A/B Testing Notes

Overview of A/B Testing

Overview of Business Example

Audacity is a company that creates online finance courses. The customer funnel of their website is the following one:

However, this is a simplistic idea. Normally, the users don't usually enter, create an account and then complete a class in a consistent manner. There is a lot of back and forth between the different stages. Other users will skip some of the stages from the customer funnel.

We will run a simple experiment from start to finish. Our first hypothesis is that changing the "Start Now" button from orange to pink will increase how many students explore Audacity's courses.

Metric Choice

Based on the hypothesis, we need to select a metric for our experiment. Which metric could we use?

- **Total number of courses completed**. This is what Audacity ultimately cares about, so this could be a possible metric. However, given that it can take students weeks or months to complete a course, using this metric would simply take too much time to be practical.
- **Number of clicks** on the 'Start now' button. The assumption is that if more people click the button and thus move on exploring the site, then eventually some of them will create an account and go on to complete a course. In other words, increasing the rate at which users progress down the funnel at one level, will have an impact on the end of the funnel as well. However, what happens if more total users view the page in one version of the experiment? Suppose these dots are the total number of visitors of each group of the experiment, and the yellow dots are the one who clicked. More clicks occurred in group 1, but there was a greater percentage of clicks in group 2. So, instead, we can use the fraction of visitors who clicked.
- **Click-Through-Rate of clicks/number of pageviews).** This is the number of clicks on the view course's button divided by the number of pageviews to the homepage.
- **Click Through-Probability (unique users who clicked/unique visitors to page).** How is this metric different from the previous one? Imagine that two users view the homepage. The first leaves without clicking on the 'Start now' button, which means they clicked 0 times. The second person clicks 5 times. The next page loaded slowly, so the user impatiently clicked 5 times. In this case, the CTR is 2.5%, since there were 5 total clicks and 2 pageviews. But the CTP equals 0.5%, since half of the users who visited the page clicked. Then, we will choose this metric for our study case.

Now, the updated hypothesis is that changing the 'Start now' button from orange to pink will increase the CTP of the button. We assume that will increase the final business metric, which is total courses completed.

When to use CTR and CTP

Generally speaking, you use a rate when you want to measure the usability of the site and a probability when you want to measure the total impact. So, for example, if you want to measure the usability of a particular button, you use a rate because the users have a variety of different places on the page that they can actually choose to click on. The rate will say how often they actually click that button.

Now, if you just want to know how often users went to the second level page on your site, you use a probability because you don't want to count if users double-clicked or did, they reload. To compute the probability, you are going to have to match each pageview with all of the child clicks, so that you count, at least, one child clicks per page.

Which Distribution?

How do you know how variable your estimate is likely to be? For this data specifically, we are going to work with a binomial distribution instead of normal distribution because instead of having continuous data, we have successes and failures. Now, for the binomial, we are in good shape for this example because we can call a click a success and no-click visit to the page a failure. You could also use this distribution if your data is red and blue, any two exclusive outcomes.

Binomial Distribution

A binomial distribution can be thought of as simply the probability of a SUCCESS or FAILURE outcome in an experiment or survey that is repeated multiple times. The binomial is a type of distribution that has two possible outcomes (the prefix "bi" means two, or twice). For example, a coin toss has only two possible outcomes: heads or tails and taking a test could have two possible outcomes: pass or fail.

Binomial distributions must also meet the following three criteria:

1. The number of observations or trials is fixed. In other words, you can only figure out the probability of something happening if you do it a certain number of times. This is common sense—if you toss a coin once, your probability of getting a tail is 50%. If you toss a coin 20 times, your probability of getting a tail is very, very close to 100%.
2. Each observation or trial is independent. In other words, none of your trials have an effect on the probability of the next trial.
3. The probability of success (tails, heads, fail or pass) is exactly the same from one trial to another.

Confidence Interval

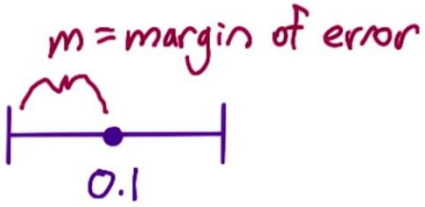
We expect CTP to follow a binomial distribution, but how do we usually use this to decide which values should surprise us? The benefit of knowing it should follow a binomial distribution is that we can use the formula we have for sample standard error for the binomial to estimate how variable we expect our overall probability of a click to be. What that means is that for a 95% confidence interval, if we theoretically repeated the experiment over and over again, we would expect the interval we construct around our sample mean to cover the true value of the population 95% of the time.

Now, let's take a look at how to compute our confidence interval for our sample. We've already found the center of the interval, that is the probability of a click. The equation was the following one:

$$p' = x \cdot N$$

where x was the number of users who clicked and N was the total number of users who visited the page. In this example, p' is 100 over 1000, which is 0.1. So, the center of the confidence interval is 0.1.

Next, I want to calculate the width of the confidence interval, which is also called the margin of error (m). To do this, I'll need to use the standard error of the binomial distribution. Recall that if the sample is large enough, instead of using the binomial distribution, I can assume the distribution is normal.



A good rule of thumb is that if N times p' is greater than 5, then it's safe to assume a normal distribution. You should also check that N(p'-1) is greater than 5. However, for small CTPs like we usually see, this is the strictest condition. In our case, N times p' is 100, so the assumption should be safe here.

When we use the normal approximation, then the width of the confidence interval (margin of error) will be equal to the z-score of the confidence level times the standard error:

$$m = z \cdot SE$$

For the binomial distribution, the standard error is the following one:

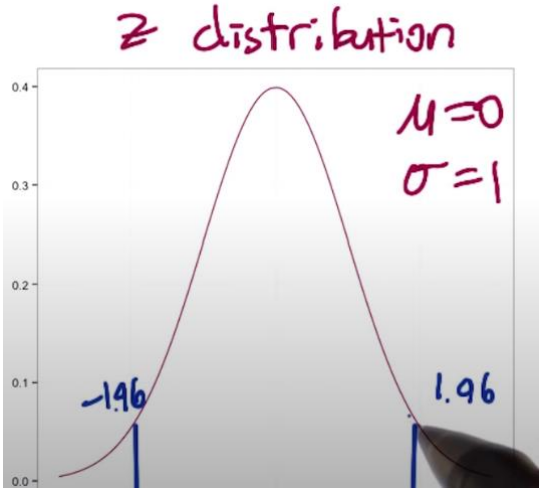
$$m = z \cdot \sqrt{\frac{p'(1-p')}{N}}$$

If you look at this formula, you will notice a few things. The amount of random variation we expect in our sample (width of the confidence interval) is a function of both the proportion of successes and the size of the sample. This means we need to consider the proportion of successes when we decide how many samples to collect. When the success probability is farther from p' = 0.5, then the standard error will be smaller, which means the distribution is tighter, which means the confidence interval will be smaller. Similarly, if the number of samples is larger, the standard error and the confidence interval will also be smaller.

Now, we need to find the z-score for the boundary of a 95% confidence interval. If we have a normal distribution with a mean of 0 and a standard deviation of 1, this is called a z-distribution. With a 95% confidence, the true value would be within 1.96 and -1.96 of the estimates we observed. So, we said the 95% confidence level corresponds to a z-score of 1.96. How can we get that? First, we need to calculate the area on the curve to the left. We will use the following formula:

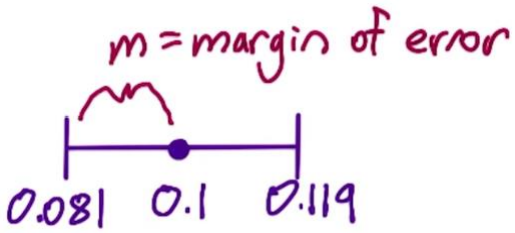
$$A = \frac{1 + CL}{2} = \frac{1 + 0.95}{2} = 0.975$$

If we look at the z-score table and find the 0.975 value, we will see the column header is 0.06 and the row 1.9. If we sum them, we find that 0.06+1.9=1.96, that is the value we were trying to find.



In this case, since we are doing a two-tailed test, each tail will contain 2.5% of the distribution. So, 1.96 is the z-score for 97.5% or 100-2.5.

In our case, the margin of error comes out to about 0.019. So, we'll add this margin of error to the point estimate to get the upper bound of 0.119 and the lower bound of 0.081. This means that if you'd run the experiment again with another thousand pageviews, you'd maybe expect to see between 80 and 120 clicks, but more or less than that would be pretty surprising.



Establishing Statistical Significance

We've estimated the CTP of the Start Now button, but we haven't actually changed the button yet. Once we do that and we run the experiment, how we will analyze the results?

There is a concept in Statistics called hypothesis testing which is a quantitate way to establish how likely it is that your results occurred by chance. So, the first thing we need what we called null hypothesis or baseline. In our case, that's the theory that there is no difference in CPT between our control and our experiment.

Then we need to think about what's called the alternative hypothesis. Are we interested in whether the CTR is different? Or just whether it's higher or lower? Or are we interested in any kind of difference at all?

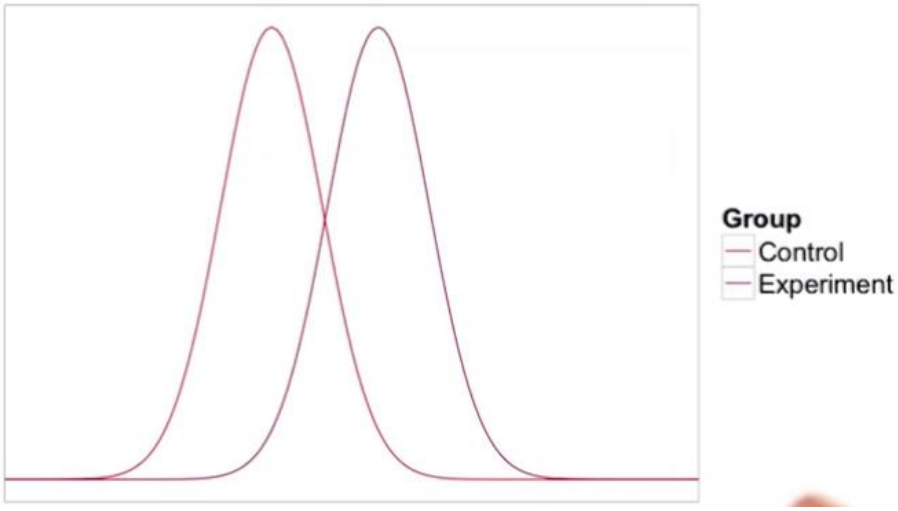
Null and Alternative Hypothesis

To establish that our results are statistically significant using hypothesis testing, you need to calculate how likely it is that your results are due by chance. In order to calculate this probability, you need to have a hypothesis about what the results would be if your experiment had no effect.

In the Audacity example, we will collect two samples: a control group with the orange button and an experimental group with the pink button. We've already assume that

each group follows a binomial distribution, but the probabilities might be different for each group.

We will call the probability that someone in the control group clicks as P_{cont} and the probability than someone in the experiment group clicks as P_{exp}. This potentially give two different distributions. If changing the button color had no effect, then we would expect the two group to have the same probability distribution, so they would be on top of each other. In other words, P_{cont} and P_{exp} would be equal. Another way of saying this is that $P_{exp} - P_{cont} = 0$.



These hypothesis of what our results would look like if the experiment had no effect is called the null hypothesis and it's represented by H_0.

We also need a hypothesis about the results if the experiment does have an effect. This is called the alternative hypothesis and it's represented by H_A. In this case, if changing the button color does have an effect, we expect that $P_{exp} - P_{con} \neq 0$.

Once we've estimated these hypotheses, we can estimate P_{con} and P_{exp} from the data we've collected. Then, we will calculate the difference between these and compute the probability that this difference would have arisen by chance if the null hypothesis were true.

Then, we want to reject the null hypothesis and conclude that our experiment has an effect if this probability is small enough.

This is the same type of significant threshold as a confidence interval, so it makes sense to choose the same cutoff (also called alpha). So, we reject the null hypothesis if this probability is less than 0.05.

Two-tailed vs. one-tailed tests

The null hypothesis and alternative hypothesis proposed here correspond to a two-tailed test, which allows you to distinguish between three cases:

1. A statistically significant positive result
2. A statistically significant negative result
3. No statistically significant difference.

Sometimes when people run A/B tests, they will use a one-tailed test, which only allows you to distinguish between two cases:

1. A statistically significant positive result
2. No statistically significant result

Which one you should use depends on what action you will take based on the results. If you're going to launch the experiment for a statistically significant positive change, and otherwise not, then you don't need to distinguish between a negative result and no result, so a one-tailed test is good enough. If you want to learn the direction of the difference, then a two-tailed test is necessary.

Comparing Two Samples

Now that we've set up a hypothesis test, how will we decide whether to reject the null? Well, we're going to need to look at our confidence intervals in a slightly different way.

In this case, we actually have two samples. We have the control side which may have a different number of users from the experiment side. So, we'll need to compare the proportion of clicks estimated on the control side with the proportion estimated on the experiment side.

Then, the quantitative task tells us whether it's likely that the results we got, the difference we observed, could have occurred by chance or if it would be extremely unlikely to have occurred if the two sides were actually the same.

Pooled Standard Error

Because we have two samples, we'll need to choose a standard error that gives us a good comparison of both. The simplest thing we can do is calculate what is called a pooled standard error. Recall that we'll measure the users who click in each group (X_{con} and X_{exp}) as well as the total number of users in each group (N_{con} and N_{exp}).

The first thing we'll calculate will be what's called the pooled probability of a click which is the total probability of a click across groups, following this formula:

$$p' = \frac{X_{con} + X_{exp}}{N_{con} + N_{exp}}$$

Then, we'll calculate the pooled standard error, which is given by this formula:

$$SE_{pool} = \sqrt{p'_{pool} \cdot (1p'_{pool}) \cdot \left(\frac{1}{N_{con}} + \frac{1}{N_{exp}}\right)}$$

Now, recall that we're going to estimate the difference between P_{con} and P_{exp} :

$$d' = p'_{exp} - p'_{con}$$

Under the null hypothesis d, this true difference is equal to zero.

$$H_0: d' = 0$$

Then, we would expect our estimation, d', to be distributed normally, with a mean of zero and a standard deviation of the pooled standard error:

$$d' \sim N(0, SE_{pool})$$

If d'>1.96·SE_{pool} or d'<-1.96·SE_{pool} , then we can reject the null hypothesis as unlikely and say tat our difference represent a statistically significant difference.

Practical or Substantive Significance

We can use hypothesis testing to see whether a difference we observe in our experiment is actually significant. What comes next? So, what we have to do next is decide from a business perspective what change in the CTP ability is practically significant. In other words, what size change matters to us?

What you want to observe is repeatability. You want to make sure when you setup your experiment that you get that guarantee that yes, these results are repeatable, so it's statistically significant. However, you want also make sure that if you see a change in your experiment that you're interested in from a business standpoint, so it's practically significant. So, you want to size your experiment appropriately, such that the statistical significance bar is actually lower than the practical significance bar.

