1. A (True)
2. A (Central Limit Theorem)
3. B (Modeling bounded count data)
4. D (All of the mentioned)
5. C (Poisson)
6. B (False)
7. B (Hypothesis)
8. A (0)
9. C (Outliers cannot conform to the regression relationship)
10. **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 graph form, normal distribution will appear as a bell curve.

It is a bell-shaped frequency distribution curve which helps describe all the possible values a random variable can take within a given range with most of the distribution area is in the middle and few are in the tails, at the extremes. This distribution has two key parameters: the mean (µ) and the standard deviation (σ) which plays key role in assets return calculation and in risk management strategy.

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

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

The imputation method develops reasonable guesses for missing data. It’s most useful when the percentage of missing data is low. If the portion of missing data is too high, the results lack natural variation that could result in an effective model.

The other option is to remove data. When dealing with data that is missing at random, related data can be deleted to reduce bias. Removing data may not be the best option if there are not enough observations to result in a reliable analysis. In some situations, observation of specific events or factors may be required.

Three ways:

**Missing Completely At Random (MCAR):** When missing values are randomly distributed across all observations, then we consider the data to be missing completely at random. A quick check for this is to compare two parts of data – one with missing observations and the other without missing observations. On a t-test, if we do not find any difference in means between the two samples of data, we can assume the data to be MCAR.

**Missing At Random (MAR):** The key difference between MCAR and MAR is that under MAR the data is not missing randomly across all observations, but is missing randomly only within sub-samples of data. For example, if high school GPA data is missing randomly across all schools in a district, that data will be considered MCAR. However, if data is randomly missing for students in specific schools of the district, then the data is MAR.

**Not Missing At Random (NMAR):** When the missing data has a structure to it, we cannot treat it as missing at random. In the above example, if the data was missing for all students from specific schools, then the data cannot be treated as MAR

Different methods:

Do Nothing

Imputation Using (Mean/Median) Values

Imputation Using (Most Frequent) or (Zero/Constant) Values

Imputation Using k-NN

Imputation Using Multivariate Imputation by Chained Equation (MICE)

Imputation Using Deep Learning (Datawig)

1. **A/B Testing** :

A/B testing is basically statistical hypothesis testing, or, in other words, statistical inference. It is an analytical method for making decisions that estimates population parameters based on sample statistics.

The population refers to all the visitors coming to your website (or specific group of pages), while the sample refers to the number of visitors that participated in the test.

Let’s say, you make a decision to implement some change on your product pages based on A/B test results that tested a “sample” of the visitors to your website. Ultimately, only a percentage of the visitors saw the challenger, so that of course means not all the visitors. However, with A/B testing, you assume if the challenger (i.e. variation) in the test increased conversions for a group of visitors on product pages, it will thus have the same result for all the visitors of your product pages (we will delve into the accuracy of a variation’s validity later).

You start the A/B testing process by making a claim (hypothesis). You launch your test to gather statistical evidence to accept or reject a claim (hypothesis) about your website visitors. The final data shows you whether your hypothesis was correct, incorrect or inconclusive.

This is your hypothesis in “normal words.”. But how would it look like in statistics?

In statistics your hypothesis breaks down into:

Null hypothesis

Alternative hypothesis

The null hypothesis states the default position to be tested or the situation as it is (assumed to be) now, i.e. the status quo.

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

1. **Is mean imputation of missing data acceptable practice?**

Mean imputation is the practice of replacing null values in a data set with the mean of the data. Mean imputation is generally bad practice because it doesn’t consider feature correlation.

If just estimating means: mean imputation preserves the mean of the observed data

Leads to an underestimate of the standard deviation

Distorts relationships between variables by “pulling” estimates of the correlation toward zero

1. **What is linear regression in statistics?**

Simple linear regression is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables: One variable, denoted x, is regarded as the predictor, explanatory, or independent variable.

The other variable, denoted y, is regarded as the response, outcome, or dependent variable.

Simple linear regression gets its adjective "simple," because it concerns the study of only one predictor variable. One variable causes the other (for example, higher SAT scores do not cause higher college grades), but that there is some significant association between the two variables. A scatterplot can be a helpful tool in determining the strength of the relationship between two variables. If there appears to be no association between the proposed explanatory and dependent variables (i.e., the scatterplot does not indicate any increasing or decreasing trends), then fitting a linear regression model to the data probably will not provide a useful model. A valuable numerical measure of association between two variables is the correlation coefficient, which is a value between -1 and 1 indicating the strength of the association of the observed data for the two variables.

A linear regression line has an equation of the form Y = a + bX, where X is the explanatory variable and Y is the dependent variable. The slope of the line is b, and a is the intercept (the value of y when x = 0).

1. **What are the various branches of statistics?**

Statistics have majorly categorized into two types:

Descriptive statistics

Inferential statistics

**Descriptive Statistics**

In this type of statistics, the data is summarized through the given observations. The summarization is one from a sample of population using parameters such as the mean or standard deviation.

Descriptive statistics is a way to organize, represent and describe a collection of data using tables, graphs, and summary measures. For example, the collection of people in a city using the internet or using Television.

Descriptive statistics are also categorized into four different categories:

Measure of frequency

Measure of dispersion

Measure of central tendency

Measure of position

The frequency measurement displays the number of times a particular data occurs. Range, Variance, Standard Deviation are measures of dispersion. It identifies the spread of data. Central tendencies are the mean, median and mode of the data. And the measure of position describes the percentile and quartile ranks.

**Inferential Statistics**

This type of statistics is used to interpret the meaning of Descriptive statistics. That means once the data has been collected, analyzed and summarized then we use these stats to describe the meaning of the collected data. Or we can say, it is used to draw conclusions from the data that depends on random variations such as observational errors, sampling variation, etc.

Inferential Statistics is a method that allows us to use information collected from a sample to make decisions, predictions or inferences from a population. It grants us permission to give statements that goes beyond the available data or information. For example, deriving estimates from hypothetical research.