In this case study, you'll apply what you've learned on confidence intervals and hypothesis testing to help a company decide whether to launch two new features on their website. To do this, you'll analyze results from A/B testing, a valuable and widely practiced method in industry.

A/B tests are used to test changes on a web page by running an experiment where a **control group**sees the old version, while the **experiment group** sees the new version. A **metric** is then chosen to measure the level of engagement from users in each group. These results are then used to judge whether one version is more effective than the other. A/B testing is very much like hypothesis testing with the following hypotheses:

* **Null Hypothesis:** The new version is no better, or even worse, than the old version
* **Alternative Hypothesis:** The new version is better than the old version

If we fail to reject the null hypothesis, the results would suggest keeping the old version. If we reject the null hypothesis, the results would suggest launching the change. These tests can be used for a wide variety of changes, from large feature additions to small adjustments in color, to see what change maximizes your metric the most.

A/B testing also has its drawbacks. It can help you compare two options, but it can't tell you about an option you haven’t considered. It can also produce bias results when tested on existing users, due to factors like change aversion and novelty effect.

* **Change Aversion:** Existing users may give an unfair advantage to the old version, simply because they are unhappy with change, even if it’s ultimately for the better.
* **Novelty Effect:** Existing users may give an unfair advantage to the new version, because they’re excited or drawn to the change, even if it isn’t any better in the long run.

You'll learn more about factors like these later.

In this case study, you’ll analyze A/B test results for Audacity. Here's the customer funnel for typical new users on their site:

**View home page > Explore courses > View course overview page > Enroll in course > Complete course**

Audacity loses users as they go down the stages of this funnel, with only a few making it to the end. To increase student engagement, Audacity is performing A/B tests to try out changes that will hopefully increase conversion rates from one stage to the next.

We’ll analyze test results for two changes they have in mind, and then make a recommendation on whether they should launch each change.

The first change Audacity wants to try is on their homepage. They hope that this new, more engaging design will increase the number of users that explore their courses, that is, move on to the second stage of the funnel.

The metric we will use is the click through rate for the Explore Courses button on the home page. **Click through rate (CTR)** is often defined as the the number of clicks divided by the number of views. Since Audacity uses cookies, we can identify unique users and make sure we don't count the same one multiple times. For this experiment, we'll define our click through rate as:

**CTR: # clicks by unique users / # views by unique users**

Now that we have our metric, let's set up our null and alternative hypotheses:

H\_0: CTR\_{new} \leq CTR \_{old}*H*0​:*CTRnew*​≤*CTRold*​

H\_1: CTR\_{new} > CTR \_{old}*H*1​:*CTRnew*​>*CTRold*​

Our alternative hypothesis is what we want to prove to be true, in this case, that the new homepage design has a higher click through rate than the old homepage design. And the null hypothesis is what we assume to be true before analyzing data, which is that the new homepage design has a click through rate that is less than or equal to that of the old homepage design. As you’ve seen before, we can rearrange our hypotheses to look like this:

H\_0: CTR\_{new} - CTR\_{old} \leq 0*H*0​:*CTRnew*​−*CTRold*​≤0  
H\_1: CTR\_{new} - CTR\_{old} > 0*H*1​:*CTRnew*​−*CTRold*​>0

Let's recap the steps we took to analyze the results of this A/B test.

1. We computed the **observed difference** between the metric, click through rate, for the control and experiment group.
2. We simulated the **sampling distribution** for the difference in proportions (or difference in click through rates).
3. We used this sampling distribution to simulate the **distribution under the null** hypothesis, by creating a random normal distribution centered at 0 with the same spread and size.
4. We computed the **p-value** by finding the proportion of values in the null distribution that were greater than our observed difference.
5. We used this p-value to determine the **statistical significance** of our observed difference.

Experiment 2

The second change Audacity is A/B testing is a more career focused description on a course overview page. They hope that this change may encourage more users to enroll and complete this course. In this experiment, we’re going to analyze the following metrics:

1. **Enrollment Rate:** Click through rate for the *Enroll* button the course overview page
2. **Average Reading Duration:** Average number of seconds spent on the course overview page
3. **Average Classroom Time:** Average number of days spent in the classroom for students enrolled in the course
4. **Completion Rate:** Course completion rate for students enrolled in the course

First, let's determine if the difference observed for each metric is statistically significant individually.

Again, let's recap the steps we took to analyze the results of this A/B test.

1. We computed the **observed difference** between the metric, average reading duration, for the control and experiment group.
2. We simulated the **sampling distribution** for the difference in means (or average reading durations).
3. We used this sampling distribution to simulate the **distribution under the null** hypothesis, by creating a random normal distribution centered at 0 with the same spread and size.
4. We computed the **p-value** by finding the proportion of values in the null distribution that were greater than our observed difference.
5. We used this p-value to determine the **statistical significance** of our observed difference.

The more metrics you evaluate, the more likely you are to observe significant differences just by chance - similar to what you saw in previous lessons with multiple tests. Luckily, this [**multiple comparisons problem**](https://en.wikipedia.org/wiki/Multiple_comparisons_problem) can be handled in several ways.