Size vs. Power Trade-Off

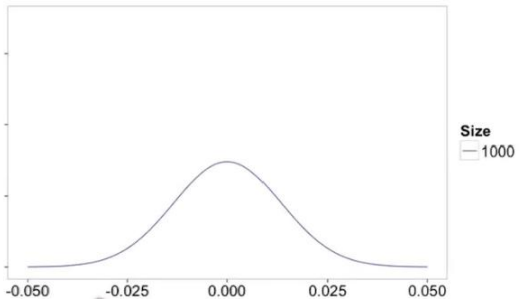
Then, we need to design our experiment. Now, the main question that we have to decide is, given that we have control over how many page views go into our control and experiment groups, how many page views we need to get a statistically significant result. This is called statistical power.

The key thing is that if we see something interesting, we want to make sure that we have enough power to conclude with high probability that the interesting result is, in fact, statistically significant.

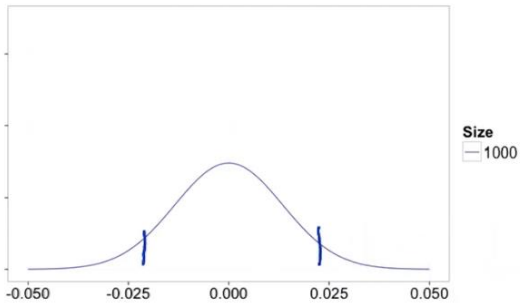
It's important to keep in mind that the power has an inverse trade-off with size. The smaller the change that you want to detect or the increased confidence that you want to have in your result, means that you have to run a larger experiment, so more page views in your control and your experiment groups.

How Page Views Affect Sensibility

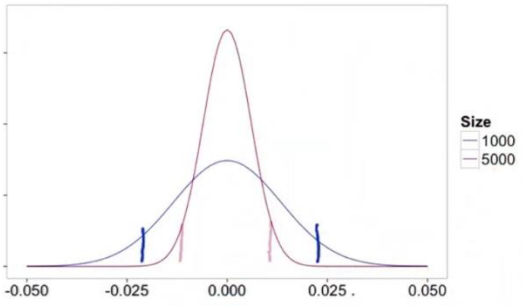
Before you run your experiment, you'll need to decide how many page views you plan to collect before you make your conclusion. First, let's look at how the distribution changes when you increase the sample size. This is what the distribution of results would look like if you collected 1000 samples and there was no true difference between the two groups. That's why the mean is zero.



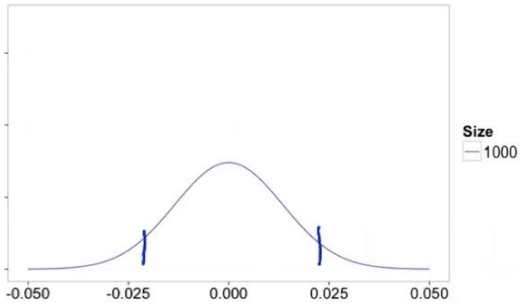
You'll reject the null and conclude that there was a difference if you measure a value less than -0.025 or 0.025. So, your probability of falsely concluding there was a difference, which is often called alpha, is 0.05.



If you increase your number of samples to 5000, then the standard error will decrease, so the distribution of results will look much narrower. To keep alpha the same, that means the cutoffs for rejecting the null will be closer to zero.

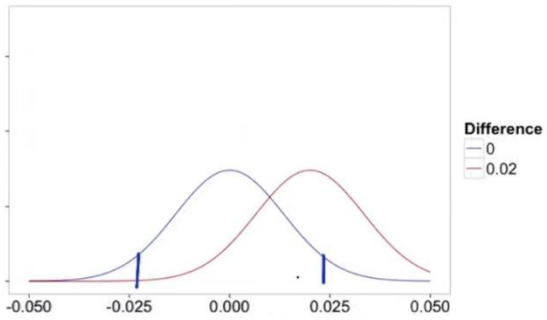


Let's say you decide to only collect 1000 page views, so that you could make a conclusion quickly. Recall that this is the distribution of results if there is no true difference and alpha is 0.05.



Now, let's consider the case where there is a difference. Specifically, the difference is equal to the practical significance level of 0.02. As the maximum of the second probability distribution is inside the confidence interval, then you'll fail to reject the

null and conclude there was not a significant difference in the same cases. Now, the probability of failing to reject the null when the null was in fact false, which is often called beta, is pretty high. Even if there is a true difference, it's pretty likely that you'll be inside this range.

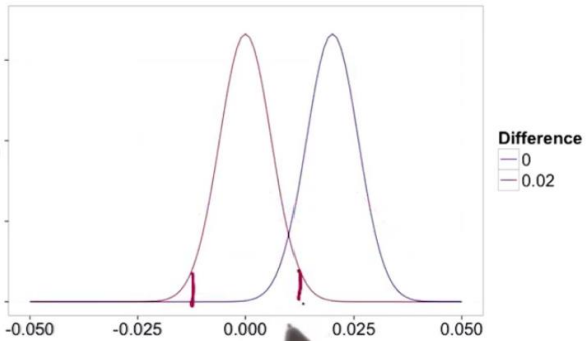


By collecting a small sample, alpha is low, so you are unlikely to launch a bad experiment. However, beta is high, so you are likely to fail on launching an experiment that actually did have a difference you care about. Beta, or the probability of falsely failing to draw a conclusion, depends on how big your effect really was.

For well-behaved distributions like the normal distribution, as your true change gets larger and larger, then beta will go down. So, you typically consider beta at your practical significance boundary since you don't care about any smaller changes, and any larger changes will have a lower beta which means lower chance of error.

People also refer to one minus beta at the sensitivity of the experiment. In general, you want your experiment to have a high level of sensitivity at the practical significance boundary. People often choose 80%.

Now again, let's look what happens if you collect 5000 samples. Both distributions get tighter and alpha doesn't change. However, in the case where there is a true difference, you are much less likely the maximum of the second probability distributions falls within the confidence interval. You are more likely to reject the null and conclude there was a difference. So, beta has gone down and your power has increased.



Calculating Number of Pages Views Needed

We can use an online calculator to get the number of page views we need for our experiment. You can access the online calculator **here**. The explanation on how to use it it's on this video.

How Does Number of Page Views Vary?

Now that you know how to calculate the number of pageviews you'll need for an experiment, I want you to consider the following possible changes to your experiment and decide whether each will increase or decrease the number of page views.

Case 1

Suppose the CTP in your control group had been higher but still less than 50%. To see how the CTP would affect the number of page views needed, recall that the standard error depends on the CTP. Specifically, the standard error follows this formula.

$$SE = \sqrt{\frac{p \cdot (1-p)}{N}}$$

So, for example, if the probability were 0.5, the standard error would be proportional to the square root of 0.5 times 0.5, which comes out 0.5:

$$SE = \sqrt{p \cdot (1-p)} = \sqrt{0.5 \cdot (1-0.5)} = \sqrt{0.5 \cdot 0.5} = 0.5$$

On the other hand, if the probability were either 0.1 or 0.9, then the standard error would be proportional to 0.3.

$$SE = \sqrt{p \cdot (1-p)} = \sqrt{0.1 \cdot (1-0.1)} = \sqrt{0.1 \cdot 0.9} = 0.3$$

It turns out that as your probability gets closer to 0.5 and further away from extremes like 0.1 or 0.9, then your standard error increases. I'll need to increase the number of page views to reduce the standard error back to its original level.

Case 2

Suppose you decided to increase your practical significance level. You would no longer care about the 2% change but you would want to see a 3% in order to launch the experiment. If you increase your practical significance level, you say you no longer care about detecting a 2% change. You would need the change to be larger than 2% before you cared detecting it. Larger changes are easier to detect, so you shouldn't need as many pageviews.

Case 3

Suppose you decided to increase your confidence level. For example, maybe instead of the 95% confidence, you would choose a 99% confidence level. If you increase your confidence level, you're saying that you want to be more certain that a change has occurred before you reject the null. In essence, you're being more conservative. You could accomplish that by rejecting the null less often, but **then your sensitivity would go down**. If you want to keep sensitivity the same, you'll need to increase the number of page views you collect.

Case 4

Finally, suppose you wanted to get a higher sensitivity at a practical significance boundary. If you want to increase the sensitivity of your experiment, you'll need to collect more page views to narrow the distribution.

Change	**Increase Page Views**	**Decrease Page Views**
Higher CTP in control (but still less than 0.5)	X	
Increased practical significance level (d_{min})		X
Increased confidence level (1-alpha)	X	

Higher sensitivity (1-beta)	X	

Calculating Results

Suppose the control group had 10,072 total page views and the experimental group had 9,886 page views. These numbers aren't exactly the same because of how people were randomly assigned to groups.

In the control group, 974 of those users clicked the Start Now button, and in the experimental group 1,242 did. So, it certainly looks like more people clicked in the experimental group, but was this due to random variations?

To answer that, we'll need to calculate the confidence interval for the difference. We'll start by calculating the pooled probability, that is the total number of clicks in both groups divided by the total number of users.

$$p'_{pool} = \frac{974 + 1242}{10072 + 9886} = 0.111$$

Next, we'll calculate the poll to standard error using the following formula:

$$SE_{pool} = \sqrt{p'_{pool} \cdot (1 - p'_{pool}) \cdot \left(\frac{1}{N_{con}} + \frac{1}{N_{exp}}\right)}$$

$$SE_{pool} = \sqrt{0.111 \cdot (1 - 0.111) \cdot \left(\frac{1}{10072} + \frac{1}{9886}\right)} = 0.00445$$

We defined the estimated difference as the experimental probability minus the control probability:

$$d' = \frac{X_{exp}}{N_{exp}} - \frac{X_{con}}{N_{con}}$$

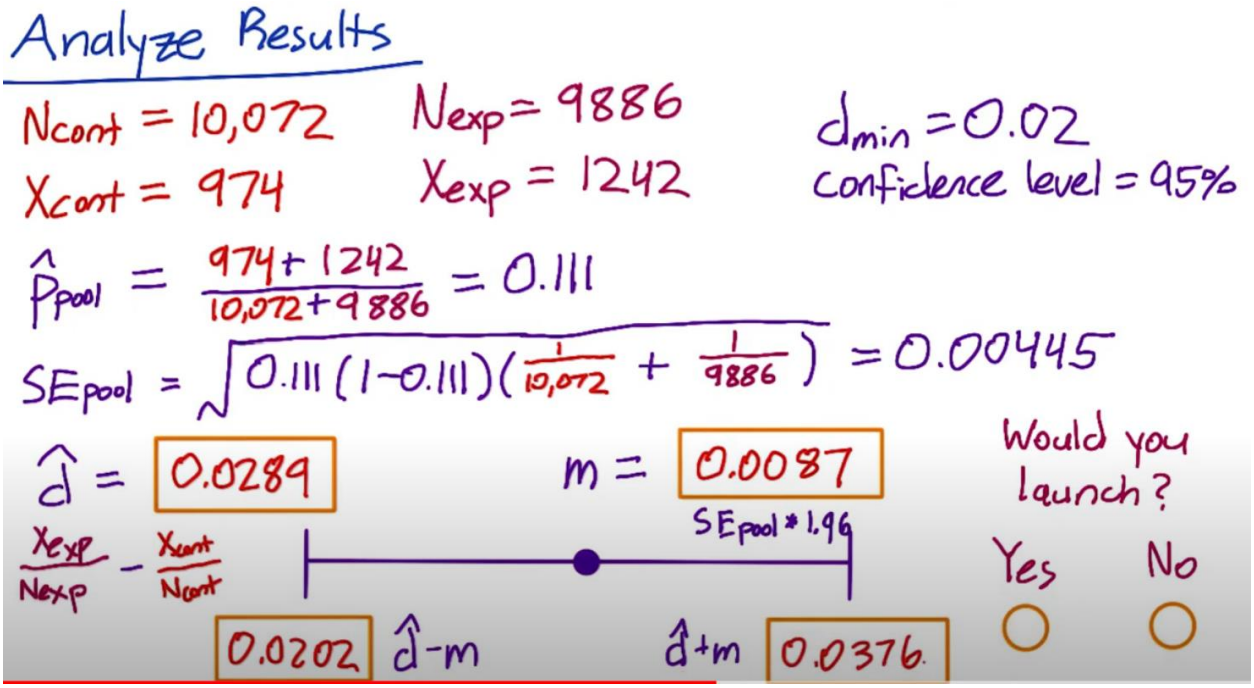
$$d' = \frac{1242}{9886} - \frac{974}{10072} = 0.00289$$

Then, the margin of error is equal to multiply SE times 1.96, which is our Z score for our confidence level of 95%.

$$m = SE_{pool} \cdot 1.96 = 0.0087$$

The lower bound of the confidence interval will be d' - m, which is 0.0202, and the upper bound will be d'+m, which comes to 0.376.

As d_{min} = 0.02 and d'-m = 0.0202, we can conclude that we will have at least a change of 2%, as the interval is between 2.02% and 3.76%.



Confidence Interval Case Breakdown

Now, let's look at some different cases than could have come up in our results.

Our point estimate, that is d', for the example we just went over greater than the practical significance boundary. And, in fact, both ends of the confidence interval were greater than the practical significance boundary.

Based on this, it's highly probable that the click-through probability did, in fact, change by more than the practical significance level. Now, let's go through some other possible cases of what your results could look like and practice recommending a decision in each case.

Case 1

The first case is the one we just saw, when the CTP changed by more than the practical significance level. In this case, we would launch this change.

Case 2

The second result is often called neutral. There's no statistically significant change from 0, since the confidence interval includes 0, and you're also confident that there's not a practical significant change. Given this, it's not worth the effort to launch the change.

Case 3

In the third case, your result is statistically significant. You're confident there was a positive change, but it's not practically significant. In fact, you're confident there was not a practically significant change. There was a change, but you don't care about the magnitude of the change. It's not worth the effort to launch this change.

Case 4

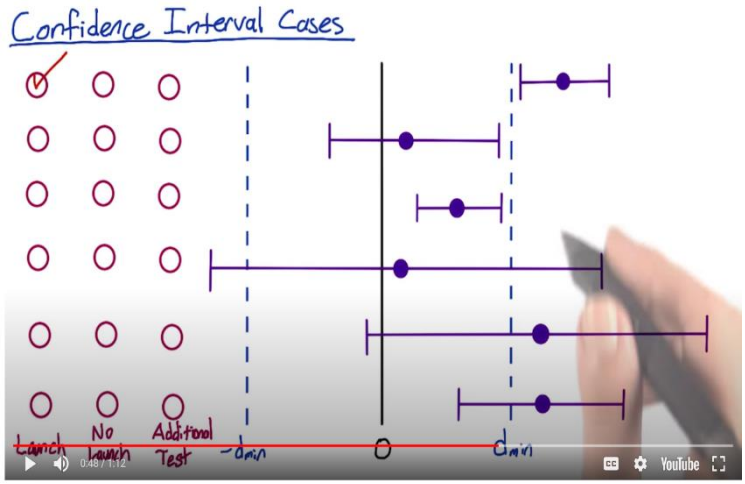
The fourth case is one of the cases that can be the most difficult to handle as an analyst. Some people call this a neutral change just like change two but the confidence interval bounds are outside of what's practically significant. If you ran an experiment and found that it could be causing your number of users to increase by 10% or it could be causing them to decrease by 10%, would you say that change is neutral? Instead, it would be better to say that you do not have enough power to draw a strong conclusion. Running an additional test with greater power would be recommended in this situation, if possible.

