - Range
        - Percentile
        - Skewness
        - Standard Deviation
- Inferential Statistics:
    - Zscore
    - Ttest
    - Hypothesis Testing
    - Chi-square Test
    - Anova
    - Manova

- Correlation

1. **Descriptive Statistics:** this is a procedure used to organize, summarize, and make sense of a dataset.
2. **Inferential Statistics:** this procedure used to generalize observations/data from the samples/subsets taken from the large population.

Descriptive Statistics: Inside descriptive statistics, we have central of tendency also known as measures of centres and spread of data also known as measures of dispersion.

Central of tendency is mean, median, and mode. Spread of data is variance, range, percentile, standard deviation, and skewness.

Inferential Statistics: Under this, we have zscore, Ttest, hypothesis testing, anova, manova, chi-square test, and correlation.