

## DS Homework 5

Mihir Kshirsagar  
msk9917  
N16144406

### 1. How do you assess the statistical significance of an insight?

The usual method for determining statistical significance is hypothesis testing. The following steps are involved in the process:

- Create an alternative hypothesis ( $H_1$ ) that reflects the effect you wish to investigate, and a null hypothesis ( $H_0$ ) that represents the status quo or no effect.
- Gather information and run a statistical test, such as a chi-squared or t-test, based on the ideas and the type of data you have.
- Determine a test statistic and contrast it with a p-value, which is the likelihood of witnessing the data under the null hypothesis.
- Your insight is statistically significant if the test statistic is more than the crucial value or the p-value is less than a selected significance level (e.g., 0.05). In these cases, the null hypothesis is rejected.

### 2. What is the Central Limit Theorem? Explain it. Why is it important?

An essential idea in statistics is the Central Limit Theorem (CLT). It says that regardless of the underlying distribution of the population, the distribution of sample means, when you take a sizable random sample from it and compute the sample means (or other sample statistics), will resemble a normal distribution. As long as the sample size is big enough, this is valid.

The CLT's application to inferential statistics is what gives it its significance. Based on the characteristics of the sample means, it enables data scientists to draw conclusions about the population. It is an essential tool for data analysis because it serves as the foundation for many hypothesis testing and confidence interval estimates.

### 3. What is the statistical power?

Statistical power is the likelihood of correctly rejecting the null hypothesis when it is in fact wrong. It is commonly represented as  $1 - \beta$  (beta). Stated otherwise, it assesses the capacity of a statistical test to identify a true impact or difference, should one exist. Because it lowers the possibility of making a Type II error—missing a true effect—high statistical power is preferred.

### 4. How do you control for biases?

The following are some ways to control for biases in data analysis: Use random sampling to lessen selection bias.

- To lessen observer bias, use double-blind and blinding protocols in your experiments.
- Transparently gather and evaluate data, clearly outlining your technique and underlying presumptions.
- Propensity score matching is one statistical technique that can be used to control for confounding variables.

- To prevent confirmation bias, use acceptable statistical techniques and refrain from p-hacking or data snooping.

5. What are confounding variables?

In a statistical analysis, confounding factors are third variables that have the potential to influence both the independent and dependent variables, creating an erroneous correlation or false link between them. Research must account for confounding variables in order to ascertain the true link between the variables.

6. What is A/B testing?

A controlled and randomized experiment can be used to evaluate the efficacy of changes or interventions using A/B testing, also referred to as split testing. It entails splitting a sample into two or more groups, with the current version (control) given to group A and a modified version (treatment) given to group B. You can ascertain whether the changes had a statistically significant effect by comparing the group results.

7. What are confidence intervals?

A range of values known as a confidence interval is used to calculate, with a given degree of confidence, the likely range of a population parameter (such as a mean or a proportion). Compared to a single point estimate, they offer a measure that is more informative. A confidence interval is commonly computed at a designated confidence level (e.g., 95%) and denotes the range that the true parameter is most likely to fall within. A more accurate estimate is indicated by a smaller confidence interval.