1. Bernoulli random variables take (only) the values 1 and 0.
a) True
b) False

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?
a) Central Limit Theorem
b) Central Mean Theorem
c) Centroid Limit Theorem
d) All of the mentioned

Ans. a) Central Limit Theorem

3. Which of the following is incorrect with respect to use of Poisson distribution?
a) Modeling event/time data
b) Modeling bounded count data
c) Modeling contingency tables
d) All of the mentioned

Ans. b) Modeling bounded count data

4. Point out the correct statement.
a) The exponent of a normally distributed random variables follows what is called the log- normal distribution
b) Sums of normally distributed random variables are again normally distributed even if the variables are dependent
c) The square of a standard normal random variable follows what is called chi-squared distribution
d) All of the mentioned

Ans. d) All of the mentioned

5. _____ random variables are used to model rates.
a) Empirical
b) Binomial
c) Poisson
d) All of the mentioned

Ans. c) Poisson

6. Usually replacing the standard error by its estimated value does change the CLT.
a) True
b) False

Ans. b) False

7. Which of the following testing is concerned with making decisions using data?
a) Probability
b) Hypothesis
c) Causal
d) None of the mentioned

Ans. b) Hypothesis

8. Normalized data are centered at_____and have units equal to standard deviations of the original data.
a) 0
b) 5
c) 1
d) 10

Ans. a) 0

9. Which of the following statement is incorrect with respect to outliers?
a) Outliers can have varying degrees of influence
b) Outliers can be the result of spurious or real processes
c) Outliers cannot conform to the regression relationship
d) None of the mentioned

Ans. c) Outliers cannot conform to the regression relationship

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

Ans. A constant proportion of data points lying under the curve between the mean and a specific number of standard deviations from the mean is called Normal distribution.

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

Ans. Depending upon the nature of the missing data, we use different techniques to impute data such as :-

a. **Mean or median imputation**
   When data is missing at random, we can use list-wise or pair-wise deletion of the missing observations. However, there can be multiple reasons –
- There may not be enough observations with non-missing data to produce a reliable analysis
- In predictive analytics, missing data can prevent the predictions for those observations which have missing data
- External factors may require specific observations to be part of the analysis

b. **Multivariate Imputation by Chained Equations (MICE)**
   MICE assumes that the missing data are Missing at Random (MAR). It imputes data on a variable-by-variable basis by specifying an imputation model per variable. MICE uses predictive mean matching (PMM) for continuous variables, logistic regressions for binary variables, bayesian polytomous regressions for factor variables, and proportional odds model for ordered variables to impute missing data

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 trees to estimate missing values and outputs OOB (out of bag) imputation error estimates.

12. What is A/B testing?

Ans. A/B testing is 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.

There are some techniques that we can use for A/B testing :-

   a. Making a hypothesis
   In hypothesis, there are two types of hypothesis such as :-

      i.   Null hypothesis or H_0 :
      The null hypothesis is the one that states that sample observations result purely from chance. From an A/B test perspective, the null hypothesis states that there is no difference between the control and variant groups. It states the default position to be tested or the situation as it is now, i.e. the status quo. Here our H_0 is " there is no difference in the conversion rate in customers receiving newsletter A and B".

      ii.   Alternative hypothesis or H_0

      The alternative hypothesis challenges the null hypothesis and is basically a hypothesis that the researcher believes to be true. The alternative hypothesis is what you might hope that your A/B test will prove to be true.

   b. Conduct the A/B test & collect the data
   One way to perform the test is to calculate daily conversion rates for both the treatment and the control groups. Since the conversion rate in a group on a certain day represents a single data point, the sample size is actually the number of days. Thus, we will be testing the difference between the mean of daily conversion rates in each group across the testing period.

   When we run our experiment for one month, we noticed that the mean conversion rate for the Control group is 16% whereas that for the test Group is 19%

13. Is mean imputation of missing data acceptable practice?

Ans. Mean imputation is one such method in which the mean of the observed values for each variable is computed and the missing values for that variable are imputed by this mean

14. What is linear regression in statistics?

Ans. 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?

Ans. There are two types of branches in Statistics :-
   a. Descriptive Statistics
   It deals with the presentation and collection of data. This is usually the first part of a statistical analysis. It is usually not as simple as it sounds, and the statistician needs to be aware of designing experiments, choosing the right focus group and avoid biases that are so easy to creep into the experiment.

b. Inferential Statistics

Inferential Statistics allows you to make predictions from that data. With inferential statistics, we take data from samples and make generalizations about a population.