Case 5

The point estimate is beyond what's practically significant. This change is an effect you care about. However, the confidence interval overlaps 0, so there might not even be a change at all. Repeating this test with greater power would give additional confidence to the results.

Case 6

There is a practically significant positive change. However, it's also possible your change is not practically significant. How do you make the call? You should run an additional test with greater power if you have the time.



Making Decisions about Uncertain Data

What should you do if one of the last three cases comes up, but you don't have time to run a new experiment? What you have to do is to communicate to the decision-makers when they're going to have a make a judgement, and take risk, because the data is uncertain. They're going to have other factors, like strategic business issues or other elements besides the data.

Choosing and Characterizing Metrics

In this lesson, we'll be focusing on metrics for experiments. We'll first cover how to define a metric, which involves brainstorming and establish a suite of metrics. Then, we'll explain how to build intuition about your metrics and understanding your metrics sensitivity and robustness. We'll conclude on how to characterize the variability of your metric.

Metric Definition Overview

So, the first thing we need to do it actually define a metric or multiple metrics for our experiment. We need to measure whether the experiment group is better than the control group or not.

The first thing to do is to think about what you're going to use the metrics for, before you are going to define them. If you think about how you're going to use the metric, there are really two main use cases.

The first is what we called **invariant checking** or **sanity check metrics**. These are the metrics that shouldn't change across your experiment and your control. For example, the populations of both groups or if the distributions are the same. All these things are sanity checks to make sure your experiment is actually run properly. The second use case is **evaluation**, which breaks down into these overall business metrics and the more detailed business metrics.

However, first you're going to have a couple of different things to think about. First, you have what you think of as high-level business metrics, so that might be how much revenue you make, what your market share is, how many users you have. Then, you're going to want more detailed metrics that focus on the user experience with actually using your product, like how long they stage on your page.

For example, users aren't finishing a class on Audacity. Well, we don't know why, but what we want to do is to dig in to the user experience for that class. Maybe the videos are taking too long to load and we should take a look at latency. Maybe, some of the quizzes are particularly difficult and the students are having trouble with that.

How do we actually go about making a definition? There's really a bunch of different steps. The first is you want to come up with a high-level concept for a metric. This is going to be your one sentence summary that everyone is going to be able to understand, like 'active users' or 'CTP'.

The second step is to really figure out the details. For example, how do you define what active is? Is it a 7-day active? Is it a 28-day active? Which events count towards activity? An automatic notification may not count towards being an active user.

The third step is that you are taking all these individual data measurements, and now ou need to summarize them into a single metric.

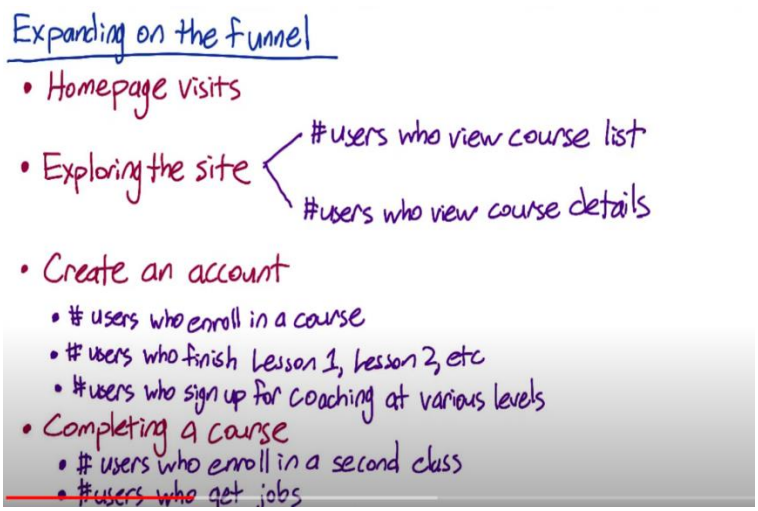
Once you have that summary, now you have an actual complete metric definition.

High Level Metrics: Customer Funnel

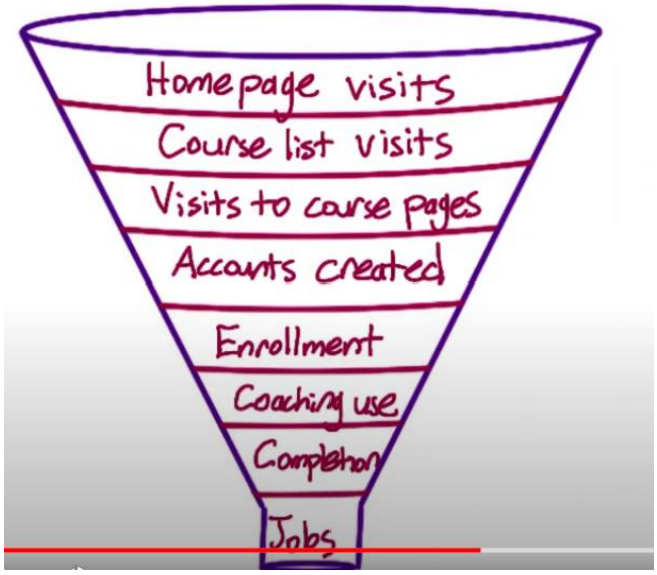
We first need to come up with high level concept for our metrics. A good place to start is with the overall business objective. What Audacity ultimately cares about is helping their students get a job in finance. Of course, they also need their business to be financially sustainable. We can break these objectives down into steps using a customer funnel.

Refining the Customer Funnel

Let's expand the funnel we used in lesson 1:



This will be our new and redefined funnel for this new study case:



Each stage in the funnel right now is a metric that you could use for an experiment.

- **Counts**. For example, the number of users who reach that point in the funnel. This type of metric is called a count. However, sometimes you'll want to use a rate or a probability. Often, you'll want to keep the counts at a few key points in the funnel. For example, the number of people who visit the homepage and the number of people who enroll in courses.

- **Rates.** For each step in the funnel, you might want to stich to a rate by dividing the number of users at that level by the number at the previous level. The reason to use a rate is because you want first to figure out how many people enter the funnel and then increase the rate at which people progress down the funnel.
- **Probabilities.** For each rate in the funnel, you might also care about the probability that a unique user progresses down the funnel.

For all those metrics, they're not yet metrics you could actually use. We start with these high-level concepts. What business objective you're tracking through the funnel, whether you care about the absolute number or the count, the usability (rate) or the progression (probability). We still have to transform these into full definitions.

Choosing Metrics

Case 1

For the updated course description, the primary metric could be the probability progressing from the course list page to the course overview page for that course. It would also be good to track continued progression down the funnel. If the updated description was accidentally misleading, you might see an increased CTR to the course overview page, but then a drop off when students realized the course wasn't right for them. So, it might be good to track the probability of enrolling in the course and maybe even how students are progressing through the course.

Case 2

By increasing the size of the Start Now button, you're making the button more prominent, presumably with the intention of making it easier to find. So, this is a good time to use the CTR of the button. CTP could also be a good choice here, but rates are usually better for assessing how easy is to find a button.

Case 3

For the experiment making the benefits of the paid service more clear, the best metric to use is the probability that the enrolled students start paying for coaching. It would also be good to measure the user retention over various time periods. For example, how long the students continue paying for coaching. You can also look at

usage metrics. For instance, among the students who are using the coaching, do they use the coaching more now that is more clear what coaches do?

Difficult Metrics

Sometimes, there are metrics that are hard to measure because you don't have access to data or it takes too long.

Case 1

Audacity definitely has access to this data. However, students could take a long time between completing one course and singing up for another. So, this metric probably takes too long to measure during most A/B tests.

Case 2

In this case, the company don't have direct access to the data they want. The shopping company doesn't know how happy its users are.

Case 3

The search engine doesn't know whether users found the information they wanted. For this reason, Audacity don't have this data.

Defining Metrics: Other Techniques

These techniques can help you get an understanding of your users, which you can use to come up with ideas for metrics, validate your existing metrics, or even brainstorm ideas of what you might want to test in your experiments in the first place. External Data You might be able to use external data, that is, data collected by people outside of your company.

What External Data Is Available?

There’s outside market share data, provided by companies such as Comscore and Hitwise, which includes things like how many users visit a site—for your site or for competitors or related sites.

1. There are also companies such as Nielsen, Forrester, and Pew Research that run and publish their own studies. For example, you can find studies surveying users on how many devices they use on a given day, or tracking the activity and recording detailed observations about the online usage of a panel of consenting users. Companies may have already done research using a variety of methods to answer questions that you may be interested in. Depending on what industry or type of site you work on, you might be able to find data from other people's experiments that is useful—and your company might already subscribe to these publications.
2. There are higher level aggregators of data, such as eMarketer, who provide summaries from all of these sources in easily consumable fashion.
3. There is also a whole treasure trove of published papers, where researchers have investigated all sorts of interesting questions in a rigorous fashion, either about how users behave in certain scenarios, how metrics are related, etc. Potentially relevant conferences include: CHI, WWW, KDD, WSDM.

What Can You Do with This External Data?

First, you can validate simple business metrics if your site or industry appears in one of these lists. For example, if you want to look at total visitors to your site, you can compare your number with the numbers provided by Comscore or Hitwise, or you could compare the fraction of shopping traffic in each "vertical" category to what you see on your site. However, the numbers you see will almost never exactly match your own data. Generally speaking, a better way to do the validation is to look at a time series of both your internally computed metric and the externally available one, and see if the trends and seasonal variability line up.

Second, you can provide supporting evidence for your business metrics, either direct measurable quantities (look, this is used by lots of sites) or to get ideas for which measurable metrics make good proxies for other harder-to-measure quantities.

Publicly available academic papers, such as the User Experience ones, often establish a general equivalence between different types of metrics. One example that Carrie worked on was a paper with Dan Russell, which compared user reported satisfaction with a search task to the duration of the task as measured on the website.

That gave a good general correlation for satisfaction with duration measured, though with some clear caveats. So, this study helped validate a metric—duration—that

could be computed at scale and then automatically converted to a metric that could not be compute at scale—user reported satisfaction.

Gathering Additional Data

There are three common techniques you can use to gather additional data about your users. They vary along two major axes. Some give more in-depth customized data and other will be possible to run on a greater total number of participants.

- **User Experience Research (UER)**. You can go really in deep with just a few users, often by observing them doing tasks of interest, like taking a lesson of a course. You get a lot of detailed and in-depth information that is useful primary for brainstorming ideas. You can also use special equipment in a UER to get information you couldn't get from your site's data. However, you'll want to make sure that you validate the results of a UER with something like a retrospective analysis.
- **Focus Groups**. You can bring a bunch of users or potential users together for a group discussion. You can talk to more total users than with a UER study, but you can't go as deep with each person. Once you bring the users together, you could show them screenshot or images, you could walk them through a demo, and then you can ask question to elicit feedback. However, you run the risk of group think and convergence on fewer opinions.
- **Surveys**. Surveys are when you recruit a population and ask them a bunch of questions, either online, in person or via telephone. Surveys are pretty cheap to run on a shole bunch of users and the data you get is much more quantitative, but it's not very deep or individually customized. They are useful for gathering metrics that you cannot directly measure, like how many students who take Audacity classes get jobs, how productive they are at their job, and whether the classes contributed to their success. However, users don't have to tell the truth and the answers can be dependent on how the questions are phrased.

Other Techniques

Now let's discuss how Audacity could apply these other techniques to either validate metrics from the customer funnel we talked about previously or brainstorm new metrics that are not covered by the funnel.

At the top of the funnel, Audacity could compare the count of how many users visit their website to externally available metrics from companies like Commscore or Hitwise. Audacity could also look at the completion rate of classes and compare that to externally available data.

If Audacity sees that a particular lesson is getting a very low completion rate, they might do a UER study to investigate that. Watching the students try to complete the lesson can help you figure out if they find everything on the screen are they progressing in order to know how they interact with a coach. You might see users twiddling their thumbs waiting for a video to load and that might give you an idea that you should be tracking the latency.

If You observe that people aren't seeing the instructor's notes below the video, then you might want to use percentage of people who click on a link as a metric. Or if latency looks like an issue, then that's the metric you could track for.

Any metric that you come up with during a UER session you might want to validate that metric with a retrospective analysis to see how that metric is varying over time or you might want to run some new experiments and see how that metric varies as you make changes

Finally, the bottom level of the funnel whether students get jobs and if they do the Audacity classes help is an example of an unmeasurable metric. In this case, surveys might help, maybe via email or on repeat students. For example, you could ask students if the material covered in class was touched on in any interview questions they were asked.

Like I mentioned before surveys are often useful for capturing metrics like these but it's not possible to measure directly. The main issue though is that you can't directly compare the numbers from a survey to what you get from any other technique. The biggest issue here is that the populations for your internal metrics and your surveys might not be comfortable.

You might be reaching a biased population with your survey relative to the population taking the class. Let's also go over the cases you saw before of metrics that were tricky to measure and go over some other techniques you could use I had mentioned before.

If Audacity wants to measure the rate of students returning to take a second course, they could definitely track this metric long-term but they might not be able to use it

for individual experiments. So, Audacity might follow up with a survey to see what causes users to return. Then, if they can find something measurable that predicts returning, they could use that as a proxy in their experiments for the shopping site trying to measure the average happiness of shoppers.

This is a metric that they probably can't track even long term but again they could try to find signals that correlate with what they really want. They could instrument their website to pop up the survey at the end of every purchase or they could run a small UER and use this data to brainstorm some metrics they could use.

For the search engine that wants to measure whether users are finding the information they're looking for there are a lot of possible proxies they might be able to use. For example, the length of time spent on the search page whether the user clicked on any results that were shown or whether there were any follow-up queries trying to get it the right information in a different way. You might be able to identify which of these proxies are more promising by looking at external data about information.

Metric Definition & Data Capture

Now, we'll step through an example of turning a high-level metric into a well-defined metric.

In lesson 1, we have chosen the high-level metric CTP. We defined this as the number of unique users who clicked the button divided by the number of unique users who visited the home page. Now, this is not actually a completely specified metric yet and there are some other possible definitions.

Case 1

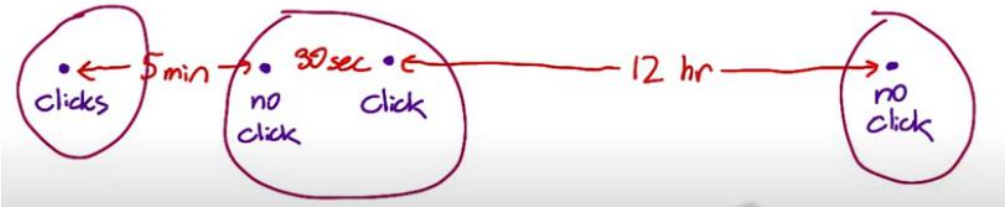
First, to use this definition, we need some way of determining whether two events are from the same user. Let's say we use cookies. Next, if the same user or cookie visits the page once and then comes back a week later or two, do we really only want to count that once? Usually, you'll want to count those visits separately, which means you'll also need to choose a period of time. Do you only count one page view per user each minute, hour, day or what?

