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)

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)

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)

FLIP ROBO

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)

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

   Ans: (c)

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

   Ans: (b)

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)

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)

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)

FLIP ROBO

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

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

Ans:

Normal distribution is also called as 'Gaussian Distribution'. In normal distribution mean=mode=median. The distribution is symmetric at the mean, which is half the values are above the mean and other half are below the mean. Its familiar bell-shaped curve is found everywhere in statistical reports. The graph of the normal distribution is characterized by two parameters – mean and standard deviation. The mean of a normal distribution is nothing but average, which is located at the center of the curve. The standard deviation determines the amount of dispersion away from the mean. A small standard deviation produces a steep graph where as a higher standard deviation produces a flat graph. Normal distribution is important for probability distribution in statistics for independent, random variables. It accurately describes the distribution of values for many natural phenomena such as distribution of heights of population. Data scientists use various methods to transform the data by cleaning so as to get the data as it would be in a normal distribution.

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

Ans:

To identify if there are any missing data one can use a heatmap in seaborn. Using a heatmap with isnull() we can know where the data is missing. In general, if we have less than 5 percent of data missing we could drop those and if there are more than 50% of the values are missing in a particular column, it can be dropped based on the importance. It would be ideal to drop that particular column than to keep the data and train the model. We can use various imputations techniques to impute data. Mean Median or Mode can be used to fill the missing data if there are only a few.

12. What is A/B testing?

Ans:

A/B testing is a basic randomized control experiment. It is mainly used to compare two versions of a variable to find out which performs better in a controlled environment. In A/B testing we create sample of an entire population and then use the Hypothesis testing mechanism to check if our Null Hypothesis is correct or our Alternative Hypothesis is right. We need to check where we are able to reject the Null Hypothesis or accept the Null Hypothesis keeping in the mind the Type 1 and Type 2 errors that can be checked via the confusion matrix. In Null Hypothesis the important values considered are the alpha (allowable percentage of error), p value and the accuracy level of the test model.

13. Is mean imputation of missing data acceptable practice?

Ans:

The fact is there is no particular standard method to deal the missing data. We can use different methods depending upon the data we have. Mean imputation is a non-standard practice but a fairly flexible imputation, as it can be applied to both continuous and categorical data. This makes the mean imputation take advantage over other methods. Mean imputation of the missing values is not much recommended as it preserves the mean of the observed data and leads to variation in the standard deviation, which distorts the relationships between variables by giving biased estimates of the correlation that are majorly towards zero. There are other better imputation methods which provide accurate data for filling the missing value. However, if every other technique fails then mean imputation can be used as a last resort instead of deleting null values especially for a smaller data set.

14. What is linear regression in statistics?

FLIP ROBO

Ans:

Linear Regression is a supervised machine learning algorithm that can be used to predict continuous data. The underlying equation used by linear regression model is y = mx+c where m is the slope and c is the intercept of the best fit line. There are 2 types of linear regression and they are Simple linear regression and Multiple linear regression techniques. Linear Regression is majorly used to predict the labels with the help of one or more feature variables ensuring that the features selected in the equation provide the necessary input to obtain the desired labels without creating an overfitting or underfitting model that provides a reasonable accuracy on future predictions.

15. What are the various branches of statistics ?

Ans:

Statistics is majorly categorized into two branches - Descriptive statistics and Inferential statistics. Both play equally important role in statistics and scientific analysis of data.

Descriptive statistics: In this type of statistics, the data is summarized through the given observations. Descriptive statistics can further be classified as central tendency category and dispersion of data. Central tendency contains parameters such as the mean, median, mode whereas the dispersion of data contains standard deviation, range, variance, skewness, percentile, etc. Descriptive statistics is a way to organize, represent and describe a collection of data using tables, graphs, and summary measures.

Inferential Statistics: Inferential statistics deals with techniques used for the analysis of data, making estimates and drawing conclusions from limited information obtained through sampling and testing the reliability of the estimates. We perform hypothesis testing in for inferential statistics.