STATISTICS

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. _______ randomvariables are used to model rates.

   c) Poisson

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

   a) True

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

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

    Normal Distribution is also called a bell shaped cure, in this case all three mean , median and mode are the same. It is a symmetrical distribution where there is no variance. Here the 3 sigma rule applies which states that 68% of the data lies between (m+_1 sd), 95% of the data lies between (m+_2 sd) and 99.7%of the data lies between (m+_3 sd). Here roughly 0.3% of the data can be outside the curve, they are referred to as outliers.

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

Missing Data can be handled by using Simple Imputer Technique which works in two ways, for numeric values both mean and median values of the respective column are accepted where as for string columns, mode of the specific column are accepted.

Simple Imputer is available under sklearn library as

```
Import sklearn
```

```
From sklearn.impute import SimpleImputer
```

Suppose the Table name is FG and the columns which we are focusing on is called HEIGHT (numeric figures) a few values are missing

```
Si=SimpleImputer(missing_values=np.nan,strategy=’mean’)
```

```
FG[‘HEIGHT’]=Si.fit_transform(FG[‘HEIGHT’].values)
```

12. What is A/B testing?

It is a tool used to analyze and interpret the data in inferential statistics.

There are 2 Hypothesis:

Null and Alternate Hypothesis:

Null Hypothesis is you make an assumption for ex avg consumption of rice of people are more than 200 gm daily and then you create another hypothesis, the alternate hypothesis stating the opposite.

With various tools like T-Test, Chi Square Test, Annova etc a p value is calculated.

If the P value is greater than 0.05 , Null Hypothesis is true and if P vale is less than 0.05 then Alternate Hypothesis is true .

STATISTICS

13. Is mean imputation of missing data acceptable practice?

Mean Imputation of missing data is acceptable only when the number of missing values in the entire column are few, if there are many missing values then its better to drop the column.

14. What Is linear regression in statistics?

Linear regression is a tool which helps in determining the predictions of a dependent variable (y) based on the independent variables (x1),(x2),(xn) etc.

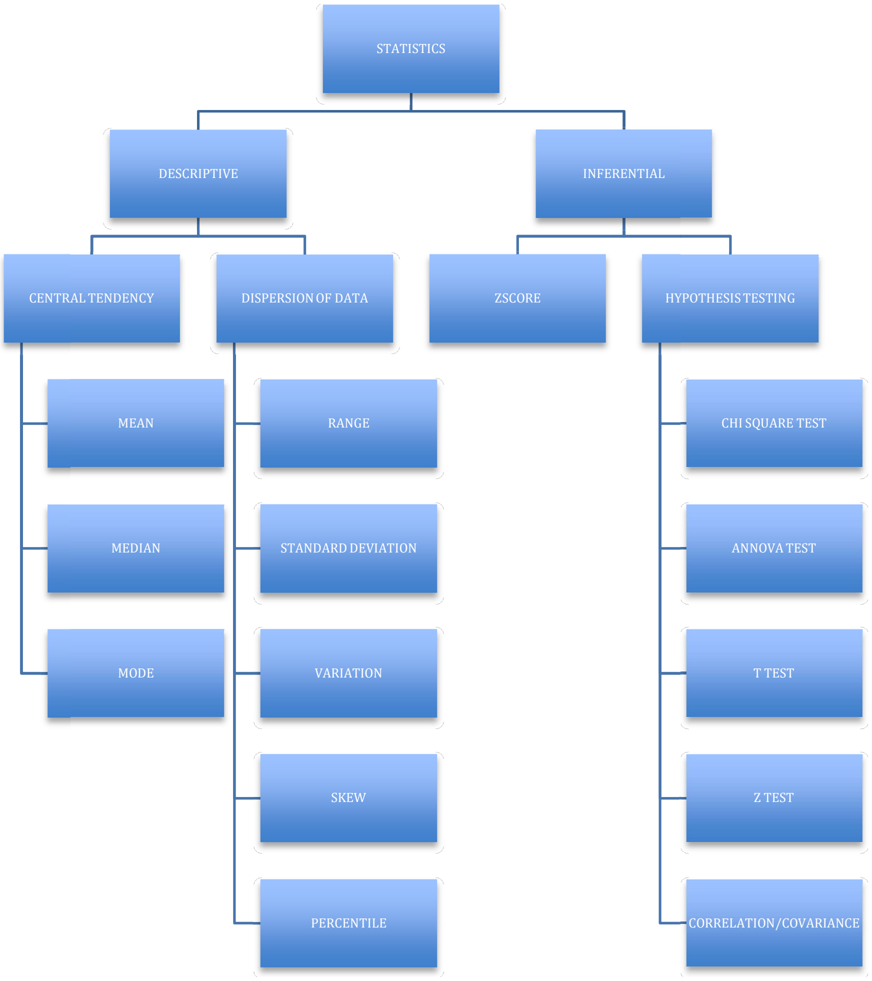
The formula is : Y=A+BX1+BX2+BX3+BXN+C

Where A is the intercept and B is the co-efficient of x. if x changes in one unit , how much change there is in y and c denotes the error.

For error calculation we use the least square method

15. What are the various branches of statistics

STATISTICS


STATISTICS
DESCRIPTIVE
INFERENTIAL
CENTRAL TENDENCY
DISPERSION OF DATA
ZSCORE
HYPOTHESIS TESTING
MEAN
MEDIAN
MODE
RANGE
STANDARD DEVIATION
VARIATION
SKEW
PERCENTILE
CHI SQUARE TEST
ANNOVA TEST
T TEST
Z TEST
CORRELATION/COVARIANCE