So, one fully specified definition would be that for each minute, you take the number of cookies that clicked during that minute divided by the number of cookies that interacted with the page at all during that minute. Now, this definition actually leads

to a whole family of definitions since you can choose any time interval. Changing this time interval could give you a different answer.

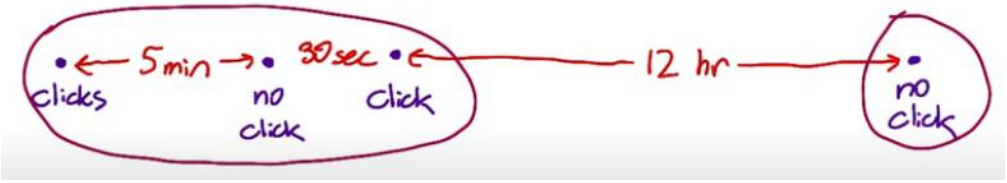
For example, suppose your site only has one user who visits the site and clicks. Then, 5 minutes later, they reload the page but don't click. 30 seconds after that, they reload the page again and do click this time. And then, 12 hours later, they reload the page one last time and don't click. Assume this user keeps the same cookie throughout this process.

If you group cookies by minute, there are three separate groups where these two events are in the same group.



The second page view after 30 seconds and the click are in the same group since they happen within a minute of each other. Since two of these groups resulted in a click and the third did not, the per-minute CTP probability is two-thirds.

Similarly, if you group by hours, there are two groups, one of had a click, so the probability is one-half.



If you group by day, everything goes in the same group and the probability is one.

Case 2

An alternative definition would be to remove the idea of a unique user and instead create a unique ID for each page view. Then, when a user clicks, record the idea of the corresponding parent page view. Then, you could define the CTP as the number of page views that eventually result in a click divided by the number of page views.

This data capture is usually easier than recording cookies and then later grouping by cookie. Now, this definition also needs a time interval. How long do you want to wait after each page view to see if it resulted in a click? Once you pick a time interval, you would count how many page views had a click within the specified time.

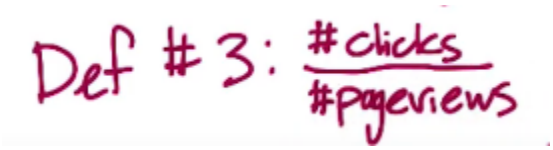
One way these two definitions could give different results if the user refreshes the page within the given time period. For example, suppose the user visits the page without clicking. 30 seconds later, they refreshed the page which generates another page view. Once second after the second page view, they click.

Then, using definition one with a time interval of one minute so that all these events are grouped together, there would be 1 cookie that clicked divided by 1 total cookie for a probability of 1.

However, for definition two with the same time interval, only 1 pageview resulted in a click over 2 total page views, so the probability would be one-half.

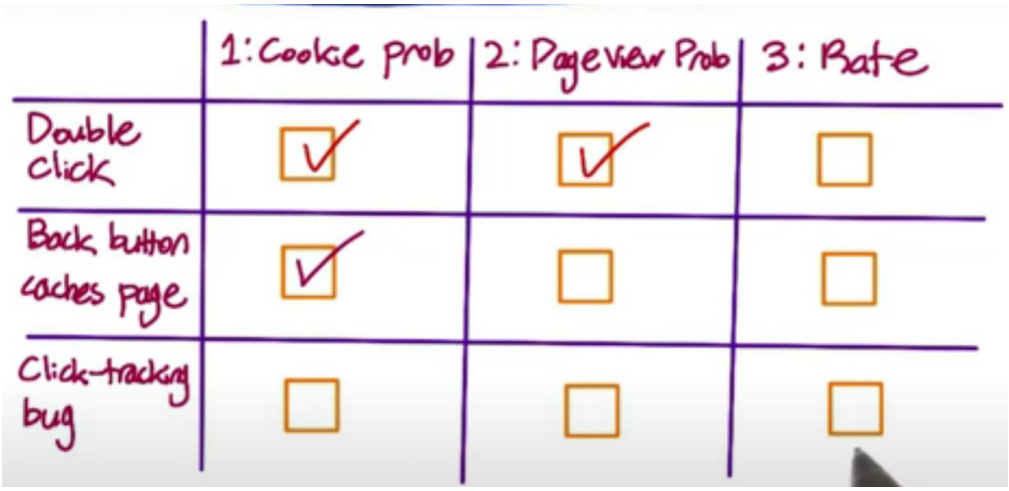
Case 3

Finally, an even simpler definition would be to count the total number of clicks and divide by the total number of page views. As you know, this would be a CTR rather than a CTP. Which definition you use will depend on your product. Often, definitions one and two will be almost indistinguishable if you choose the same relatively short time interval. So, you might want to go with definition two, since it's easier to compute.



Now, for each of the three definitions we just discussed, it would be affected by the following problems:

- **Double click.** Suppose you want your metric to be unaffected by the user double clicking. You want your metric to have the same value of the user double clicks as if they single click.
- **Back button caches page**. When the user clicks the back button, the browser may have cached the page, so another page view may or may not be generated.
- **Click-tracking bug**. Suppose you are worried that JavaScript may not be tracking clicks correctly. For example, it might send two clicks instead of one or it might fail to record a click at all.



In the first case, both probabilities would give the same result, regardless of whether the user double clicked or single clicked. For example, suppose a user load the page, which generates a page view. One minute later, they click twice, half a second apart.

Assuming that the time interval is longer than a minute, say five minutes, then the cookie definition and the page view definition probabilities will both count the single unique click and give a result of one.

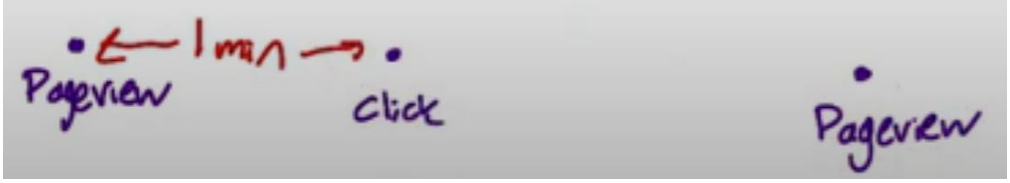
The CTR, on the other hand, would be two, since both clicks would be counted separately, which is a different answer than if the user had single clicked.

In the second case, only the cookie definition would give the same answer either way. Compare these two timelines. In the first timeline, the user loads the page, which generates a pageview, then navigate away. Then, one minute later, they use the back button to return, and it generates another pageview. And click a minute after that.

The second timeline is the same, but in this case, the browser cached the page, so the second pageview is missing. Thus, by looking at your data, all you can really see is that there was pageview and then two minutes later, there was a click.

The cookie probability will calculate one unique click and one user in both cases. The pageview probability will also calculate a unique click, but in the first case there will be two pageviews in the denominator. In the second case, there will be only one, giving a different result. The rate will also calculate a different number of pageviews in the denomiator.

In the last case, if two clicks are recorded instead of one, this is the same as double clicking. So, the rate will be affected and the probabilities will not. However, suppose, instead, that a click is completely missed. So, instead of recording one pageview with a click a minute later, just record a page view with no click. All three definitions will give one in the case where there was a click and zero in case there was not.



Filtering and Segmenting

Looking at different segments of your data can also be useful for evaluating metric definitions, since you can look at how the different definitions vary by segments. This exploration is useful when building intuition about your data and your system.

For example, let's look at total active cookies over time, since that's an important high-level business metric. So, this plot shows the number of active cookies per day for the past four weeks.

As you can see, there is a weird spike that showed up sometime last week. You can also see some weekly variation. The number of visitors looks higher over the weekend.

One way I can verify whether the spike is odd is by looking at a week-over-week plot. We'll divide each data point by the corresponding data point from a week ago. As you can see, that tends to smooth out the weekly variation. So, if I had wondered whether one of these spikes was higher than usual, I can see looking at this plot that it would stand out if one of these abnormally high given what day of the week it was.

Since we see that the big spike is still here, that makes it clear that this spike was not due to weekly variation.

By the way, this corresponding drop, a week later, happens because we divided this data point by the spike that occurred a week earlier.

Another good thing to look at is year-over-year data. This is the same as week-over-week, but dividing by the numbers from a year ago. That way, if there's an annual conference or something causing the spike, it will disappear. But, as you can see, the same spike is still here, meaning it's probably not due to a yearly variation. You can also see the weekly variation is back, since the day of week is not quite matched up to the day of the week from a year ago.

Now the question is: If we can pin down what's causing this spike, since it doesn't seem to be caused by either weekly or yearly variation. One way we can figure this out is by looking at this metric across different segments of our population to see if one segment is causing the spike.

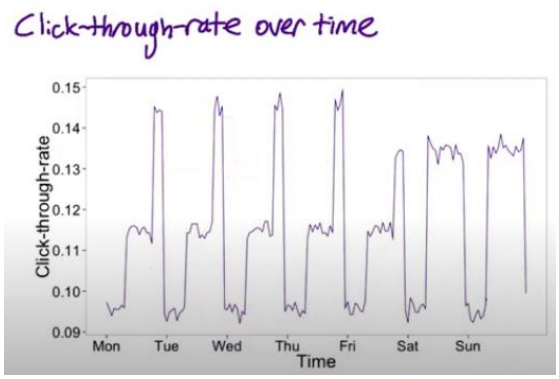
So, let's try looking at how this metric varies by country. What's interesting here is that we don't see the spike in most countries, but we do see it in Berzerkistan, so that one country was causing the entire spike.

At this point, it's a good idea to talk to the engineering team and maybe they'll be able to figure out if this spike is in fact caused by only a small number of rogue IP addresses. This is pretty likely to be spam, a row grow bot or some competitor trying to get information about classes or something of that nature.

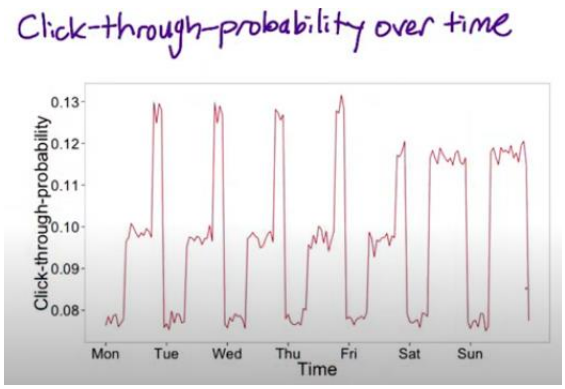
Now, suppose that you suspect there is an issue with JavaScript click tracking. Specifically, you're worried that JavaScript is counting each click event twice on mobile but not on desktop. Which of the following graphs would confirm this problem if it existed?

- **CTR over time**. The graph of CTR over time might look something like this. Now, there are some interesting patterns on this graph. You can see both weekly and daily variations as people are more likely to click through either in

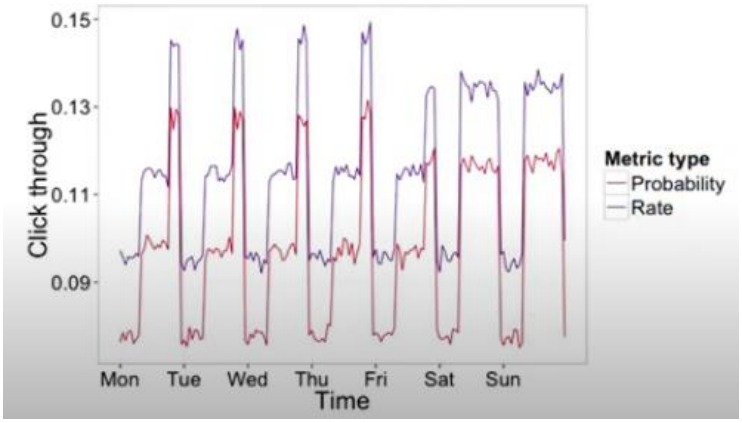
the evening or on the weekend. But this graph can't help us answer our original questions.



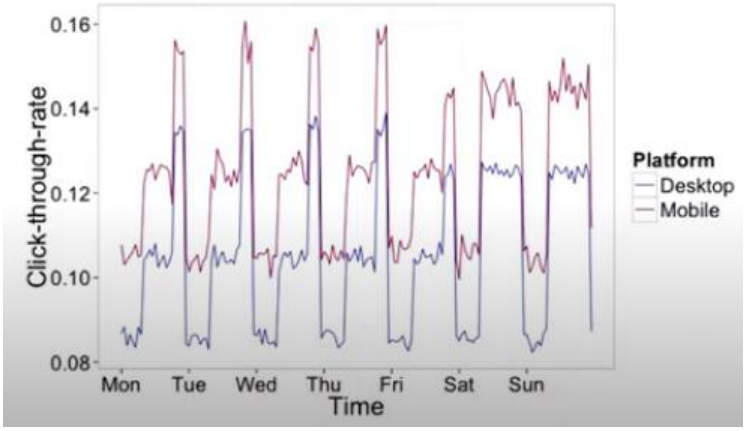
- **CTP over time.** The CTP over time might look something like this. It looks pretty similar to the rate. And again, it's hard to tell by this graph alone whether there's a problem.



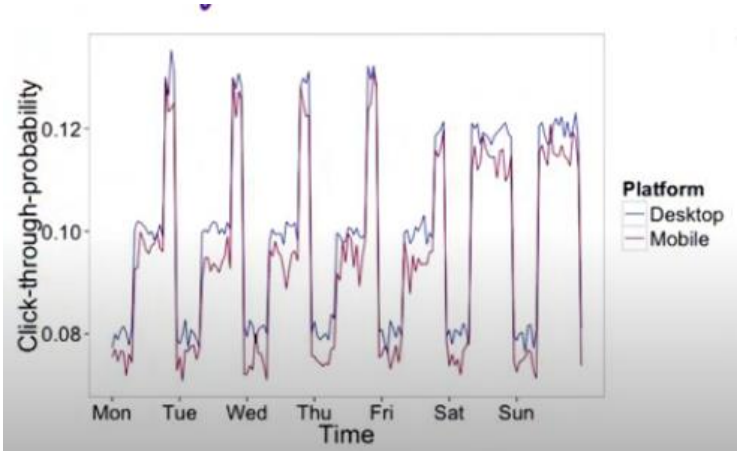
- **Both rate and probability on the same graph.** Now, let's take a look at both the rate and the probability on the same graph. You can see the rate is consistently higher than the probability. But that's to be expected whether there's a problem or not, and it's hard to know how much higher you would expect the rate to be. The difference is also consistent over time, so it's hard to see whether there is a problem.



