INTRODUCTION TO A/B TESTING

Joseph Nelson, Data Science Immersive

AGENDA

‣ What is A/B Testing?

‣ A/B Testing Design

‣ T-Tests and Z-Tests

‣ Case Studies

WHAT IS A/B TESTING?

‣ Some of you have likely seen or completed A/B tests before. What are they?

WHAT IS A/B TESTING?

‣ Some of you have likely seen or completed A/B tests before. What are they?

‣ A/B Testing is a term for a randomized experiment with two variants, A and B. These tests consist of test design, data collection, and data analysis stages.

WHAT IS A/B TESTING?

‣ The most common use of A/B testing is to audition proposed changes to a website. Once the variants are designed, data is collected by assigning users to 'test' and 'control' groups, which will dictate the version of the site they will be served.

50 % visitors see variation A → Variation A → 23% conversion

50 % visitors see variation B → Variation B → 11% conversion

WHAT IS A/B TESTING?

‣ It's very important when designing an A/B test to make the smallest change possible before testing the variant. Widespread changes introduce a slew of variables that will be impossible to track in most cases. Some examples of A/B tests that one might conduct are:

‣ Changing the number of images on a page

‣ Changing the font on a page

‣ Adding or removing single elements from a page

‣ Altering the text on a button

‣ Re-organizing a pages content

WHAT IS A/B TESTING?

‣ Consider an e-commerce site. What must be taken into account when designing, conducting, and analyzing an A/B test?

WHAT IS A/B TESTING?

‣ Consider an e-commerce site. What must be taken into account when designing, conducting, and analyzing an A/B test?

‣ The main effect in e-commerce is the flow of the user through the conversion funnel! Once users land on the site, test to see if the variant has any effect on how many products they **view**, how many products are **added to cart**, **changes in cart abandonment rates**, changes in **conversion rates**, **order volume**, **average order value**, etc.

‣ 1. What element(s) will be changed?

‣ 2. Who will be a part of the test group?

‣ 3. How long will the test run?

‣ 4. Why is this test truly necessary?

‣ 1. What element(s) will be changed?

‣ While working with a (or as a) PM, you will likely have little say in what elements are changed for a test. Keep in mind that to prevent false correlations in the data, the **smallest changes possible** will likely have the most meaningful results.

‣ 2. Who will be a part of the test group?

‣ 3. How long will the test run?

‣ 4. Why is this test truly necessary?

‣ 1. What element(s) will be changed?

‣ 2. Who will be a part of the test group?

‣ Will you be splitting the incoming traffic 50/50 between variants, or can you get away with serving the variant under test to a smaller group? Also, will the test split change? (We'll discuss one strategy for assigning test groups in a minute)

‣ 3. How long will the test run?

‣ 4. Why is this test truly necessary?

‣ 1. What element(s) will be changed?

‣ 2. Who will be a part of the test group?

‣ 3. How long will the test run?

‣ This is a very important question to ask. If the test doesn't run **long enough**, your data won't be useful. If it runs **too long**, that can impact business needs. Remember back to Week 9's Time Series Analysis lessons- ensure that you have enough data to capture across multiple periods, or seasons, but not too much data that your result will be heavily affected by trend.

‣ 4. Why is this test truly necessary?

‣ 1. What element(s) will be changed?

‣ 2. Who will be a part of the test group?

‣ 3. How long will the test run?

‣ 4. Why is this test truly necessary?

‣ A/B testing is a gamble. If the business result of the test is less valuable than the possible negative effects on churn or conversion rate, then it might be worth re-evaluating your design.

‣ Multi-arm bandit testing is an innovative way to split traffic (rather than simply 50/50) developed by Google (more on this in Resources).

‣ There are two phases:

‣ **Exploration Phase:** During the first ~10% of the test, traffic is split 50/50. This phase picks a short-term 'winner', and a short-term 'loser'.

