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

Ans:-b)False

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

Ans:- b) Hypothesis

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

Ans:- 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?

Ans:- Normal distribution, also known as the Gaussian distribution, is a probability distribution 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. In a normal distribution the mean is 0 and standard deviation is 1 . The normal distribution is the most common type of distribution assumed in technical stock market analysis and in other types of statistical analyses.

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

There are 3 types of missing data :-

- **a.** Missing Completely At Random
- **b.** Missing at Random
- **c.** Not missing at random

And there are also 3 techniqes to handle missing data:-

- a. Mean imputation
- b. Multivariate Imputation by Chained Equations (Mice)
- c. Random Forest

Random forest is a non-parametric imputation method applicable to various variable types that works well with both data missing at random and not missing at random. Random forest uses multiple decision tree to estimate missing values and outputs OOB (out of bag) imputation error estimates. So I will always recommend this technique.

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 optimization 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?

- Bad practice in general
- If just estimating means: mean imputation preserves the mean of the observed data
- Leads to an underestimate of the standard deviation

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 = m*x+c, where y = estimated dependent variable score, c = constant, m= regression coefficient, and x = score on the independent variable. Three major uses for regression analysis

are: (1) determining the strength of predictors, (2) forecasting an effect, and (3) trend forecasting.

15. What are the various branches of statistics?

Statistics is a branch of science that deals with the ability to grasp different outcomes from it .

It basically is the study of the following features:

1. Data collection

2.Classification

3.Analysis

4.Interpretation

5.Presentation

There are majorly two types of statistics:

A)Descriptive statistics:- Descriptive Statistics is used to summarise the data through a given set of observations and illustrations. This type of statistics deals with the organization and representation of data. It also describes the given collection of data in the form of tables, graphs, and summary measures.

Descriptive statistics can also be divided into four different categories, which are as follows:
Measure of frequency: Indicator of the number of times a particular data occurs.
Measure of dispersion: Range, variance, and standard deviation are measures of dispersion of the data.
***Measure of central tendency*:** Mean, median, and mode are the measure of central tendency of the given data distribution.
Measure of position: The measure of position describes the percentile and quartile ranks.

B)Inferential statistics:- The data collection and analysis are followed by the interpretation of the collected facts which is done by using inferential statistics. Inferential Statistics is dependent on the following random variations: 1. Observational errors 2. Sampling variation