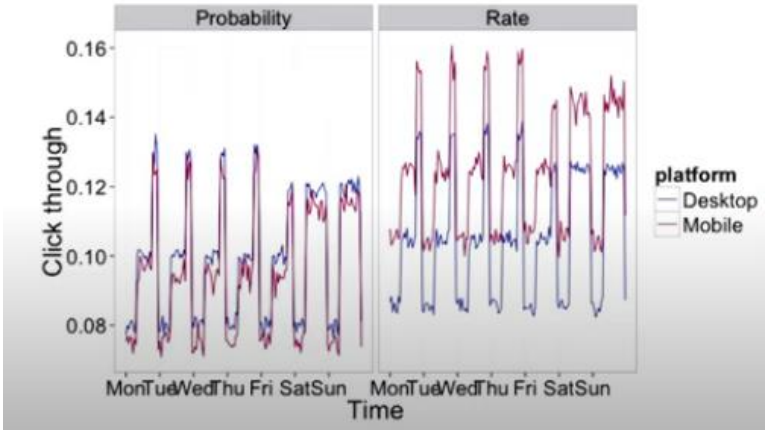
- **CTR by platform.** Now, let's take a look at the CTR by platform. Now, this graph is pretty suspicious because it shows the CTR being higher on mobile than on desktop. But users will behave different on desktop than on mobile, so it's still not clear that this is a problem with the JavaScript tracking and not just a difference in user behavior.



- **CTP by platform.** Now, if we plot the CTP by platform, you won't see anything suspicious since if JavaScript does send a duplicate ping, the CTP will eliminate that, collapsing it into one. So, here we see that the probability is fairly similar between platforms, which is maybe what you would have expected. So, again, we can't tell whether there was a problem.



- **Both rate and probability by platform.** If you plot both rate and probability by platform, you can clearly see the problem. The CTP is actually slightly lower on mobile, but the CTR is significantly higher. You won't necessarily be able to narrow it down to duplicate JavaScript pings at this point, but this point's pretty clearly to some sort of instrumentation issue.



Summary Metrics

What we want to do at this point is to summarize all of these individual events into a single summary metric. In some cases, the summary metric is really obvious. For example, if you're counting how many cookies are visiting the homepage, it's a count. You can also summarize that further into something like the average numbers of visitor per week. Now, for a rate or a probability, if you actually look at the specific computation, it's actually computing an average over the clicks per page view for every single individual effect. In all those cases, the summarization is actually part of the metric definition.

You're going to stablish a few characteristics for your metric. The first one is going to be the sensitivity and robustness. You want your metric to be sensitive enough in order to actually detect a change when you are testing your possible feature options. The second thing is what distribution of your metric looks like and that's going to help you choose.

The most ideal way to characterize a distribution is to do a retrospective analysis and to compute a histogram. When you plot the histogram, you get a shape that will be your distribution. For example, if it's a very normal shape, then the mean or median will make a lot of sense. As it becomes more one sided you might want to go more for 25th, 75th or 90h percentile.

What are the general categories of metrics you like to keep in mind? There's four main categories:

- **Sum & counts**. For example, how many cookies visit the homepage.
- **Distributional metrics**. For example, means, medians, percentiles.
- **Probabilities & rates**.
- **Ratios**. They can compute a whole range of different business models you may care about. However, they can be very difficult to characterize. For example, the probability of clicking in a revenue generating click divided by the probability of clicking in any link

Summary Metrics: Example

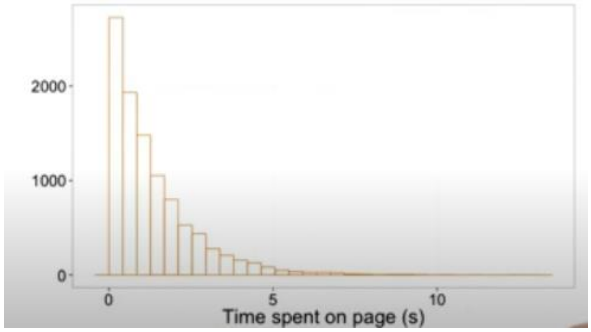
For this example, let's focus in on the second category and see how we can choose between different options looking at the distribution of our data. Let's look at an example of some simulated user data about how long users spend on a particular parge of the Audacity site.

Here, you can see the histogram of the time spend on the page where each bar is the frequency. So, out of 10,000 data points, a little less than 3,000 fall in this first bucket. The mean of this data is about 1,3 s. The median, on the other hand, is about 0.9. That means 50& of all users spend less than a second on the page.

This reason for this difference is all these points off to the right. They increase the mean proportional to their size, so if one user stays for 6 seconds, then that

outweighs a few users who don't stay long at all. This is an example of a exponential distribution.

If you are thinking about something like how many users really get information from this page, you might want to use something besides the median or the mean, like the 75th or 90th percentile.



Common Distributions in Online Data

Let's talk about some common distributions that come up when you look at real user data.

For example, let's measure the rate at which users click on a result on our search page, analogously, we could measure the average stay time on the results page before traveling to a result. In this case, you'd probably see what we call a Poisson distribution, or that the stay times would be exponentially distributed.

Another common distribution of user data is a "power-law," Zipfian or Pareto distribution. That basically means that the probability of a more extreme value, z, decreases like 1/z (or 1/z^exponent). This distribution also comes up in other rare events such as the frequency of words in a text (the most common word is really common compared to the next word on the list). These types of heavy-tailed distributions are common in internet data.

Finally, you may have data that is a composition of different distributions - latency often has this characteristic because users on fast internet connection form one group and users on dial-up or cell phone networks form another. Even on mobile phones you may have differences between carriers, or newer cell phones vs. older text-based displays. This forms what is called a mixture distribution that can be hard to detect or characterize well.

The key here is not to necessarily come up with a distribution to match if the answer isn't clear - that can be helpful - but to choose summary statistics that make the most sense for what you do have. If you have a distribution that is lopsided with a very long tail, choosing the mean probably doesn't work for you very well - and in the case of something like the Pareto, the mean may be infinite!

Why Probabilities and Rates are Averages

To see how probabilities are really averages, consider what the probability of a single user would look like - either 1 / 1 if they click, or 0/1 if they don't. Then the probability of all users is the average of these individual probabilities. For example, if you have 5 users, and 3 of them click, the overall probability is (0/1 + 0/1 + 1/1 + 1/1 + 1/1) divided by 5. This is the same as dividing the total number of users who clicked by the number of users, but makes it clearer that the probability is an average.

A rate is the same, except that the numerator for each pageview can be 0 or more, rather than just 0 or 1.

Sensitivity and Robustness

You want to choose a metric that picks up changes that you care about but is robust against changes that you don't care about. So, it doesn't move a lot when nothing interesting has happened.

The idea is that the mean is sensitive to outliers. So, if in your data you see a lot of cases of really long load times for the page, maybe due to something going on in the users machine or a bad network connection, then you want to maybe not choose the mean because it is going to be pretty heavily influenced by those types of observations. That's called not being robust.

In the other hand, you could choose the median, which tends to be much more robust to that type of behavior. However, if you only affect a fraction of your users, even if it's a fairly large fraction, even a 20%, you might not see the median move at all. In this case, the median is robust but, in this case, you might consider using other statistics.

How would you measure the sensitivity and robustness? There are two main ways. The first has to do with running experiments or using experiments you already have. So, in our latency example, we could run a few simple experiments where we

increase the quality of the video which should, in theory, increase the load time for users. And we could see if the metrics were are interested in actually respond to that.

We can also use what we called A vs A experiment to determinate if they're too sensitive. That's an experiment where you don't change anything. You just compare people who saw the same thing to each other. And you see if your metric pick up any spurious difference between the two.

The second main category of things you can look at is sort of a retrospective analysis of your logs. If you don't have experiment data or if you can't run a new experiment, then you can look back at changes you made to your site and see if the metrics you're interested in actually moved in conjunction with those changes.

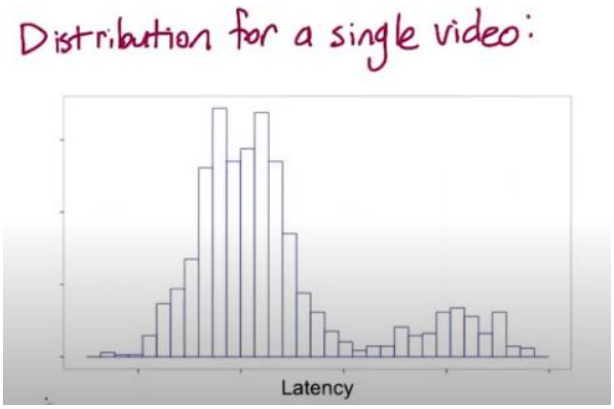
Or you can just look at the history of the metric and see if you can find a cause for any major changes that you see.

Measuring Sensitivity and Robustness

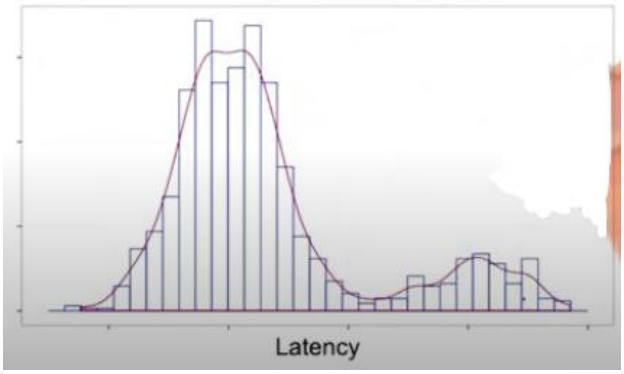
Let's talk about how to measure the sensitive and robustness of some different metrics. Specifically, we'll try to choose a summary metric for the latency of a video. There are various different summary metrics you could use. So, we'll look at the sensitivity and robustness for mean, median and percentiles.

We could start out by either looking at the results of a bunch of experiments or by doing a retrospective analysis. Let's do the retrospective analysis first. What we might do is de-segment the data by different videos. In other words, look at the distribution of load times per video.

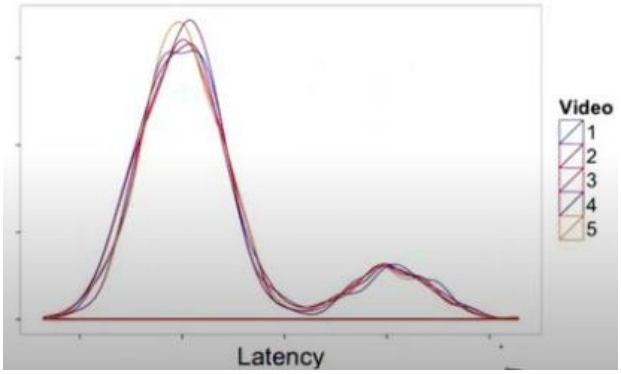
If we wanted to look at the distribution of a single video, we could plot it as a histogram like this. However, this can get hard to see when we have multiple different videos to compare.



Instead, we could draw something called a density line over the histogram that roughly approximates the shape of the histogram.



Then, we could plot only this density line for several different videos to compare. If we do that, we get something like this:



Suppose that I've picked five roughly comparable videos of the same size, so I get a roughly similar distribution of load times for the different videos. You can see two peaks, a fairly long load time and then more people with a shorter load time. This

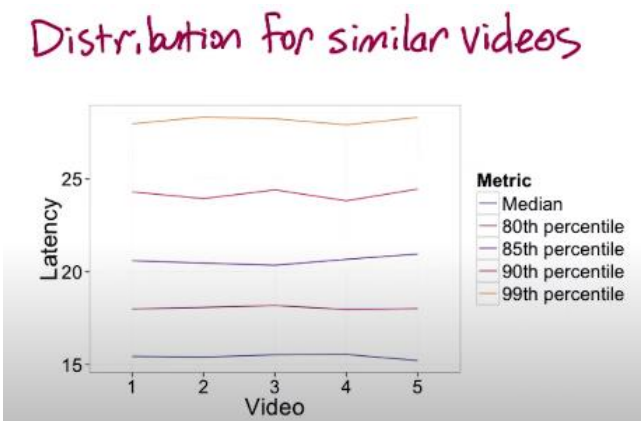
could have if you had people with different types of internet access, a slower internet access an a faster one.

In order to characterize the sensitivity and robustness of different summary metrics, you can see how they vary across videos.

You can find here a few different summary metrics by video. In theory, as these videos are all comparable, there should be not too much difference between the different videos for a good metric.

We can see, the median, the 80th and 85th percentiles don't move around too much. They're pretty good. However, the 90th and 99th percentile are zigzagging around a bit. This is a good indication than the 90th and 99th percentile are not robust enough as summary metrics, since they're moving around quite a bit, even for videos that are pretty comparable.

Of course, you have to be careful. Maybe these metrics are moving around for some other reason because the videos aren't actually comparable. For example, maybe the videos are at different resolutions or have a different encoding scheme.

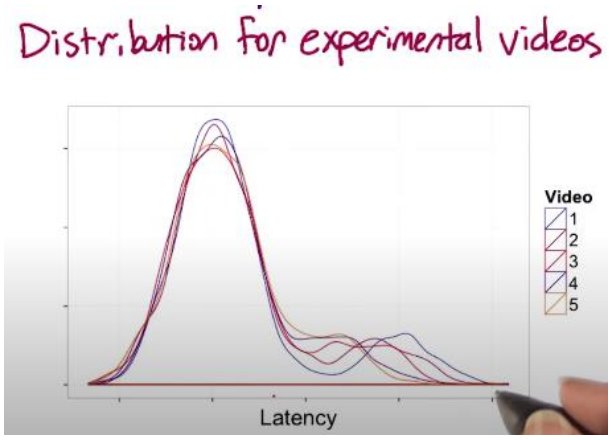


In this case, let's say we're pretty sure that these videos are comparable and that we've checked for those things. However, in general, if you think your metric might be too sensitive, then it's a good idea to dig and ensure that there's not some underlying factor that you haven't taken into account.

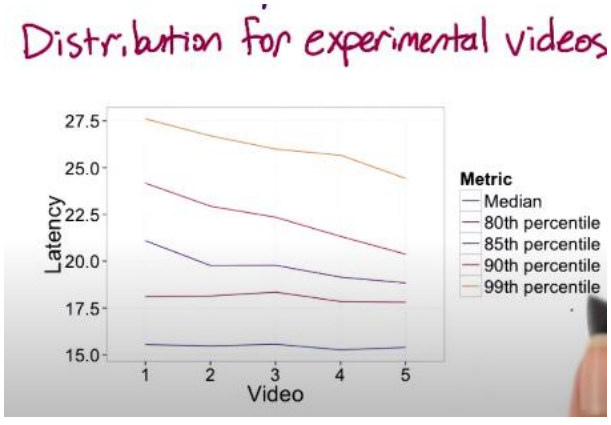
The other technique we can use is looking for previous experiments. For this example, it would be great if we had experiments that changed the resolution. That should impact latency and if it doesn't, then our metric isn't sensitive enough.

Let's take a look at data from five different experimental groups that have a range of resolutions. Video 1 has the highest resolution, which means it should have the highest load time. In fact, you do see that video 1 is off to the right a bit more.

You can also see that people who already have the slow internet are a lot more affected by the resolution than people with the faster connection type.



