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**
   b) Poisson

6. **Usually replacing the standard error by its estimated value does change the CLT**
   c) False

7. **Which of the following testing is concerned with making decisions using data?**
   b) Hypothesis

8. **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?**
   d) Outliers cannot conform to the regression relationship

10. **What do you understand by the term Normal Distribution?**
    The normal distribution, also known as the Gaussian or standard normal distribution, is the probability distribution that plots all of its values in a symmetrical fashion, and most of the results are situated around the probability's mean. Values are equally likely to plot either above or below the mean.

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

The most common ways of handling missing data

- We can replace the missing value with zero or most frequent data irrespective of everything.
- We can replace the missing value with the minimum or maximum value of a feature.
- If it's time-series data, we can fill it with the previous value

- We can replace missing value with mean or median or most frequent feature value.

- By using KNN- k nearest neighbours is an algorithm that is used for simple classification. The algorithm uses feature similarity to predict the values of any new data points.

- We can replace the value of the missing cell with the previous cell's value. This kind of technique is popular while inputting time series data.

- By using Multivariate Imputation - because Multiple Imputations are much better than a single imputation as it measures the uncertainty of the missing values in a better way. The chained equations approach is also very flexible and can handle different variables of different data as well as complexities such as bounds or survey skip patterns.

But there is no perfect way to compensate for the missing values in a dataset. Each strategy can perform better for certain datasets and missing data types it depends on types of datasets. There are some set rules to decide which strategy to use for particular types of missing values, but beyond that, we need to experiment and check which model works best for particular dataset.

12. What is A/B testing?

A/B testing in its simplest sense is an experiment on two variants to see which performs better based on a given metric.

For example if we own a company and want to increase the sales of our product. Here, either we can use random experiments, or we can apply scientific and statistical methods. A/B testing is one of the most prominent and widely used statistical tools.

Before conducting an A/B testing we need to follow few steps

1-Make a Hypothesis

null hypothesis is 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 group.

alternative hypothesis is one that states that sample observations are influenced by some non-random cause. From an A/B test perspective, the alternative hypothesis states that there is a difference between the control and variant group.

While developing null and alternative hypotheses, we should follow a PICOT format.

Population: the group of people that participate in the experiment

- Intervention: refers to the new variant in the study
- Comparison: refers to what you plan on using as a reference group to compare against your intervention
- Outcome: represents what result you plan on measuring
- Time: refers to the duration of the experience (when and how long the data is collected)
- Once we are ready with our null and alternative hypothesis, the next step is to decide the group of customers

2. **Create Control Group and Test Group**

- we have two groups – The Control group, and the Test group
- Randomly selecting the sample from the population is called random sampling. It is a technique where each sample in a population has an equal chance of being chosen. Random sampling is important in hypothesis testing because it eliminates sampling bias, and it's important because we want the results of A/B test to be representative of the entire population rather than the sample itself.
- Another important aspect is the Sample size. In this we will take the minimum sample size for our A/B test before conducting it so that we can eliminate under coverage bias. It is the bias from sampling too few observations.

3. **Conduct the test, compare the results, and reject or do not reject the null hypothesis**

- Once we conduct our experiment and collect our data, if we want to determine the difference between our control group and variant group is statistically significant. There are a few steps in determining this:
- First, we want to set our alpha, the probability of making a type 1 error. Typically the alpha is set at 5% or 0.05
- Next, we want to determine the probability value (p-value) by first calculating the t-statistic using the formula above.
- At the end compare the p-value to the alpha. If the p-value is greater than the alpha, do not reject the null!

13. Is mean imputation of missing data acceptable practice?

- mean imputation is depends on data and scenario.

14. What is linear regression in statistics?

Linear regression is the simplest and most extensively used statistical technique for predictive modelling analysis. It is a way to explain the relationship between a dependent variable (target) and one or more explanatory variables (predictors) using a straight line. Linear regression is only dealing with continuous variables instead of Bernoulli variables.

Continuous variables - variables that can take on any value within a range.

Bernoulli variables- finite or infinite sequence of binary random variables which takes only two values 0 or 1.

15. What are the various branches of statistics?

There are two branches of statistics are descriptive statistics and inferential statistics.

Descriptive Statistics:- Its considered as the first part of statistical analysis which deals with collection and presentation of data. Descriptive statistics can be categorized into

- **Measures of central tendency**- measures of tendency are:

  Mean

  Median

  Mode

- **Measures of variability**- measure of variability help statisticians to analyse the distribution spread out of a given set of data like measures of variability include quartiles, range, variance and standard deviation.

Inferential Statistics:- Inferential statistics are techniques that enable statisticians to use the gathered information from a sample to make inferences, decisions or predictions about a given population. Inferential statistics often talks in probability terms by using descriptive statistics. Different types of calculation of inferential statistics include:

- Regression analysis
- Analysis of variance (ANOVA)
- Analysis of covariance (ANCOVA)
- Statistical significance (t-test)
- Correlation analysis