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.

b) 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?

c) Outliers cannot conform to the regression relationship

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

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 graphical form, the normal distribution appears as a "bell curve".

- he normal distribution is the proper term for a probability bell curve.
- In a normal distribution the mean is zero and the standard deviation is 1. It has zero skew and a kurtosis of 3.
- Normal distributions are symmetrical, but not all symmetrical distributions are normal.
- Many naturally-occurring phenomena tend to approximate the normal distribution.
- In finance, most pricing distributions are not, however, perfectly normal.

The normal distribution is the most common type of distribution assumed in technical stock market analysis and in other types of statistical analyses. The standard normal distribution has two parameters: the mean and the standard deviation. The normal distribution model is important in statistics and is key to the Central Limit Theorem (CLT). This theory states that averages calculated from independent, identically distributed random variables have approximately normal distributions, regardless of the type of distribution from which the variables are sampled (provided it has finite variance). The normal distribution is one type of symmetrical distribution. Symmetrical distributions occur when where a dividing line produces two mirror images. Not all symmetrical distributions are normal, since some data could appear as two humps or a series of hills in addition to the bell curve that indicates a normal distribution.

The Formula for the Normal Distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where:

- x = value of the variable or data being examined and f(x) the probability function
- μ = the mean
- σ = the standard deviation

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

Real-world data is messy and usually holds a lot of missing values. Missing data can skew anything for data scientists and, A data scientist doesn’t want to design biased estimates that point to invalid results. Behind, any analysis is only as great as the data. **Missing data appear**

when no value is available in one or more variables of an individual. Due to Missing data, the statistical power of the analysis can reduce, which can impact the validity of the results.

Deletion: The Deletion technique deletes the missing values from a dataset. followings are the types of missing data .

List wise deletion: List wise deletion is preferred when there is a Missing Completely at Random case. In List wise deletion entire rows(which hold the missing values) are deleted. It is also known as complete-case analysis as it removes all data that have one or more missing values. List wise deletion is not preferred if the size of the dataset is small as it removes entire rows if we eliminate rows with missing data then the dataset becomes very short and the machine learning model will not give good outcomes on a small data set.

Pair wise Deletion: Pai rwise Deletion is used if missingness is missing completely at random i.e MCAR. Pair wise deletion is preferred to reduce the loss that happens in List wise deletion. It is also called an available-case analysis as it removes only null observation, not the entire row. All methods in pandas like mean, sum, etc. intrinsically skip missing values.

Dropping complete columns If a column holds a lot of missing values, say more than 80%, and the feature is not meaningful, that time we can drop the entire column.

Imputation techniques:

The imputation technique replaces missing values with substituted values. The missing values can be imputed in many ways depending upon the nature of the data and its problem. Imputation techniques can be broadly they can be classified as follows:

Imputation with constant value: As the title hints — it replaces the missing values with either zero or any constant value.

Imputation using Statistics: The syntax is the same as imputation with constant only the Simple Imputer strategy will change. It can be “Mean” or “Median” or “Most Frequent”

“Mean” will replace missing values using the mean in each column. It is preferred if data is numeric and not skewed. “Median” will replace missing values using the median in each column. It is preferred if data is numeric and skewed. Most frequent” will replace missing values using the most frequent in each column. It is preferred if data is a string(object) or numeric.

Before using any strategy, the foremost step is to check the type of data and distribution of features (if numeric).

Advanced Imputation Technique: Unlike the previous techniques, Advanced imputation techniques adopt machine learning algorithms to impute the missing values in a dataset. Followings are the machine learning algorithms that help to impute missing values.

K_Nearest Neighbor Imputation: The KNN algorithm helps to impute missing data by finding the closest neighbors using the Euclidean distance metric to the observation with missing data and imputing them based on the non-missing values in the neighbors.