Now, let's also look at the same summary metrics for these experimental videos. What we should see is the latency going down as we increase the video number (those have a lower resolution). In factor, for some of these metrics we do see that, but median and 80th percentile, they don't really seem to be moving. This is a good indication that the median and the 80 percentile are not sensitive enough. They're not showing a change when we do make a change that we care about.



Absolute or Relative Difference?

Now, we need to actually decide how we're going to compute the comparison. We have a value for our experiment and another for your control group. We have to

actually decide how we are going to compute the comparison between the experiment and the control.

Couldn't you take the difference like we did in lesson 1? The simplest way is just to take the difference. If you're just getting started with experiments or you're building up your knowledge of a whole bunch of different metrics, that's probably the way to go.

However, if you are running lots of experiments, you may want to consider computing the relative change (percentage change), as opposed to the absolute change.

The main advantage of computing the percent change is that you only have to choose one practical significance boundary to get stability over time.

The main situations that I really see this being applicable are basically with regards to seasonality. Let's say you have a shopping site and in June they're not shopping a lot. So, you have fewer users, so you probably have a lower CTR. However, in December, you've got loads of users and your CTR increases. If you have the same practical significance boundary and across the same times, you can basically have the same comparison.

If you are actually running lots of experiments and your system is changing over time, your metrics are probably changing over time as well. Again, if you are using the relative difference, you can stick with one practical significance boundary as opposed to having to change it as your system changes.

The main disadvantage is really variability. Ratios, such as relative differences, are not always as well behaved as absolute differences. If you're just starting out with this or if you have some metrics, you don't understand that well, it's often good to start with the absolute different.

Absolute vs. Relative Difference

Suppose you run an experiment where you measure the number of visits to your homepage, and you measure 5000 visits in the control and 7000 in the experiment. Then the absolute difference is the result of subtracting one from the other, that is, 2000. The relative difference is the absolute difference divided by the control metric, that is, 40%.

Relative Differences in Probabilities

For probability metrics, people often use percentage points to refer to absolute differences and percentages to refer to relative differences. For example, if your control click-through-probability were 5%, and your experiment click-through-probability were 7%, the absolute difference would be 2 percentage points, and the relative difference would be 40 percent. However, sometimes people will refer to the absolute difference as a 2 percent change, so if someone gives you a percentage, it's important to clarify whether they mean a relative or absolute difference!

Variability

Up until now, we've talked sort of about developing intuition for the metric, about sensitivity and robustness. However, now we're going to need a really more rigorous statistical definition of variability.

We also want to check that the practical significance level we're interested in is really realistic for our metric. If we have a metric that varies a lot under normal circumstances, that may not really work for us in practice because the practical significance level we're interested in just may not be feasible with this metric. How would we figure that out?

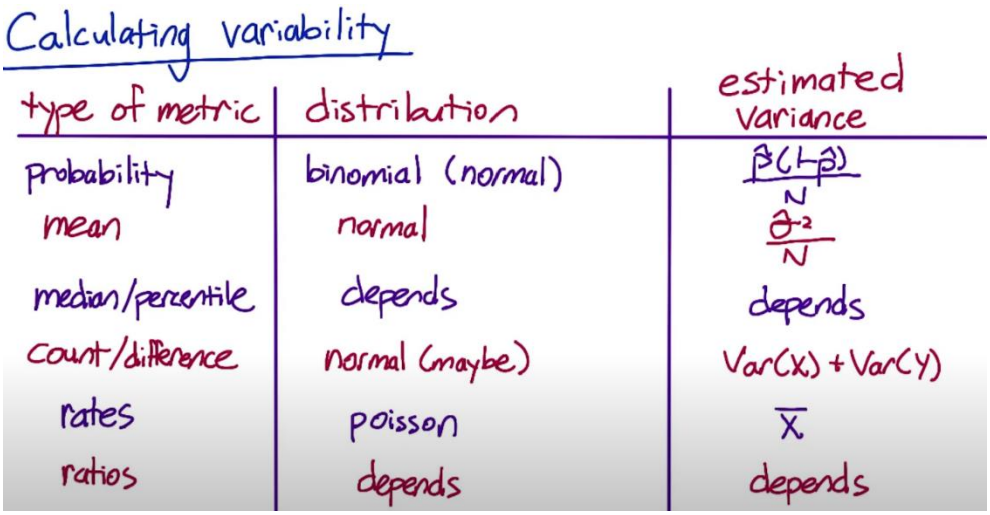
In lesson one, we did a simple example of our CTP, where we looked at user data which was whether the user clicked on a specific link or not. Our summary statistic for this case was the overall CTP. In that case, we were able to do an analytic or a theoretical computation of the variance that we expected from our overall probability.

Now, for other types of metrics, the same thing works. For example, if you have nice normal data, like demographic data, you have counts or probabilities, then usually you can do the confidence interval.

However, in some other cases, you may actually have to do this another way. So, if you're using something like a count of a probability, then you're only really dealing with the variability of a single measurement, or of a constrained one, in the case of probability. If you move on to using ratios or percentiles, or if your data is pretty lumpy, then you probably want to actually compute the variability empirically.

Calculating Variability

We've looked at a lot of the distributions you might see in your data and used that information to choose a specific summary metric or to understand the sensitivity and robustness of your summary metric. Now, it's time to add a level of rigor.



In order to calculate a confidence interval for your metric, you are going to need to things. On one hand, you'll need to know the variance or equivalently the standard deviation of your metric and you'll need to know its distribution.

In lesson one, we calculated the confidence interval for a probability metric, which we assumed followed a binomial distribution. Then, we calculated the standard error, which was our estimate of the standard deviation using the following formula:

$$SE = \sqrt{\frac{p \cdot (1-p)}{N}}$$

Then, we calculated the width of the confidence interval or the margin of error the following way:

$$m = z \cdot SE$$

So, we definitively needed to know the standard error. We used the fact this was a binomial distribution in two ways. First, we use the fact that this was a binomial distribution to get the former formula for the standard error. Also, this formula for the margin of error depends on the assumption that this is a normal distribution. Remember that the binomial approaches a normal distribution as N gets larger.

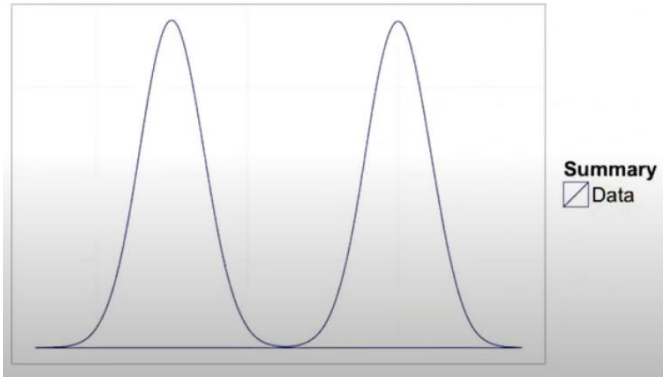
So, to summarize, if your type of metric is a probability metric, you can usually assume a binomial distribution, which approximates to a normal for a large enough sample size and the estimated variance will be the standard deviation squared we saw in the previous formula.

Now, let's talk about some other types of metrics you might see. If you metric is a mean, then by the central limit theorem, your metric will follow a normal distribution if the sample size is large enough. The variance of the metric will be given by the variance of the sample:

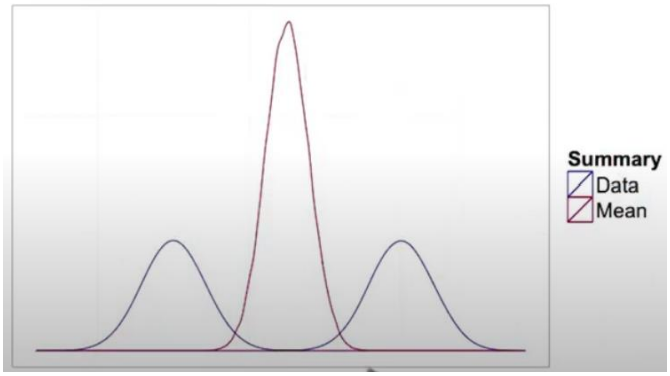
$$SE = \frac{\sigma^2}{N}$$

This estimated variance is an analytic result. For other types of summary metrics, estimating the variance analytically may not be so easy. A good example of this is the median. If the underlying data is normal and the sample is large, then the median will be approximately normal. However, if the underlying data is not normal, then the median might not be normal either.

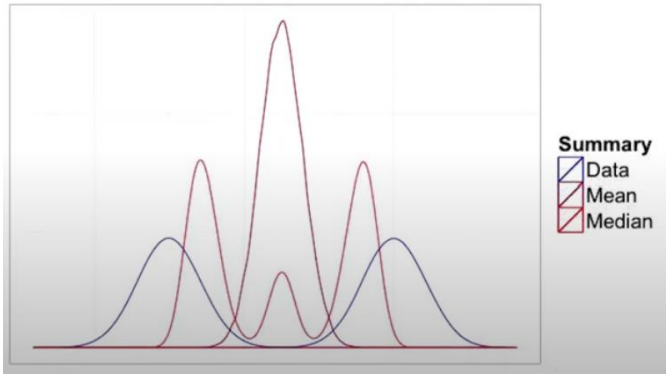
Suppose, for example, you're considering latency and your latency is a bimodal distribution since some people have a slower internet connection and some people have a faster type.



Then, for a large enough sample, the mean would be normally distributed. Where this blue line is the distribution of the underlined data and the pink line is the distribution of the mean.



The median, on the other hand, might not be normally distributed. In this example, you can see it looks pretty crazy.



Other percentiles besides the median might also not be normally distributed. To estimate the variance of the median, you'd need to make an assumption about how the underlying data was distributed. Depending on this assumption, it might not be easy to estimate the variance analytically.

So, for a median or other percentile, the distribution of the metric depends on the distribution of the underlying data and your estimate for the variance will depend on this assumption as well.

Another type of metric that's often easier to analyze is a count or rather you're usually looking at the difference between two counts, like the experiment group minus the control group. This difference won't always be normally distributed, but it often is especially for things like demographic data. The estimated variance of the difference is the sum of the variances of the individual variables:

$$SE = Var(X) + Var(Y)$$

Other types of metrics, such as rates, have a more unusual distributions, but they can have analytic solutions. Rates tend to follow a Poisson distribution and the variance of the Poisson distribution is equal to the mean. So, you could estimate it by taking the mean of your sample

$$SE = \bar{x}$$

For your experiment though, you would be interested in the difference between two rates. So, you would need to estimate the variance of the difference between two Poisson distributions. Or, alternatively, you could test that the ratio of the means is close to one and compute the variance of that. Unlike for the normally distributed data, these difference in rates aren't likely to be either Poisson or a normal distribution. There are several options to do this analytically.

But most people probably don't want to do and it can be a big investment and hard to analyze.

Finally, businesses often want to use general ratios. For example, you might want to use the ratio of CTP in you experiment and control group instead of the difference. Like for the median, the distribution and estimated variance of a ratio will depend on the distribution for the numerator and the denominator.

Often, you won't work those distributions and for anything more complicated than two normal, you often won't have an analytic result for the variance.

Confidence Interval for a Mean

You've already calculated a confidence interval for a probability in less than one using the binomial distribution. Now, you'd like you to calculate a confidence interval for a mean, which will follow a normal distribution. Suppose that Audacity wants to measure the mean number of students that visit their homepage each week, with the following numbers:

87029, 113407, 84843, 104994, 99327, 92052, 60684

From this, they can compute a mean, which I'll represent as N bar by summing the data points and dividing by 7. Now, what is the confidence interval for this measurement? To answer this, you'll need to calculate the standard deviation of the seven counts Audacity collected, which I'll call sigma. Remember we will use a 95% confidence interval.

To calculate the confidence interval, I'll need the mean first:

$$N = \frac{87029 + 113407 + 84843 + 104994 + 99327 + 92052 + 60684}{7} = 91762.29$$

Then, we can compute the standard deviation:

$$\sigma = 17000$$

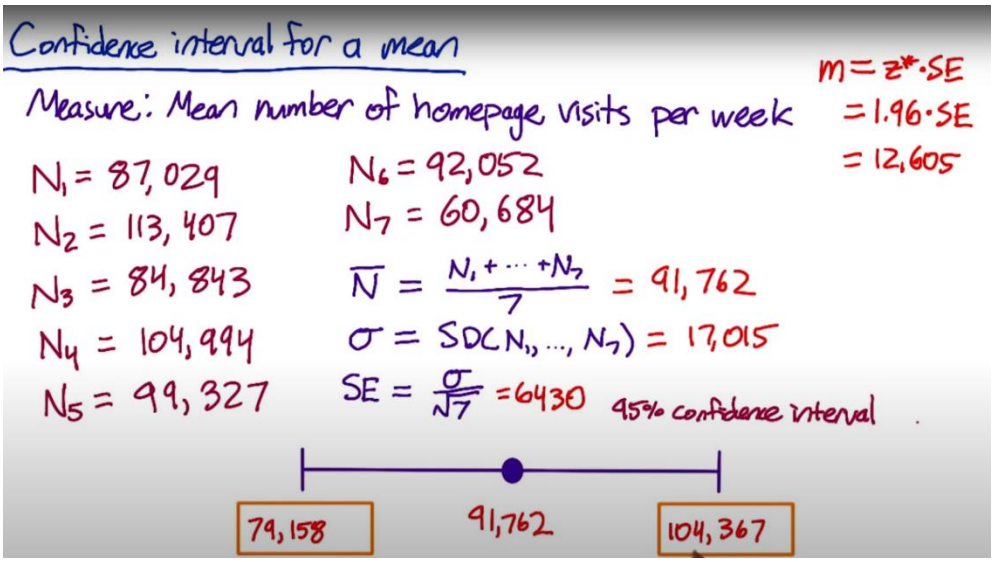
Finally, the standard error for the sample will be the standard deviation:

$$N = \frac{91762{,}29}{\sqrt{7}} = 6430.99$$

Now, to calculate the confidence interval, the estimated mean will be the center of the interval. Then, we will need to calculate the margin of error. As this metric is normally distributed, the margin of error follows this formula. Since we are using a 95% confidence interval, our Z score 1.96:

$$m = 1.96 \cdot 6430 = 12605$$

The upper bound of the confidence interval will be the center plus the margin of error and the lower bound will be the center minus the margin of error. So, if Audacity repeated the measurement for another seven weeks, they might expect to get anywhere from 79k to a 104k as the mean homepage visits per week.



Difference Between Poisson Variables

The difference in two Poisson means is not described by a simple distribution the way the difference in two Binomial probabilities is. If your sample size becomes very large, and your rate is not infinitesimally small, sometimes you can use a normal confidence interval by the law of large numbers. But usually, you have to do something a little more complex. For some options, see **here** (for a simple summary), **here** section 9.5 (for a full summary) and **here** (for one free online calculation). If you have access to some statistical software such as R (free distribution), this is a good time to use it because most programs will have an implementation of these tests you can use.

If you aren't confident in the Poisson assumption, or if you just want something more practical - and frankly, more common in engineering, see the Empirical Variability section of this lesson, which starts **here**.

