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

a) True

b) False

**Ans-** b) False

7. 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.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

**WORKSHEET**

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

**Ans-**Normal distribution is the most important probability distribution in statistics. It is also known as Gaussian distribution and bell curve. It describes how the values of a variable are distributed. It is a symmetric distribution where most of the observations cluster around the central peak and the probabilities for values further away from mean taper off equally in both directions.

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

**Ans-** There are three main approaches to handle a missing data in statistics.They are

i)Imputation-Here values are filled in place of missing data

ii)Omission-samples with invalid data are discarded from further analysis

iii)Analysis-By directly applying methods unaffected by the missing values.

There are different imputation techniques to handle this.They are:

i)Imputation using mean/median –here missing values are replaced by calculating the mean /median of the non-missing values in each column

ii)imputation using most frequent values or zero–Here missing values are either replaced by the most frequent value in each column or replaced by zeros.

iii)Imputation using KNN- The new point is assigned a value based on how closely it resembles the points in the training set. Missing values are predicted by k’s closest neighbours to the observation with missing data.

12. What is A/B testing?

**Ans-** A/B test is an example of statistical hypothesis testing. A process where a hypothesis is made about the relationship between two datasets and those datasets are then compared against each other to determine if there is statistically significant relationship or not.

13. Is mean imputation of missing data acceptable practice?

**Ans-** It is a non-standard, but fairly flexible imputation algorithm. Mean imputation reduces the variance of the imputed variables. It shrinks the standard error which invalidates most hypothesis tests and the calculation of confidence interval. Also it doesn’t preserve the relationship between variables such as correlation.

14. What is linear regression in statistics?

**Ans-**It is a linear approach to modelling the relationship between a scalar response and one or more explanatory variables. The case of one explanatory variable is called simple linear regression and for more than one it is called multiple linear regression.

15. What are the various branches of statistics

**Ans-**The two main branches of statistics are descriptive statistics and inferential statistics. Descriptive statistics deals with the presentation and collection of data. Inferential statistics involves drawing the right conclusions from the statistical analysis performed using descriptive statistics. Good scientific methodology needs to be followed in both these steps of statistical analysis and both these branches are equally important.