DS Bootcamp - Week 5

1. How do you assess the statistical significance of an insight?
   Statistical significance if an insight is means determining whether the insight is purely out of chance or due to some real meaningful effect. Following are the steps to access the statistical findings of an insight:
   1. Create a hypothesis: Create the null hypothesis (H0), which represents the absence of an effect or no change, and an alternative hypothesis (H1) which represents the effect we are testing.
   2. Decide on a significance level: Choose a significance level (alpha) which is typically set to 0.005. This value represents the threshold below which you will consider the results statistically significant.
   3. Select a statistically test: Select a test that is suitable for the data we are testing.
   4. Perform test: perform the statistical test on the collected data. This will generate a p-value.
   5. Compare values: compare the p-value to the alpha value. If the p-value is less than alpha, you may conclude that the results are statistically significant. In other words, you have evidence to reject the null hypothesis in favor of the alternative hypothesis.

2. What is the Central Limit Theorem? Explain it. Why is it important?
   The central limit theorem states that if you have a significantly large random sample drawn from a population with a finite mean (mu) and a finite variance (sigma square), the sampling distribution of the sample mean (x) will be approximately normally distributed, regardless of the shape of the population distribution.
   The CLT helps in hypothesis testing where it is used in underlying assumptions of hypothesis tests and helps in accessing the significance of these tests. The CLT provides a foundational concept for making statistical inferences about population parameters based on sample statistics. It enables data scientists to estimate population parameters, construct confidence intervals, and perform hypothesis tests reliably.

3. What is the statistical power?
   Statistical power, also known as ‘power’ is probability that the statistical test will correctly reject the null hypothesis when the null hypothesis is indeed false. In other words, it measures the ability of a statistical test to detect a true effect or difference when it exists.

4. How do you control for biases?
   There are several ways to control for bias, they are as follows:
   1. Data collection: Ensure that the data collected or the data being used for analysis is diverse and represents all the classes equally. One way to ensure this is to use a random sampling method to minimize possible selection bias.
   2. Bias Assessment: Do a bias assessment on the data before beginning analysis. This may include conducting demographic analysis.

3. Data Preprocessing: Use data preprocessing methods to reduce bias. This could involve handling missing values, addressing outliers, standardizing data collection methods and/or normalizing the data.

5. What are confounding variables?
   Confounding variables, also known as confounders, are variables in a research study that are not the primary independent or dependent variables of interest but can influence the outcome or the relationship between the variables of interest. In other words, confounding variables are third variables that may lead to a false association or misleading conclusions if not properly controlled for in a study. They can introduce bias and make it difficult to determine the true cause-and-effect relationship between the variables of interest.

6. What is A/B testing?
   A/B testing, also known as split testing, is a method of experimentation used in marketing, product development, and user experience (UX) design to compare two versions of a webpage, app, email, or other content to determine which one performs better. The "A" represents the control group or the current version (the "baseline"), and the "B" represents the treatment group or the new version with one or more changes (the "variant").

7. What are confidence intervals?
   A confidence interval is a statistical concept used to estimate a range of values within which a population parameter (e.g., a population mean or proportion) is likely to fall with a certain level of confidence. It provides a range of values rather than a single point estimate, allowing for uncertainty in the estimation.
   1. Point Estimate: A point estimate is the best guess of the population parameter based on a sample of data. For example, the sample mean is a point estimate of the population mean.
   2. Margin of Error: The margin of error (MOE) is a measure of the uncertainty in the estimation. It represents the range within which the true population parameter is likely to fall.
   3. Level of Confidence: The level of confidence (often denoted as $1 - \alpha$, where α is the significance level) is the probability that the confidence interval contains the true population parameter.