AB Testing - Introduction

Introduction

In this section, you'll get to develop your skills regarding AB testing. Before diving in, take a minute to note some key points which you should keep in mind when completing the various labs and conducting your own hypothesis tests in practice.

Experimental Design

You've seen that a lot goes into the proper design of statistical tests. You've learned about Goodheart's law as well as the multiple comparisons problem. Additionally, you've also seen that a p-value by itself is prone to misinterpretation if not presented with other relevant design parameters such as effect size, sample size, and α. With that, here are three overarching considerations to keep in mind.

Well Formulated Questions

A well-formulated question is essential to a good statistical experiment. This includes careful thought of unintended consequences, as you saw in the discussion of Goodheart's law. Additionally, the question should also be specific and measurable.

Choosing Appropriate Parameters

It cannot be stressed enough how important the relationship between α, power, sample size and effect size is. While larger sample sizes provide more powerful tests, one should also realize that tiny effects can produce significant p-values with large samples. While this may be interesting, such small practical differences might have little to no applicable value. Furthermore, avoiding pitfalls such as the multiple comparisons problem is also important. Recall that if you perform multiple t-tests, The probability of encountering a type I error will continue to increase with additional tests. (Each test will still have the corresponding alpha value set, but collectively, the chance that a false positive type I error exists in your conclusions increases.)

Preprocessing, Data Anomalies and Outliers

On the other end of problem formulation is formatting the data to actually answer said question. You'll encounter this most explicitly in the final lab of this section. There, you'll have to transform your data into an appropriate format before conducting the statistical test. Furthermore, it's important to note how idiosyncrasies in your data can impact results. For example, monumental outliers can drastically impact the outcome of statistical tests. Whether or not to remove these data points can be a source of contention and will vary upon the circumstance. Similarly, it should go without saying that erroneous data or faulty data will clearly degrade statistical tests. All in all, it's always important to get familiar with the structure of the data and the context of the question being asked before diving into the statistics themselves.

Summary

Time to have at it! Dive in and start practicing some hypothesis testing!

A/B Testing

Introduction

You've now seen all of the statistical techniques and background to design and conduct your own A/B tests in practice! To do this, you'll go through the process of stating the null hypothesis and the alternative hypothesis which will include some test statistic for comparison. For example, you might compare the average purchase price between customers on two different versions of your online store, or a pharmaceutical researcher might compare the blood pressure of patients before and after taking a prescription. You've also seen that good test design requires multiple decisions. Recall both the multiple comparisons problem and Goodheart's law: you can't effectively measure everything (without increasing the risk of mistakes), and incentivizing measurements can lead to unforeseen consequences. With that, this section will give you a chance to put your new statistical techniques into practical applications.

Objectives

You will be able to:

- List the steps required to design, structure, and run an A/B test

Choosing a Metric

Any hypothesis testing will start with a given scenario. This will determine everything from what metrics are deemed important to realistically obtainable sample sizes. Goodheart's law can also be an important consideration when performing ongoing tests in attempts to optimize performance metrics.

Defining the Null Hypothesis: H_0

Once an appropriate metric has been selected, it's time to formally define the experiment with a null hypothesis. Typically, the null hypothesis is the claim that a researcher is hoping to refute. For example, a medical researcher might hope to show that a new drug is more effective than a previous treatment option. A common practice is then to define the null hypothesis as the contrary: there is no difference between the two drugs. The researcher hopes to refute the null hypothesis thereby proving their claim by contradiction.

You might start with something like "drug_a is more effective than drug_b".

While this is a good start, proper formulation of the null hypothesis should ensure that it is written with a quantitative measurement. Perhaps the drugs are for high blood pressure and so the statement becomes, "drug_a at dose_a lowers blood pressure more than drug_b at dose_b".

Formulating the null-hypothesis like this is apt to lead you into conducting a paired t-test for the mean blood pressure of two groups: one representing drug_a, and another representing drug_b.

Alternatively, if one of these two medications were more heavily researched, one might wish to compare the effectiveness of the new medication with a predetermined metric such as the average drop in blood pressure from medication. This would lead to a 1-sample t-test, as opposed to a two-sample t-test.

Investigating α, power, effect size, and sample size

Finally, one must formulate the various surrounding parameters required to conduct the test. You've seen that there is an intimate relationship between α, power, sample size, and effect size. With that, questions such as "How costly is sample size?" are instrumental in experiment design. For example, in an online scenario, it might be quite easy to conduct experiments at scale. On the other hand, in medical research, larger sample sizes are apt to be extremely costly. Investigating the relationships between α, power, effect size, and sample size is important for all experiments, and a suitable combination will depend on these contextual factors regarding implementation.

Summary

When researching, you are often presented with two choices for stating a question. One is to estimate a parameter in question, such as the procedures previously examined for estimating the mean of a population. Alternatively, you may wish to test the validity of a claim—whether you can refute that claim, or whether you should withhold judgment. In practice, it is up to the practitioner to determine the appropriate alpha, beta, and sample size that is determined to be both satisfactory confidence and a viable sample

size to attain. In the upcoming labs, you'll get to practice this setup and design work-flow for a few scenarios.

AB Testing - Recap

Introduction

In this section, you got a chance to further practice hypothesis testing with some real-world scenarios. Here's a brief recap of some of the key takeaways from the section.

A Brief Review

When conducting statistical tests, there's simply no substitute for critical thinking. As you've seen, there are numerous considerations from formulating appropriate hypotheses to thinking about the context of the problem itself. While the techniques presented are extremely powerful if properly employed, you've also seen how each technique comes with its own assumptions, and ignoring these can invalidate results. Similarly, how the test is structured can lead to widely different results.

There are also additional statistical tests not touched upon here. To date, your primary method has been conducting t-tests. This the most common statistical test used in practice and is very effective at comparing averages (of any metric) between groups. You've also seen how ANOVA can generalize this process to multiple groups. If you wish to continue to strengthen your knowledge of other statistical tests, take a look at some of the resources below.

Additional Resources

- Choosing Statistical Tests
- How to choose the right statistical test?
- Hypothesis Testing

Summary

Well done! You've not only learned about a variety of statistical tests and the theory behind them, but you've now also put these techniques into practice in order to carry out hypothesis testing!