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?

Answer: Normal distribution is kind of the probability distribution which is symmetric about the mean. It shows that the data near the mean are more frequent in occurrence than the data which is far from the mean. It is also called as Gaussian distribution. Its graphical representation appears like a bell curve. In the other words, we can say as normal distribution is the exact term we can use for a probability bell curve. So, in the Normal distribution the mean is 0 and the standard deviation is 1. Normal distribution are symmetrical distributions, whereas not all the symmetrical distributions are normal distribution.

11. How do you handle missing data? What imputation techniques do you recommend?
Answer: There are 3 methods to handle missing handle:

1. Mean or median imputation: When data is missing at random, we can use list-wise or pair-wise deletion of the missing observations. In such cases, we impute values for missing data. A common technique is to use the mean or median of the non-missing observations. This can be useful in cases where the number of missing observations is low. However, for large number of missing values, using mean or median can result in loss of variation in data and it is better to use imputations.
2. Multivariate Imputation by Chained Equations MICE: MICE assumes that the missing data are missing at Random MAR. It imputes data on a variable-by-variable basis by identifying an imputation model per variable. MICE uses predictive mean matching method PMM for continuous variables, logistic regressions for binary variables, bayesian polytomous regressions for factor variables, and proportional odds model for ordered variables to impute missing data.
   To set up the data for MICE, it is important to note that the algorithm uses all the variables in the data for predictions. In this case, variables that may not be useful for predictions, like the ID variable, should be removed before implementing this algorithm.
3. Random Forest: Random forest is a non-parametric imputation method which is applicable to various variable types that works well with both data missing at random and not missing at random. Random forest uses multiple decision trees to gauge missing values and outputs OOB i.e. out of bag imputation error estimates.
   So, random forest works best with large datasets and using random forest on small datasets runs the risk of overfitting. The extent of overfitting leading to incorrect imputations will depend upon how closely the distribution for predictor variables for non-missing data resembles the distribution of predictor variables for missing data.

   Imputation Technique
   I recommend Complete Case Analysis technique, as is is quite direct method of handling the missing data, which straighjforward removes the missing rows of the data, that is we consider only those rows where we have entire data which means data is not missing. So, this is also known as Listwise deletion. This method is used if the data is completely removed from the table. It is very easy to handle and implement, and there is no need of data manipulation required in this technique.

12. What is A/B testing?
Answer: An AB testing is a primary randomized control experiment. AB testing is used to compare the two versions of a variable to detect which performs better in a controlled environment. So, is an example of statistical hypothesis testing, a process in which a hypothesis is made about the relationship between two data sets and those data sets are then compared against each other to find out if statistically significant relationship is there or not.

13. Is mean imputation of missing data acceptable practice?
Answer: Yes, imputing the mean safeguard the mean of the observed data but it’s the bad practice in general. So, if the data are randomly missing completely, then the estimate of the mean remains impartial or unbiased. It also leads to an underestimate of the standard deviation, and deform relationships between variables by pulling estimates of the correlation toward 0.

14. What is linear regression in statistics?
Answer: Linear regression in statistics is basic basic and frequently used type of predictive analysis.

Basically, its entire idea is to examine whether a set of predictor variables do a good job in predicting an outcome that is dependent variable and which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable. Linear regression estimates are used to explain the relationship between one dependent variable and one or more than one independent variables. The simplest form of the regression equation is defined by the formula y = c + b*x, where y is estimated dependent variable score, c is constant, and b is regression coefficient, and x is score on the independent variable.

15. What are the various branches of statistics?

Answer: Statistics plays the key role in the field of research, it helps in collecting, analyzing and presenting the data. There are two main branches of statistics:

1. Descriptive Statistics: As its name suggest, it describes the data that is already known. It organize, analyze and present the data in the meaningful manner, as it is more concerned about the describing the target population. So, the outcomes are shown in two basic forms: presenting aspects of the data either visually via graphs, charts, etc. or numerically via averages or mean/median/mode and so on. The basic aim of descriptive statistics is to ‘present the data’ in an understandable way.
2. Inferential Statistics: Inferential statistics, make inferences or conclusions from the data we have and generalize them to the population. So, it compares, test and predicts the future results or outcomes, so its results are like the probability scores. Basically, it tries to make inferences about the population that is beyond the sample available. Tools used for this are Analysis of variance, hypothesis tests, etc.