Nonparametric Answers

What can we do if you want to use one of these more difficult metrics, like a median or ratio, or anything where you can't calculate the variants analytically? There's actually a broader class of methods here called non-parametrics method. That means that you have a way to analyze the data without making an assumption about what the distribution is. These methods can be noisier and they can be more computationally intensive, but they can also be very useful.

A great simple example is called the sign test. Now, let's imagine that we ran our A/B experiment for 20 days and on 15 of those days, the experiment side had a higher measurement than the control side. That seems unlikely to have arrived by chance, and what you can do is use what we learned about the binomial distribution to actually calculate how likely it is that that occurred, if there was realy no difference between the two sides.

The downside of doing this is that it doesn't help you estimate. You can't say you're confident this is at least 2% change in your metric. The upside is it's pretty easy to do and you can do it under a lot of different circumstances. So, if you want to launch any positive change in your experiment, then you could figure out whether there was one using a sign test.

However, sometimes you won't want to launch unless you meet some threshold, your practical significance level. Later we will talk about how to actually compute the variance empirically from the sample data.

Once you've done that, you have two choices. First, if you look at your summary statistic distribution and it's pretty normal, you can do what we've done for normal distributions and use the normal confidence interval just with the variants you estimated empirically. If your data is a little funnier than that or if you want to be really robust, you can actually go ahead and compute a nonparametric confidence interval.

Empirical Variability

For more complicated metrics, you might need to estimate the variance empirically instead of computing it analytically. The key rule of thumb to keep in mind is that the standard deviation is going to be proportional to the square root of the number of samples. That's great if you have enough traffic and you run your own experiments so that you can run that many. However, what if you can't for some reason?

Another option is to run one really big A/A experiment. There's a method in statistics called the bootstrap, where what you do is you take that big sample and you randomly divvy it up into a bunch of small samples and you do the comparison within those random subsets.

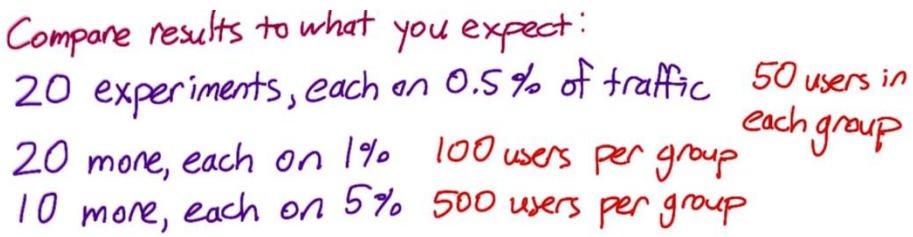
Empirical Variability: Sanity Check

Let's look at the results of some A/A tests on CTP. Since we've already doone the analytics calculation for CTP, we'll be able to compare the empirical results to the analytic results. Remember you can also use A/A tests in case where you weren't able to do an analytic calculation. Now, there are three main things you can do using A/A testing:

- **Compare results to what you expect (sanity check)**. If you have already an analytic calculation of your confidence interval, you can check your A/A test results to see if you're getting what you expect.
- **Estimate variance and calculate confidence**. If you are willing to make an assumption about the distribution of your metric, but you weren't able to estimate the variance analytically, you can estimate the variance empirically and then use your assumption about the distribution to calculate the confidence interval the same way we did before.

- **Directly estimate confidence interval**. If you don't want to make any assumptions about your data, you can directly esimate a confidence interval from the results of the A/a tests.

Case 1: Compare Results to What you Expect



Let's way we run 20 experiments, each running on 0.5% of our traffic. Then, we run another 20, each on 1% of our traffic and 10 more, each on 5%. All of these are A/A tests, with no difference between the two groups. Each of these 20 experiments had 50 users in each group. Similarly, each of these experiments had 100 users per group and the last 10 experiments had 500 users per group.

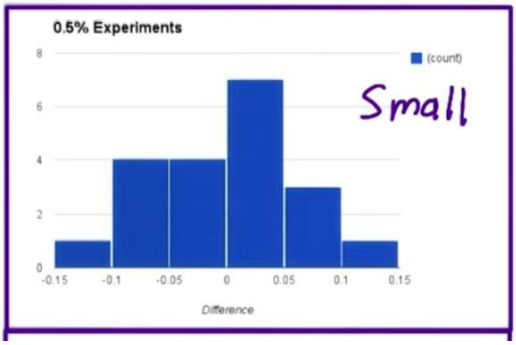
Let's say we analyzed each of these experiments using the same methods as in lesson one. How many experiments will show statistically significant difference at 95% level? Even there was no difference between the two groups, remember that in a 95% confidence interval, the true value, in this case zero, will be only captured 95% of the time. That means that our of 20 experiments, we expect to see one significant difference on average.

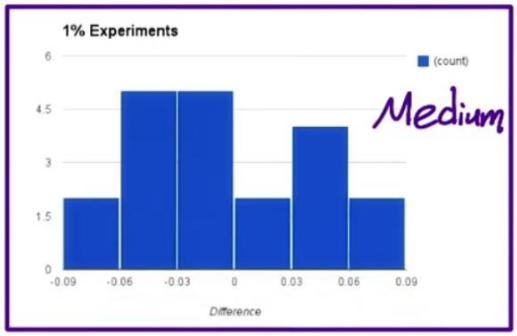
Now, we have the CTP measured for both groups for all the 50 experiments we ran. Now, if we analyze these numbers using the methods from lesson one, then we find that one experiment in the smallest group is significant and zero in teach of the other two. That's not terribly surprising. If you had seen 5 positives in one of the groups, then that would be a sign something was wrong with the set-up of the experiment or in the assumptions that you made. **This spreadsheet** contains the data, calculations, and graphs shown in the video.

Another thing we can check about A/A tests in whether the difference follow the distribution we expect. To do that, I'll insert a third column which contains the difference between the two groups for each experiment. One thing to check is whether the differences are following a normal distribution as we expect.

For the smallest experiments, the distribution looks fairly normal, but for the other two it doesn't. However, I'd say that this is probably due to the fact that we didn't run

that many experiments. Another thing that these plots show is that the distribution is getting tighter as the experiment size increases. It is hard to see that from the actual width of the distributions since these plots are scaled to the range of the values.





Case 2: Empirical Confidence Intervals

Now, let's move on to the second case I mentioned for A/A tests, where we can estimate the variance empirically if we weren't able to calculate it analytically. **This spreadsheet** contains the data and calculations shown in the video. We will actually compute the standard deviation instead since that's the direct analog of standard error. We do this by taking the standard deviation of the difference from all the experiments.

Since we expect that our metric follows roughly a normal distribution, we can compute the margin of error the following way:

$$m = z * \cdot SD$$

This is the same equation you saw in lesson one, but with the empirical standard deviation instead of the analytic standard error. If you didn't know beforehand whether to expect your metric to follow a normal distribution, then you might look

at histograms like the ones we just looked at to see whether the metric looked like it followed a normal distribution.

If we do that for the small experiment, we just looked at with a 95% interval, then we get:

$$m = 0.059 \cdot 1.96 = 0.116$$

Remember if we had done this analytically, the standard error depends the pooled probability, which will be different for each experiment:

$$SE_{pool} = \sqrt{p'_{pool} \cdot (1 - p'_{pool}) \cdot \left(\frac{1}{N_{con}} + \frac{1}{N_{exp}}\right)}$$

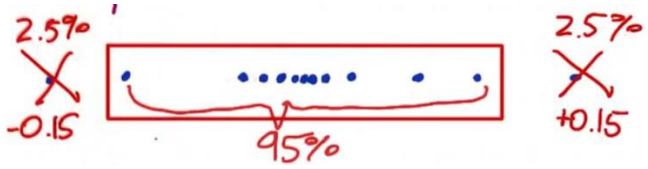
That means we actually would have gotten a slightly different margin of error for each experiment. Whereas, empirically, we calculated one margin of error across all the experiments.

If I do calculate the pooled probability and the pooled standard error for each experiment, I see that the pooled standard error varies, but it's always close to 0.059. So, we're getting roughly the same results either way.

Case 3: Directly Estimate Confidence Interval

However, if your metric doesn't seem to follow a normal distribution, then what can we do? You can also directly estimate a confidence interval from the results of your A/A tests.

The way to do this is take all your differences and put them in order. Then, if you want a 95% confidence interval, select a box hat includes only 95% of the values, discarding 2.5% of values of each side. Then, the range of your remaining data points will give you a 95% confidence level. Since we have 20 data points, dropping the highest and the lowest gives a 90% confidence level. For our data, that gives us a lower bound of -0.1 and an upper bound of 0.06.



Recall that the empirical standard deviation we calculated a minute ago was 0.059. Now, if we multiply that by 1.65 which is the z-score for a 90% confidence level, then that comes to about 0.097.

$$m = 0.059 \cdot 1.65 = 0.097$$

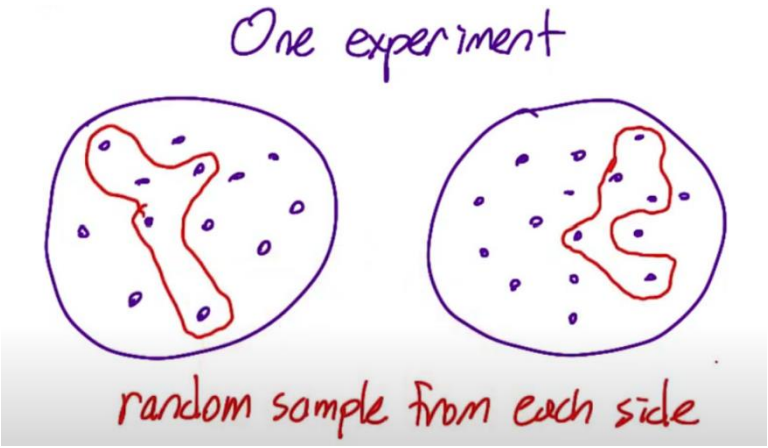
So, if the true difference were 0 that would give us a confidence interval of -0.097 and 0.097.

These two methods give a sort of close answer, but it's not that close. The main reason for that is that we only have 20 data points. You'd probably to run more A/A tests to actually trust this confidence interval, unless each test was run for a pretty long time.

Empirical Variability: Bootstrapping

If you don't want to run a lot of A/A tests, you can run one experiment and then use it to estimate the variability of your metric using called **Bootstrap** method.

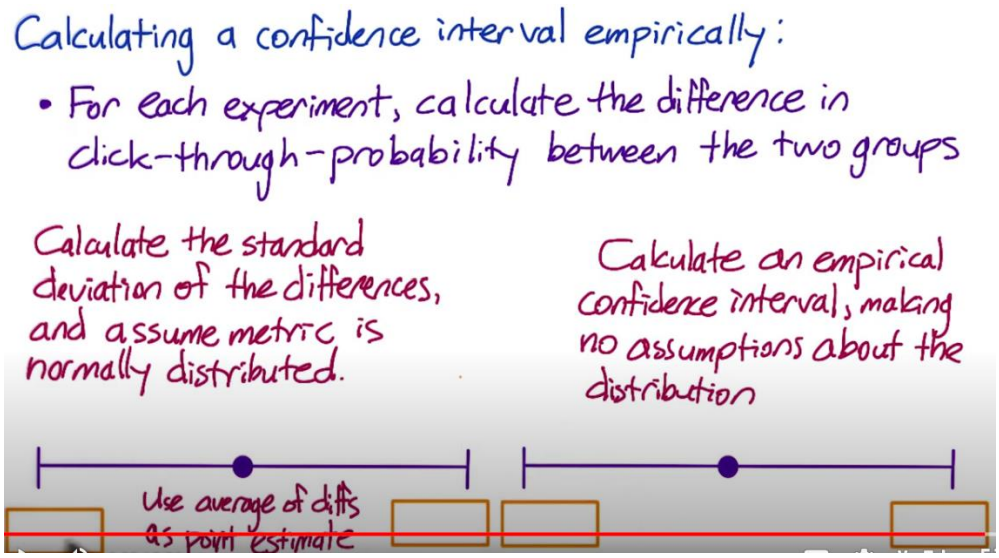
The idea of bootstrapping is that you run one experiment. Then, you take a random sample of those data points from each side of the experiment and calculate the CTP based on that random sample as if it were a full experimental group. Then, you record the difference in the CTP and use that as a simulated experiment. Then, you repeat this process over and over, recording the results. You can use the results as if you had run multiple experiments, even though you really only ran one big experiment.

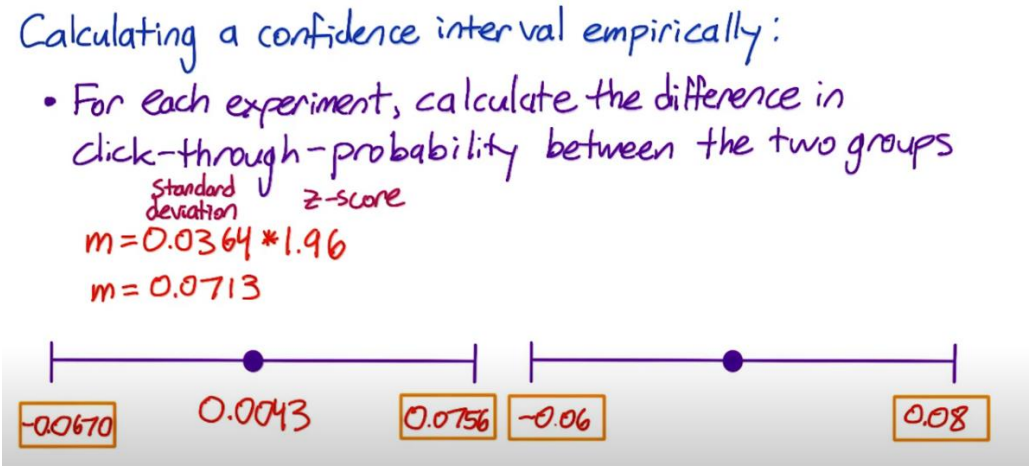


So, the numbers in **this spreadsheet** we've been assuming that they came from multiple A/A tests. But they actually have come from one big experiment, from which

we drew many bootstrap samples. For each experiment, calculate the difference in CTP between the two groups. Then, calculate the 95% interval in two different ways:

- Calculate the standard deviation of the differences and assume metric is normally distributed. Use the average of the differences as a point estimate in the center.
- Calculate an empirical confidence interval, making no assumptions about the distribution.





Designing an Experiment

Unit of Diversion

The unit of diversion is how we define what an individual subject is in the experiment. There are tree commonly used categories of unit of diversion:

User ID

A user identifier can be the login that people create on websites or apps. For example, your email address, your mobile or your username. While a person could have mor than one login, typically a login is a pretty good proxy for a user and it's stable and unchanging.

If you use a user ID as your unit of diversion, what that means is that all the events correspond to the same used ID are either in the control group or the experiment group. However, they are not mixed between the two groups.

Whether the user is using an app on their phone, visiting a website on their phone or visiting a website on their desktop computer, it's a consistent experience.

It's worth noting that a user ID is considered personally identifiable. Some sites use your email address as your login and for sites where you create a user name, many people use some variation on their name.

