**Course Four**

# From Data to Insight: The Power of Statistics



# Instructions

Use this PACE strategy document to record decisions and reflections as you work through this end-of-course project. As a reminder, this document is a resource that you can reference in the future, and a guide to help you consider responses and reflections posed at various points throughout projects.

# Course Project Recap

Regardless of which track you have chosen to complete, your goals for this project are:

* Complete the questions in the Course 4 PACE strategy document
* Answer the questions in the Jupyter notebook project file
* Compute descriptive statistics
* Conduct a hypothesis test
* Create an executive summary for external stakeholders

# Relevant Interview Questions

Completing this end-of-course project will empower you to respond to the following interview topics:

* How would you explain an A/B test to stakeholders who may not be familiar with analytics?
* If you had access to company performance data, what statistical tests might be useful to help understand performance?
* What considerations would you think about when presenting results to make sure they have an impact or have achieved the desired results?
* What are some effective ways to communicate statistical concepts/methods to a non-technical audience?
* In your own words, explain the factors that go into an experimental design for designs such as A/B tests.

**Reference Guide**

This project has four tasks; the visual below identifies how the stages of PACE are incorporated across those tasks.



**Data Project Questions & Considerations**

**PACE: Plan Stage**

* What is the main purpose of this project?

To understand whether there is a statistically significant difference in the means of the verified and unverified accounts in terms of viewership count.

This will provide insight into their differences and how to treat these two groups differently.

* What is your research question for this project?

In the long term, to understand the differences between reports that are claims and opinions

* What is the importance of random sampling?

Often, we are not able to take the entire population when trying to calculate its parameter. So, instead what we do is take a small subset of the population, called a sample, calculate the corresponding sample statistic and then infer the population mean using the sample statistic. The process used to find the subset is called subset and random sampling is when we randomly select people from the population, thereby, giving everyone an equal chance. Otherwise, we might encounter bias in our sample which would then make the findings innacurate.

* Give an example of sampling bias that might occur if you didn’t use random sampling.

Under coverage bias – is the bias that comes when a certain group in the population is not represented enough in the dataset



 **PACE: Analyze & Construct Stages**

* In general, why are descriptive statistics useful?

They help in getting a quick overview of a large dataset. This can then be used to drill down to better understand the data.

* How did computing descriptive statistics help you analyze your data?

It helped to evaluate my dataset in terms of many field.

It contained the datasets central tendancies like mean, median and mode

Then it allowed me to understand the skewness by providing the measures of dispersion like range and standard deviation.

* In hypothesis testing, what is the difference between the null hypothesis and the alternative hypothesis?

The null hypothesis is considered to be true unless there is strong proof against the statement.

This usually represents the typical system of things. It contains the equality sign.

The alternate hypothesis contradicts the null hypothesis and is taken to be true only when we have a lot of proof to support that claim

* How did you formulate your null hypothesis and alternative hypothesis?

Since we have two samples and we are trying to find whether the mean values are same or not, we can just use a two sample t test.

The null hypothesis will be that the means are same while the alt hypothesis will be that the means are not the same, this is because we are not interested in knowing whether the mean is less or greater, we just want to know if the two means are equal or not.

* What conclusion can be drawn from the hypothesis test?

We can see that there may be a relationship between the two variables view count and verified status, with the unverified authors getting significantly more views. Thus, the difference in the means between the verified and unverified groups is statistically significant and cannot be left upto chance.

**PACE: Execute Stage**

* What key business or organizational insight(s) emerged from your A/B test?

Found out that there may be a relationship between the verified status of authors and their video view counts.

* What recommendations do you propose based on your results?

Try to create a regression model so that we can reliably predict whether a report is a claim or opinion as this is the end goal of the project.