# 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?
11. How do you handle missing data? What imputation techniques do you recommend?
12. What is A/B testing?
13. Is mean imputation of missing data acceptable practice?
14. What is linear regression in statistics?
15. What are the various branches of statistics?

Ans 10: Normal Distribution is a probability distribution that is symmetric about the mean which shows the data near the mean . Graphically Normal distribution is in bell curve shape. In this shape , Half of the data will fall to left of the mean and half of the data will fall to the right. Normallly we convert normal distribution into the standard normal distribution for many reasons like to find the probability of observations in a distribution falling above and below a given value.

Ans 11. If there is nissing data , we can handle by using deletion methods to eliminate missing data. This method only work for certain datasets where participants have missing fields.Another we can use regression analysis to systematically eliminate data. Data scientists can use many data imputation techniques, but the best technique is:

1) By ignoring the records with missing values: Many tools ignore records with missing values. When the percentage of records with missing values is small, we could ignore those records.
2) By Substituting the value such as Mean: When the percentage is large and also when it makes sense to do something to avoid bias modelling results , substituting a value is commonly used way.
3) By predicting missing values: It depends on the type of the imputed variable and missing data pattern . If we plan to do it in SAS, there are SAS codes thet we can write to identify the missing data pattern.

Ans 12. A/B testing is that tool which is used to compare a webpage or application withdifferent variants to determine which is better for example : when a webpage is modified to create any changes made to it , This change can be as simple as a single headline or button, or be a complete redesign of the page
With the help of this, half of the traffic is shown the original version of the page and half of the are shown the modified version of the page. The engagement of the users with both the classes of webpages is measured and collected in an analytics dashboard. Now, The changes in user experience , whether positive or negative can be determined.

Ans 13. Yes, That is true. Imputing the mean secures the mean of the observed data . So, if the data are missing completely at random, the estimate of the mean remains unbiased. Outliers data points will have a significant impact on the mean and in that cases, it is not recommended to use the mean for replacing the missing values. Using mean values for replacing missing values may not create a great model and hence gets ruled out.

Ans 14. First thing, Linear Regression is the most widely used statistical technique . It is a way to model a relationship between two sets of variables. The result is a linear regression equation that can be used to make predication about data . Linear means Line and linear relationaships are easier to work with and most phenomenon are naturally linearly related. If variables are not linearly related, then some math can transform that relationship into linear , so that it is easier for the researcher to understand.

Ans 15.There are two branches of statistics . One is Descriptive and another one is Inferential statistics. Both of these are used in scientific analysis of data and both are equally important for the sattistics. Descriptive statistics deals with the presentation and collection of the data and this is used with designing experiments, choosing right focus group and avoid biases where as Inferential Statistic involves drawing the right conclusion from the statistical analysis that has been performed using descriptive statistics. Lastly we can say that both branches go side by side and one cannot exist without the other. .

FLIP ROBO