Statistics for the Apps Project

Overview

To complete the Apps project (and succeed as a data scientist,) there are a couple of important statistical concepts you'll need to become familiar with. You can give this article a read now, or read through it when you begin working on the statistics portion of this project (which you'll be starting in on in just a moment.)

Below, we've highlighted the integral statistics concepts you'll use in this project (don't worry — they're fairly lightweight!)

Statistics and Data Science

Most models in data science involve making a statistical hypothesis and then using statistical methods to test that hypothesis. Some people even define the term 'data science' as just the combination of applied ***descriptive statistics*** (saying how things are on the basis of statistical data) and applied ***predictive statistics*** (saying how things will be on the basis of that data).

While statistics aren't something you can learn overnight, everything you need to know about statistics for the App project is contained in this document.

The Use of Statistics in the Apps Project

As you know, for this final course project, you've assumed the role of a data scientist at a consulting a firm. Your client wants to know if they should integrate Goolge Play or Apple Stores apps into their operating system. For the stats portion of the project, you'll use statistical methodology to check whether the difference between the mean rating of

Google Play apps and the mean rating of Apple Store apps is statistically significant. If it is, you'll be able to conclude that the platform does affect ratings, which means you'll be able to advise your client to integrate the platform that maximizes ratings into their operating system.

In this project, you'll calculate the mean rating for Google Play apps and the mean rating for Apple Store apps, and the difference between these means. This difference (also called our ***'observed difference***') will be this number: 0.14206. It's the task of statistical testing to check whether this difference is ***significant***, and hence whether it justifies your conclusion that platform does impact ratings.

Details

We already saw more or less how our statistical test is going to go (refer back to the Jupyter Notebook for the specific steps.) We're going to assume, at the start, a hypothesis called the ***Null hypothesis*** (also called 'H0'):

> ***H0***: the observed difference in the mean rating of Apple Store and Google Play apps is due to chance (and thus not due to the platform).

The more interesting hypothesis is called the ***Alternate hypothesis*** (also called 'HA'):

> ***HA***: the observed difference in the average ratings of Apple and Google users is not due to chance (and is actually due to platform)

We'll then work out how things would look were the Null true by shuffling (or *permuting)* the ratings column many times but keeping the platform constant. The motivation for this is as follows: if the Null were true, platform would have no bearing on ratings, so shuffling ratings but keeping the platforms constant shouldn't impact the difference in the mean rating of Google apps and the mean rating of Apple apps. We can then compare the observed difference with the average difference of these shuffles, and if they're significantly different, we reject the Null.

Permutation Difference and Observed Difference

For each shuffle, we'll calculate the mean rating for Apple apps, the mean rating for Google apps, and the difference between these means. Then we'll get the mean of all these differences. This mean will be our ***permutation difference*** and will be compared to our actual ***observed difference***. If they are sufficiently close, we accept our initial assumption of the Null; but if they are sufficiently different, we will reject the Null and accept the Alternate hypothesis.

What counts as sufficiently close? Who decides? Fascinatingly, as statisticians, we do. But this doesn't mean that no choice of this value is better than any other.

Normally, if the probability of seeing our observed difference on the assumption of the Null hypothesis is less than 0.05, and yet we still saw that observed difference, we conclude that the Null is false. This makes sense, right? In general, if the probability of seeing a given thing on the assumption of some hypothesis is super low, and yet we still definitely saw that thing, we can actually conclude that that hypothesis is likely to be wrong.

P-Values

The ***p-value*** of our observed data is just the probability of seeing it on the assumption of the Null. Since we're assuming probabilities are just proportions, here's an equivalent definition: the p-value of the observed data is the proportion of the data given the Null that's at least as extreme as that observed data.

Significance Level

The ***significance level*** is just how improbable our observed data has to be for it to be deemed *significant* and thus as sufficient ground for rejecting the Null. For example, choosing a significance level of 0.05 is just saying that any observed data whose probability given the Null is less than or equal to 0.05 is significant and thus is sufficient ground on which to reject the Null. In other words, choosing a significance level of 0.05 is just saying that if less than or equal to 5% of the data given the Null are at least as extreme as the observed data, we reject the Null.

There are customary significance levels: 0.05, 0.01, 0.005 and 0.001. The lower the significance level, the more improbable our observed data has to be given the Null for it to be taken as significant and thus as sufficient ground on which to reject the Null.

Type 1 and Type 2 Errors

Fascinatingly, we might opt for different significance levels in different contexts, depending on how important it is to avoid ***Type 1 Errors*** (rejecting a true Null) or ***Type 2***

Errors (not rejecting a false Null). In actual fact, we can only reduce the chance of one type of error by increasing the chance of the other.

For example, in a medical context where the Null is just that the patient is free from a given disease, we might push the significance level up; thereby making the Null easier to reject, and thus Type 2 Errors less likely. Notice that this will be at the expense of making Type 1 Errors more likely: we'll reject more true Nulls than if the significance level were lower.

By contrast, in a judicial context where the Null is that the defendant is innocent, we might want to push the significance level down; thereby making it harder to reject the Null and the imprisonment of innocent defendants less likely. Note that this will come at the expense of a higher rate of Type 2 Error: in more cases, we won't correctly identify judge guilty defendants.

Histograms

One of the visualizations we'll look at in the App Project is the ***histogram***. A histogram of a given column simply reveals the frequency with which the various values of that column occur. Remind yourself about histograms here.

Other Statistical Tests

The test we're using here is called the Permutation test. But there are actually many different statistical tests, all with different assumptions. You'll generate an excellent judgment about when to use which statistical tests over the Data Science Career Track course. A great resource to read in regard to the conditions under which different statistical tests should be used is David Spiegelhalter's *The Art of Statistics.* It will be the perfect accompaniment to the DSC and can be read slowly as you become increasingly comfortable in Python.