**Bonferroni Correction**

If you remember from the previous lesson, the [**Bonferroni Correction**](http://en.wikipedia.org/wiki/Bonferroni_correction) is one way we could handle experiments with multiple tests, or metrics in this case. To compute the new bonferroni correct alpha value, we need to divide the original alpha value by the number of tests.

Since the Bonferroni method is too conservative when we expect correlation among metrics, we can better approach this problem with more sophisticated methods, such as the [**closed testing procedure**](http://en.wikipedia.org/wiki/Closed_testing_procedure), [**Boole-Bonferroni bound**](http://en.wikipedia.org/wiki/Bonferroni_bound), and the [**Holm-Bonferroni method**](http://en.wikipedia.org/wiki/Holm%E2%80%93Bonferroni_method). These are less conservative and take this correlation into account.

If you do choose to use a less conservative method, just make sure the assumptions of that method are truly met in your situation, and that you're not just trying to [**cheat on a p-value**](http://freakonometrics.hypotheses.org/19817). Choosing a poorly suited test just to get significant results will only lead to misguided decisions that harm your company's performance in the long run.

# Difficulties in A/B Testing

Being able to determine the statistical significance of performance differences in A/B test results is valuable. However, there are many other factors to consider to ensure your A/B tests are successful. In the real world, designing, running, and drawing conclusions from an A/B test to lead you to the right decisions can be tricky.

In the following quizzes, you'll find three scenarios where Audacity conducted an A/B test that led to a poor decision. Think about what went wrong and what could've been done to avoid these outcomes.

## Scenario #1

* **EXPERIMENT:** Audacity tests a new layout in the classroom to see if it would help engage students. After running an A/B test for two weeks, they find that average classroom times and completion rates decrease with the new layout, and decide against launching the change.
* **REALITY:** What they don't know, is that the classroom times and completion rates actually increase significantly for new students using the new layout. In the long run, the layout would help existing students too, but they are currently experiencing change aversion.

### QUESTION 1 OF 3

Based on the information above, which of the following contributed to this misguided decision? Select all that apply.

* 

A change in layout is not an appropriate change to A/B test.

* 

The sample of students tested in this experiment was too small.

* The experiment included existing users who would bias results in a short time frame.
* The experiment wasn't run long enough to allow existing users to adjust to the change.

SUBMIT

## Scenario #2

* **EXPERIMENT:** Audacity tests a new feature on their landing page that guides users through their course selection process and makes recommendations. After running an A/B test for a month, they find that the click through rates on course pages increase (enrollment rates increase) with the new feature, and decide to launch the change.
* **REALITY:** What they don't know, is that although the number of total enrollments increase with the new feature, the courses they are purchasing are almost exclusively shorter and cheaper courses, which brings down the revenue for Audacity. It turns out the feature is leading more students to choose more courses with smaller commitments.

### QUESTION 2 OF 3

Based on the information above, which of the following contributed to this misguided decision? Select all that apply.

* 

The addition of entire new features isn't something that can be A/B tested.

* Enrollment rates alone was not the best metric for this experiment.
* 

The experiment was not run long enough to have significant results.

* Their metric(s) didn't account for revenue, which is ultimately what they want to increase with their decision.

SUBMIT

## Scenario #3

* **EXPERIMENT:** Audacity tests a new description for a difficult course that gets very few enrollments. They hope this description is more exciting and motivates students to take it. After running an A/B test for five weeks, they find that the enrollment rate increases with the new description, and decide to launch the change.
* **REALITY:** What they don't know, is that although the enrollment rate appears to increase with the new description, the results from this A/B test are unreliable and largely due to chance, because fewer than 40 out of thousands of visitors enrolled during this experiment. This makes even one new student for the course substantially impact the results and potentially even the conclusion.

### QUESTION 3 OF 3

Based on the information above, which of the following contributed to this misguided decision? Select all that apply.

* This course page had too little traffic and conversions to produce significant and repeatable results in this time frame.
* 

A significant increase in enrollment rate is not a good enough reason to launch this new description for this course.

* 

The experiment did not target a group of similar enough users.

* 

The experiment was run for too long.

SUBMIT

# Difficulties in A/B Testing

As you saw in the scenarios above, there are many factors to consider when designing an A/B test and drawing conclusions based on its results. To conclude, here are some common ones to consider.

* Novelty effect and change aversion when existing users first experience a change
* Sufficient traffic and conversions to have significant and repeatable results
* Best metric choice for making the ultimate decision (eg. measuring revenue vs. clicks)
* Long enough run time for the experiment to account for changes in behavior based on time of day/week or seasonal events.
* Practical significance of a conversion rate (the cost of launching a new feature vs. the gain from the increase in conversion)
* Consistency among test subjects in the control and experiment group (imbalance in the population represented in each group can lead to situations like [**Simpson's Paradox**](https://en.wikipedia.org/wiki/Simpson%27s_paradox))