# STATISTICS WORKSHEET-1

## Answer Sheet:

1.Bernoulli random variables take (only) the values 1 and 0.

**Ans:** 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?

**Ans:** central limit theorem

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

**Ans:** Modeling bounded count data

4. Point out the correct statement

**Ans:** All of the mentioned

5. _____ random variables are used to model rates.

**Ans:** Poisson

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

**Ans:** False

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

**Ans:** Hypothesis

8. Normalized data are centered at_____and have units equal to standard deviations of the original data.

**Ans:** 0

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

**Ans:** Outliers cannot conform to the regression relationship

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

**Ans:** As with any probability distribution, the normal distribution describes how the values of a variable are distributed. It is the most important probability distribution in statistics because it accurately describes the distribution of values for many natural phenomena.

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

**Ans:** a) Use deletion methods, Pair wise Deletion. List wise Deletion/ Dropping rows.

b) Use regression analysis to systematically eliminate data.

c) Using data imputation techniques

Types of Imputation technique,

1. Complete Case Analysis(CCA): This is a quite straight forward method of handling the Missing Data, which directly removes the rows that have missing data

2. Arbitrary Value Imputation.

3. Frequent Category Imputation

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

**Ans:** 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 Testing process:

1. Start the A/B testing process by making a claim (hypothesis).
2. Launch your test to gather statistical evidence to accept or reject a claim (hypothesis) about your website visitors.
3. The final data shows you whether your hypothesis was correct, incorrect or inconclusive.

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

**Ans:** Not Acceptable,

- Bad practice in general.

- If just estimating means: mean imputation preserves the mean of the observed data.

- Leads to underestimate of the standard deviation.

- Distorts relationship between variables by "pulling" estimates of the correlation toward zero.

## 14. What is linear regression in statistics?

**Ans:** Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable

Linear regression method can be performed in a variety of programs and environments,
- R linear regression
- MATLAB linear regression
- Sklearn linear regression
- Linear regression Python
- Excel linear regression

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

**Ans:** a) Data collection.
　　b) Descriptive statistics.
　　c) Inferential statistics.

Data collection: is all about how the actual data is collected. There are significant issues to be considered when actually collecting data.
For data such as marks in a class test, this is fairly straightforward but sometimes, data is harder to collect.

Descriptive statistics: is the part of statistics that deals with presenting the data we have. This can take two basic forms – presenting aspects of the data either visually (via graphs, charts, etc.) or numerically (via averages and so on).

Inferential statistics: is the aspect that deals with making conclusions about the data. This is quite a wide area, essentially you are asking 'What is this data telling us, and what should we do?'