**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 iid 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) Modeling bounded count data

4. Point out the correct statement.

d) All of the mentioned

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

c) Poisson

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

b) False

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

b) Hypothesis

8. 4. Normalized data are centered at\_\_\_\_\_\_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?

c) Outliers cannot conform to the regression relationship

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

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

Normal distribution, also known as the Gaussian distribution, is a  [probability distribution](https://www.investopedia.com/terms/p/probabilitydistribution.asp) that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graph form, normal distribution will appear as a [bell curve](https://www.investopedia.com/terms/b/bell-curve.asp).

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

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

Imputaion Techniques

Mean or Median Imputation

## Multivariate Imputation by Chained Equations (MICE)

## Random Forest

12. What is A/B testing?

A/B testing also known as split testing, refers to a randomized experimentation process wherein two or more versions of a variable (web page, page element, etc.) are shown to different segments of website visitors at the same time to determine which version leaves the maximum impact and drive business metrics.

Essentially, A/B testing eliminates all the guesswork out of [website](https://vwo.com/website-optimization/) optimisation and enables experience optimizers to make data-backed decisions. In A/B testing, A refers to ‘control’ or the original testing variable. Whereas B refers to ‘variation’ or a new version of the original testing variable.

13. Is mean imputation of missing data acceptable practice?

There are a few advantages, but many serious drawbacks

1. Mean substitution leads to **bias in multivariate estimates** such as correlation or regression coefficients. Values that are imputed by a variable’s mean have, in general, a correlation of zero with other variables. Relationships between variables are therefore biased toward zero.
2. **Standard errors and variance** of imputed variables are biased. For instance, let’s assume that we would like to calculate the standard error of a mean estimation of an imputed variable. Since all imputed values are exactly the mean of our variable, we would be too sure about the correctness of our mean estimate. In other words, the [confidence interval](https://www.mathsisfun.com/data/confidence-interval.html) around the point estimation of our mean would be too narrow.
3. If the [response mechanism is MAR or MNAR](https://statisticsglobe.com/missing-data/#response-mechanisms), even the **sample mean of your variable is biased** (compare that with point 3 above). Assume that you want to estimate the mean of a population’s income and people with high income are less likely to respond; Your estimate of the mean income would be biased downwards.

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**two main branches** of statistics  
- Inferential Statistic.  
- Descriptive Statistic.

**Inferential Statistics:**  
Inferential statistics used to make inference and describe about the population. These stats are more useful when its not easy or possible to examine each member of the population.

**Descriptive Statistics:**  
Descriptive statistics are use to get a brief summary of data. You can have the summary of data in numerical or graphycal form.