# Statistics worksheet 1

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

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

Ans.

Loss of data is an everyday problem that a data professional needs to deal with. Missing data is defined as a missing value, Missing data can be from anything missing sequence Incomplete feature, missing files, incomplete information, data entry error etc.

**Handle missing data -** In the case of multivariate analysis, if there are a large number of missing values, it may be better to discard those cases (rather than impute) and replace them. On the other hand, in univariate analysis, imputation can reduce the amount of bias in the data if the values are missing at random.

**There are two forms of randomly missing values:**

- MCAR: Missing completely at random – The first form is completely at random (MCAR) missing. This form exists when the missing values are randomly distributed across all observations.
- MAR: Missing at random - That second form is missing at random (MAR). In MAR, the missing values are not randomly distributed across observations but rather in one or more sub-samples. This form is more common than the previous one.

**Imputation techniques –**

I. Complete Case Analysis(CCA)- This is a very simple way to handle missing data, which directly deletes those rows in which data is missing i.e. we consider only those rows where we have complete data i.e. data is not missing. This method is also popularly known as "list wise deletion".

When to Use:-

- Data is MAR(Missing At Random).
- Good for Mixed, Numerical, and Categorical data.
- Missing data is not more than 5% – 6% of the dataset.
- Data doesn't contain much information and will not bias the dataset.

II. Arbitrary Value Imputation- This is an important technique used in Imputation as it can handle both the Numerical and Categorical variables. This technique states that we group the missing values in a column and assign them to a new value that is far away from the range of that column. Mostly we use values like 99999999 or -9999999 or "Missing" or "Not defined" for numerical & categorical variables.

When to Use:-

- When data is not MAR(Missing At Random).
- Suitable for All.

III. Frequent Category Imputation- This technique says to replace the missing value with the variable with the highest frequency or in simple words replacing the values with the Mode of that column. This technique is also referred to as Mode Imputation.

When to Use:-

- Data is Missing at Random(MAR)
- Missing data is not more than 5% – 6% of the dataset.