**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**

2**. Which of the following theorem states that the distribution of averages of id variables, properly normalized, becomes that of a standard normal as the sample size increases?**

**a) Central Limit Theorem**

3. Which of the following is incorrect with respect to use of Poisson distribution?

**b) Modelling 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**

**5. \_\_\_\_\_\_ random variables are used to model rates**

**c) Poisson**

**6. Usually replacing the standard error by its estimated value does change the CLT.**

**a) True**

**7. Which of the following testing is concerned with making decisions using data?**

**b) Hypothesis**

**8. Normalized data are centered at\_\_0\_\_\_\_and have units equal to standard deviations of the original data.**

**a) 0**

**9. Which of the following statement is incorrect with respect to outliers**

**b) Outliers can be the result of spurious or real processes**

**c) Outliers cannot conform to the regression relationship**

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

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

 It represents a**symmetric distribution** where most of the observations cluster around the central peak called as **mean** of the distribution.

A normal distribution can be thought of as a bell curve or Gaussian Distribution which typically has two parameters: **mean** and **standard deviation (SD)**.  The parameter used to measure the variability of observations around the mean is called as **standard deviation**. The probabilities for values occurring near mean are higher than the values far away from the mean. The **parameters** of the normal distribution plot defining the **shape** and the probabilities are **mean**and **standard deviation.** The area of the plot between two different points in the normal distribution plot represents the probability of the value occurring between those two points.

While training machine learning models, the normal distribution is very important because it helps normalize the data. Many machine learning algorithms assume normal distribution in the data. If you have normally distributed inputs, use a normal probability function to calculate them. If your inputs are not normally distributed, transform them by applying log or square root transformations until they become normally distributed before feeding into an algorithm that assumes normal distribution (such as linear regression).

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

The deletion methods only work for certain datasets where participants have missing fields. There are several deleting methods – two common ones include Listwise Deletion and Pairwise Deletion. It means deleting any participants or data entries with missing values.

This method is particularly advantageous to samples where there is a large volume of data because values can be deleted without significantly distorting readings. Alternatively, data scientists can fill out the missing values by contacting the participants in question. The problem with this method is that it may not be practical for large datasets.

Furthermore, some corporations obtain their information from third-party sources, which only makes it unlikely that organisations can fill out the gaps manually. Pairwise deletion is the process of eliminating information when a particular data point, vital for testing, is missing. Pairwise deletion saves more data compared to likewise deletion because the former only deletes entries where variables were necessary for testing, while the latter deletes entire entries if any data is missing, regardless of its importance.

**12. What is A/B testing?**

A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment. A/B tests, also known as split tests, allow you to compare 2 versions of something to learn which is more effective. Simply put, do your users like version A or version B? The concept is similar to the scientific method. If you want to find out what happens when you change one thing, you have to create a situation where only that one thing changes.

**13. Is mean imputation of missing data acceptable practice?**

True, mean imputation is not a good solution.

**14. What is linear regression in statistics?**

Linear regression is a basic and commonly used type of predictive analysis.  The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable?  (2) 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?  These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables.

 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.

**15. What are the various branches of statistics?**

There are three real branches of statistics:

1. Data collection
2. Descriptive statistics
3. Inferential statistics