‣ **Exploitation Phase**: For the remainder of the test, shift the majority of traffic to the higher performing variant. Continue to adjust traffic as performance increases/decreases.
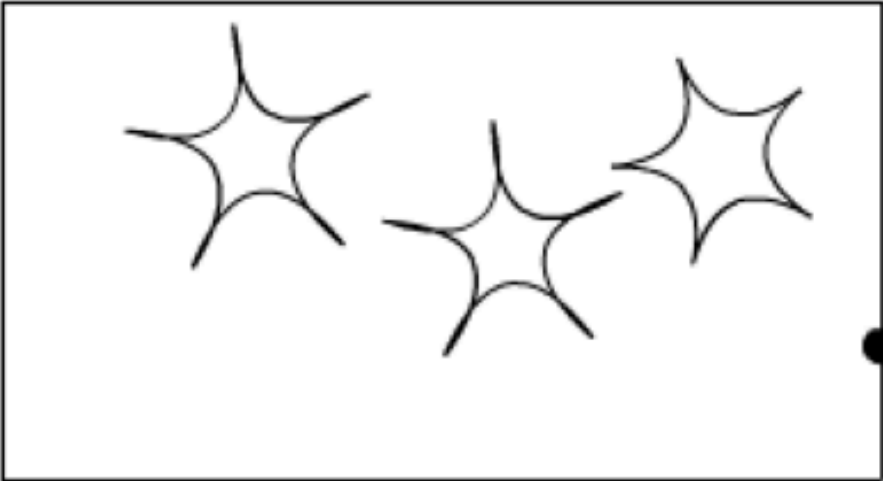
‣ Pros/Cons?

‣ Pros/Cons?

‣ In practice, Multi-Arm Bandit testing does a fairly good job of optimizing conversion rates. The downside to this method, however, is increased difficulty in evaluation of results. Simply picking a 'winner' variant is not always the best strategy, especially since the 'loser' variant often gets so little traffic that it can be hard to validate the statistical significance of the lift.

Variant	Image	Title
Control	Stars	General Assembly DSI
Test 1	Flowers	General Assembly DSI
Test 2	Stars	Data Science Immersive
Test 3	Flowers	Data Science Immersive

▸ REMEMBER: The t-test is one of the most commonly used techniques for testing a null hypothesis on the basis of a difference between sample means. By testing the means of two samples derived from the same source, the t-test determines a probability that two populations are the same with respect to the variable tested. In an A/B test, this will tell us if the difference in the target metric is accidental, or statistically significant.
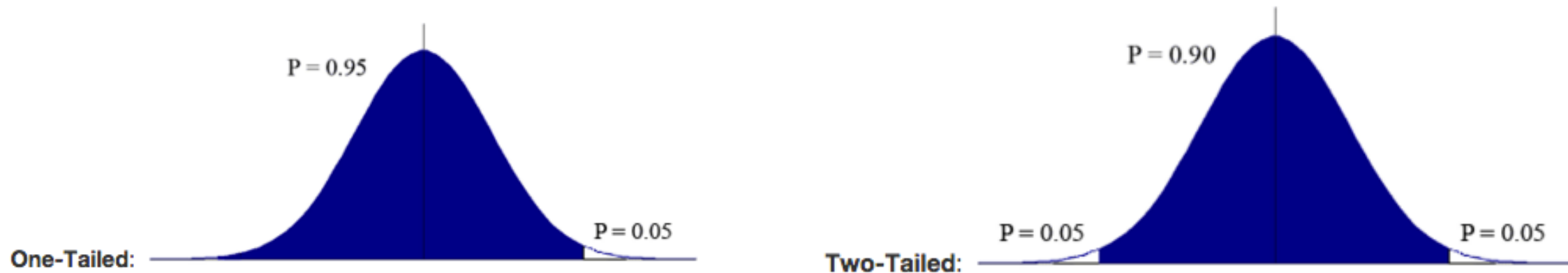
$$t = \sqrt{p}\frac{Z}{s} = \sqrt{p}\frac{(\bar{X} - \mu)/(\sigma/\sqrt{n})}{s}$$

‣ T-tests may be one-tailed or two-tailed.

‣ One-tailed t-tests examine whether a test population varies in a single direction vs the potential for two directions.

‣ If you're checking a new drug vs an old drug, which would you choose? Why?

- Z-tests are another method used to analyze test results. Use of a z-test is possible when the observed data can be decided to follow a Normal distribution with unknown mean and known variance. The output of a z-test is the z-statistic, which represents the number of standard deviations and its corresponding p-value. It is defined as such:

$$z \text{ test} = \frac{P_h - P_m}{\sqrt{P(1-P)(\frac{1}{N_h} + \frac{1}{N_m})}}$$

CASE STUDY ONE: OBAMA FUNDRAISING

‣ https://blog.optimizely.com/2010/11/29/how-obama-raised-60-million-by-running-a-simple-experiment/

CASE STUDY TWO: SOCIAL SHARING

‣ https://vwo.com/blog/amd-3600-social-sharing-increase/

‣ To the repo…