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. Typically, two consumer groups are exposed to two different versions of the same thing to see if there is a significant difference in metrics like sessions, click-through rate, and/or conversions. Using the visual above as an example, we could randomly split our customer base into two groups, a control group and a variant group. Then, we can expose our variant group with a red website banner and see if we get a significant increase in conversions. It's important to note that all other variables need to be held constant when performing an A/B test. Getting more technical, A/B testing is a form of statistical and two-sample hypothesis testing. Statistical hypothesis testing is a method in which a sample dataset is compared against the population data. Two-sample hypothesis testing is a method in determining whether the differences between the two samples are statistically significant or not.

Formulate your hypothesis

Before conducting an A/B testing, you want to state your null hypothesis and alternative hypothesis: The 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. The 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. When developing your null and alternative hypotheses, it's recommended that you follow a PICOT format. Picot stands for:

- **P**opulation: the group of people that participate in the experiment
- **I**ntervention: refers to the new variant in the study
- **C**omparison: refers to what you plan on using as a reference group to compare against your intervention
- **O**utcome: represents what result you plan on measuring
- **T**ime: refers to the duration of the experience (when and how long the data is collected)

Create your control group and test group Once you determine your null and alternative hypothesis, the next step is to create your control and test (variant) group. There are two important concepts to consider in this step, random samplings and sample size.

Random Sampling Random sampling 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 to eliminate bias because you want the results of your A/B test to be representative of the entire population rather than the sample itself.

Sample Size It’s essential that you determine the minimum sample size for your A/B test prior to conducting it so that you can eliminate under coverage bias, bias from sampling too few observations. There are plenty of online calculators that you can use to calculate the sample size given these three inputs.

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

$$T - statistic = \frac{Observed\ value - hypothesized\ value}{Standard\ Error}$$

$$Stamdard\ Error = \sqrt{\frac{2 * Variance(sample)}{N}}$$

Once you conduct your experiment and collect your data, you want to determine if the difference between your control group and variant group is statistically significant.

13. Is mean imputation of missing data acceptable practice?

The process of replacing null values in a data collection with the data’s mean is known as mean imputation.

Mean imputation is typically considered terrible practice since it ignores feature correlation. Consider the following scenario: we have a table with age and fitness scores, and an eight-year-old has a missing fitness score. If we average the fitness scores of people between the ages of 15 and 80, the eighty-year-old will appear to have a significantly greater fitness level than he actually does. Second, mean imputation decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate and the confidence interval is narrower

14. What is linear regression in statistics?

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable. This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. There are simple linear regression calculators that use a “least squares” method to discover the best-fit line for a set of paired data. You then estimate the value of X (dependent variable) from Y (independent variable).

Linear-regression models are relatively simple and provide an easy-to-interpret mathematical formula that can generate predictions. Linear regression can be applied to various areas in business and academic study. You'll find that linear regression is used in everything from biological, behavioral, environmental and social sciences to business. Linear-regression models have become a proven way to scientifically and reliably predict the future. Because linear regression is a long-established statistical procedure, the properties of linear-regression models are well understood and can be trained very quickly.

15. What are the various branches of statistics?

The two main branches of statistics are descriptive statistics and inferential statistics Both of these are employed in scientific analysis of data and both are equally important

Descriptive Statistics

Descriptive statistics 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. Different areas of study require different kinds of analysis using descriptive statistics. For example, a physicist studying turbulence in the laboratory needs the average quantities that vary over small intervals of time. The nature of this problem requires that physical quantities be averaged from a host of data collected through the experiment**.**

Inferential Statistics

Inferential statistics, as the name suggests, involves drawing the right conclusions from the statistical analysis that has been performed using descriptive statistics. In the end, it is the inferences that make studies important and this aspect is dealt with in inferential statistics. Most predictions of the future and generalizations about a population by studying a smaller sample come under the purview of inferential statistics. Most social sciences experiments deal with studying a small sample population that helps determine how the population in general behaves. While drawing conclusions, one needs to be very careful so as not to draw the wrong or biased conclusions. Even though this appears like a science, there are ways in which one can manipulate studies and results through various means. For example, data dredging is increasingly becoming a problem as computers hold loads of information and it is easy, either intentionally or unintentionally, to use the wrong inferential methods.