Anonymous ID

An anonymous ID is usually something like a cookie. On most website, whenever a user visits the website, it will write a cookie, which is usually an anonymous random identifier to a file on that device. The cookie is specific to a browser and a device though. If the users switch from Chrome to Firefox or from their laptop to their phone, they'll get a different cookie.

Also, users can choose to clear their cookies, in which case the next time they visit the website they'll get assigned a new one. It's also possible to set your preferences such that every time you close your browser, all your cookies are cleared automatically. It's much easier for a person to change their cookie than it is for a person to change or clear an account, like the user ID.

Event

Event-based diversion means that on every single event, you redecide whether that event is in the experiment or in the control. This means the user may not get a

consistent experience at all, so this is only appropriate in situations where the changes are not user visible.

Device ID

On mobile devices only, there's an option called a device ID. The device ID is between a cookie a user ID. It's typically something that's tied to a specific device and it's unchangeable by the user. It's also considered identifiable because it's immutable. However, it doesn't have the cross device or cross platform consistency that the user identifier might have.

IP Address

Any event with the same IP address will be put in the same group. If the user changes location, then they often get a new IP address.

Consistency of Diversion

How would you actually choose between units of diversion? One of the main considerations is user consistency. If you are using a user ID, then the user gets a consistent experience as they change devices as long as they're signed in. For example, if you are testing how courses are being displayed, then the user will get a consistent experience across devices.

However, if you're testing a change that crosses the sign in, a user ID doesn't work as well. For example, if you are changing the layout of the page or the location of the sign in bar. In that case, you may want to use a cookie instead, so you get consistency.

This also depends in what you want to measure. If you want to measure a learning effect, whether or not users adapt to change. You algo need a stateful unit of diversion like a cookie or user ID. For example, if you're making a latency where you are making the site slower and you're trying to see whether or not the user uses the site less. In those cases, you will need a user ID or cookie to see what happens across time. When the user doesn't notice the change, depending on what you want to measure, you may also choose a user ID or cookie.

IP based diversion is not very useful. You don't get the consistency that you get from a user ID or a cookie because the user's IP address could randomly change depending on what's happening with the provider, for example. There's a whole host of changes where IP based diversion may be your only choice. For example, you're

testing out an infrastructure change when you're testing out one hosting provider versus a different hosting provider to understand the impact of latency. In that situation, IP based diversion may really be your only choice.

Unit of Analysis vs. Unit of Diversion

The second consideration to have in mind is variability. You might recall back in lesson three when we talked about how to compute the variability of a metric empirically. And that reason why was because sometimes the empirically computed variability was much higher than the analytically computed variability.

The reasoning of this is because that's what happens when your unit of analysis is different than your unit of diversion. The unit of analysis is whatever the denominator of your metric is. For example, if you're doing a CTR and you have clicks divided by page view, then page view would be your unit of analysis.

When your unit of diversion is also a page view, so as would be the case in an event based diversion, then they analytically computed variability is likely to be very close to the empirically computed variability.

However, if your unit of diversion is a cookie or a user id, then the variability of the same metric CTR is actually going to be much higher. In those cases, you want to move to an empirically computed variability given your unit of diversion.

This difference happens when you're actually computing your variability analytically. You're fundamentally making an assumption about the distribution of the data. But also, you're making an assumption about what's considered to be independent. You're basically doing these random draws and whether they're independent or not.

When you are doing event-based diversion every single event is different random draw and your independence assumption is actually valid. When you're doing cookie or user ID based diversion, that independence assumption is no longer valid because you're actually diverting groups of events. And so, they're actually correlated together. That will increase your variability greatly.

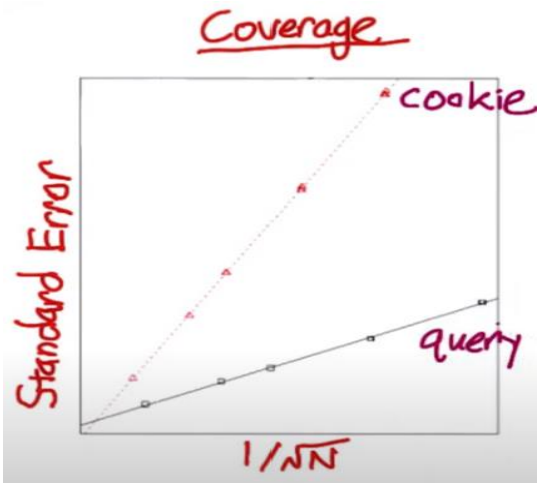
Changing the unit of diversion can change the variability of a metric, sometimes pretty dramatically. One experiment tested the unit of diversion between a query and a cookie. Diverting by query is a type of event-based diversion, since for a search engine a query is an event. The metric they measured was called coverage, which is

defined as the percentage of queries for which an ad is shown. You can calculate it following this formula:

$$\text{Metric: Coverage} = \frac{\#\text{queries with ad}}{\#\text{queries}}$$

Since the number of queries is the denominator, that means that the unit of analysis is a query in this case. So, when the unit of diversion was a query, then the unit of diversion and the unit of analysis were the same.

This plot shows the results of the experiment, with the standard error on the y-axis and 1 over the square root of the sample size on the x-axis. The red line shows the standard error for cookie-based diversion and the black line shows the standard error for query-based diversion.



To see why the x-axis is 1 over the square root of N, instead of N directly, recall that the standard error for a binomial is proportional to 1 over the square root of N. That's why both lines here are straight.

$$\text{Binomial: } SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{N}}$$

Notice that when the unit of diversion is a cookie, which is not the same as the unit of analysis, the variability is much higher than when the two units are the same. The variability might be higher by as much as four times, depending on the sample size.

The standard error for the query-based diversion was much closer to the analytical standard error. When your unit of analysis and your unit of diversion are the same,

the variability tends to be lower and closer to the analytical estimate than when they're different.

For each of the following cases, would you expect the analytic variance to match the empirical variance?

- **CTR (clicks divided by pageviews) and the unit of diversion is a cookie**. The cookie doesn't match the unit of diversion, which usually means the analytical estimate will be an underestimate of the variance. So, you wouldn't expect the analytic and empirical variance to match.
- **Cookies that view the homepage and the unit of diversion is a pageview**. There isn't really a denominator to this metric, but since the number of cookies is what's being computed, the cookie is the unit of analysis. The unit of analysis is larger than the unit of diversion in the sense that one cookie could generate multiple pageviews. This is a problem, given that the unit of diversion is a pageview, because it means the same cookie could have events in both experiment and control groups. This means the metric is actually not well defined for this experiment design. In general, you need your unit of diversion to be at least as big as your unit of analysis. In this case, cookie would work as the unit of diversion, and user ID would work, since one user ID can correspond to multiple cookies. But usually not vice versa.
- **Percentage of users that sign up for coaching divided by the total users who enrolled a course**. The unit of diversion is the user ID. The unit of analysis matches the unit of diversion. You'll never get analytic and empirical estimates that agree exactly, but they'll probably be a lot closer in this case.

Inter vs. Intra User Experiments

When you are choosing a population, if you choose cookie or user ID based as unit of diversion, you're looking at proxies for users. That means you're going to have one group of users on the A side of you experiment and one group on the B side.

If you do event-based diversion, you can end up with a mix of the same people on both sides. You have to be pretty careful in this case to make sure you haven't inadvertently mismatched your users.

We can also perform an intra-user experiment. This typically means that you expose the same user to this feature being on and off over time, and you actually analyze how they behave in different time windows. This has some pitfalls. For example, you

have to be really careful what you choose a comparable time window. You don't want to do this in two weeks before Christmas and then have them behave very differently in the second part.

For certain other types of applications, like search ranking, preferences or other thing where you have a ranked order list, you have an option of running what's called an interleaved experiment, where you actually expose the same user to A and B side at the same time. This typically works in cases where you're looking at reordering a list.

We also have the inter-user experiments. That means you've got different people on the A side and on the B side. You use cohorts with this type of experiments. In a cohort, you try to match up your entering class so at least you have roughly the same parameters in your two user groups.

Interleaved Experiments

In an interleaved ranking experiment, suppose you have two ranking algorithms, X and Y. Algorithm X would show results X1, X2, ... XN in that order, and algorithm Y would show Y1, Y2, ... YN. An interleaved experiment would show some interleaving of those results, for example, X1, Y1, X2, Y2, ... with duplicate results removed. One way to measure this would be by comparing the click-through-rate or -probability of the results from the two algorithms. For more detail, see **this paper**.

Target Population

If we assume that we're doing an inter-user experiment, is there anything else we need to decide over our population? You still need to decide who you're targeting in your users. Normally, this decision is taken in advanced because of two reasons.

The first one is that sometimes you may want to restrict who sees your experiment for a variety of different reasons. For example, if you're running a feature and you're not sure if you're going to release it and it's a pretty high-profile launch, you might want to restrict how many of your users have actually seen it. Another example is that if you're running a couple of different experiments at your company at the same time, you might not want to overlap.

The second reason is sort of a more numeric, which is you may not want to dilute the effect of your experiment across a global population. So, if you're analyzing an experiment for the first time and it only affects English, you may want to do your analysis specific on English, and be able to ignore the rest of the population.

There are some situations where you don't want to target the experiment in advanced. Sometimes you can't necessarily ID who a particular feature is going to affect. Other times you may just not care that much because it could be a feature that affects 90% of your traffic. In those cases, it's just not worth the trouble to try to target the experiment.

If you can identify what population will be affected by your experiment, you might want to target your experiment to that traffic. This means running only your experiment on the affected traffic.

Let's take a look at how this can affect the variability of your metric. As we saw, changing the unit of diverging can change the empirical estimate of variability. Filtering your traffic can change the variability as well.

Suppose you want to control the change that will only affect users in New Zealand to see whether it increases the CTP. You run the experiment and looking only at the New Zealand data, you find the following data for the control group:

$$N_{con} = 6021, X_{con} = 302$$

And the following data for the experiment group:

$$N_{exp} = 5979, X_{exp} = 374$$

Based on the equation we saw in lesson one, the estimate probability of a click in the control group or experiment group is:

$$p'_{con} = \frac{X_{con}}{N_{con}} = 5.1\%$$

$$p'_{con} = \frac{X_{exp}}{N_{exp}} = 6.3\%$$

The pooled standard error is 0.0042.

$$SE_{pool} = \sqrt{p'_{pool} \cdot (1 - p'_{pool}) \cdot \left(\frac{1}{N_{con}} + \frac{1}{N_{exp}}\right)} = 0.0042$$

Now, suppose you had analyzed the global data instead of just the New Zealand data. Since the change doesn't affect other traffic, let's say that this is all the non-New Zealand traffic.

$$N_{con} = 50000, X_{con} = 2500$$

$$N_{con} = 50000, X_{con} = 2500$$

Now, for the global data, which includes both the other data and the one from New Zealand all together, what would be the pooled standard error?

First, let's get the global calculations:

- N_{con} = 6021 + 50000 = 56021
- X_{con} = 302 + 2500 = 2802
- N_{exp} = 5979 + 50000 = 55979
- N_{exp} = 374 + 2500 = 2874

Then, let's calculate the pooled probability and the pooled standard error:

$$p' = \frac{X_{con} + X_{exp}}{N_{con} + N_{exp}} = 0.051$$

$$SE_{pool} = \sqrt{p'_{pool} \cdot (1 - p'_{pool}) \cdot \left(\frac{1}{N_{con}} + \frac{1}{N_{exp}}\right)} = 0.0013$$

For New Zeeland, the pooled error was 0.0042, which is about four times as large. In this case, the variability of the global data is lower than the filtered data. This is mostly because there is so much more data globally.

Population vs. Cohort

In the population you have a shole group of users. Within that population, you can define a cohort. This means people who enter the experiment at the same time. Cohort usually means that you define an entering class and you only look at users who entered your experiment on both sides around the same time.

When would you want to use populations versus cohorts? Cohorts are harder to analyze and they're going to take more data because you'll lose users. Typically, you only use them when you're looking for user stability. For example, if you have a learning effect or you want to measure something like increased usage of the site or increased usage of a mobile device. Those are cases where you really want to see if your change had a real effect on their behavior relative to their history.

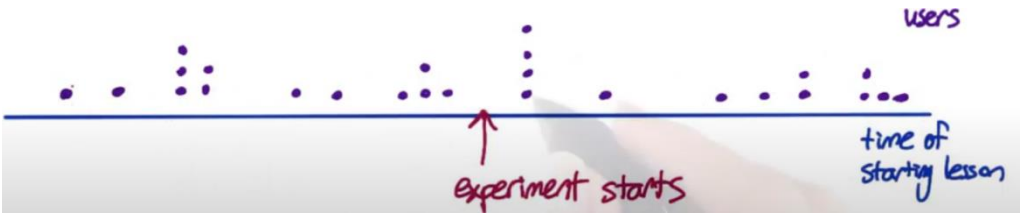
You might want to use a cohort depending on the type of problem you are looking at. A cohort makes more sense than looking at the entire population if you are looking for learning effects, if you're examining user retention, if you want to increase user activity or anything else that requires the user to be established for some reason.

Going back to the Audacity example, here's a situation where they might need to use a cohort. Suppose they have an existing course that's already up and running. Some students have completed the course, other students are midway and there are students who have not yet started. They want to try changing the structure of one of the lessons to see if it improves the completion rate of the entire course.

Now, because they want to see what happens throughout the course where students can pause or unpause lessons, switch devices, etc... the unit of diversion will need to be a user ID. Also, it doesn't make sense to just run the experiment on all the users in the course.

To see that, suppose that this blue line shows the time that students start the lesson that Audacity is changing with later times to the right. Each purple dot represents a user or student.

Now, suppose that Audacity starts running the experiment at this time. For students who started the lesson a while ago, they may actual have finished the lesson already. So, they're already past that lesson and they're not even going to see the change.



Instead, it would make sense to use a cohort and only include users who started the lesson after the experiment was started.

For the control group, Audacity needs to create a comparable cohort. They cannot just use these users, who are not included in the experiment, as the control because they may have been other system changes in that time that affected the new users. Audacity will need to split this cohort into an experiment cohort and a control cohort, so that they all have the same timing of when they started the lesson.

Experiment Design and Sizing

Now, we have to actually size our experiment and control groups. This is an iterative process. We are going to try out some decisions for our unit of diversion and our population, see what the implication is on both the size as well as the duration of our experiment. And then, if we don't like those results, we'll need to revisit our decisions and iterate.

Sizing

Before start working on the size, we need to have in mind that the choice of the metric, the unit of diversion and the population can affect the variability of the metric. You want to take all the stuff into account and then start to determine the size based on the process we talked about before. Then, you're going to have to figure out if what you've planned to do is realistic, give how long you have to run the experiment and the variability of your metric.

For example, imagine you want to measure the page load time with a 90th percentile latency. Originally, you could measure that in an event-based diversion because you just measure each page load time. Let's say we want to look at the user ID diversion where we look at whether that user uses our site more or less based on the latency that they're experiencing. That's a little more challenging as a metric because you're going to need a fair amount of user data to make that work. If you're originally planning to run this globally, you may realize looking at the variance of your metric that it's not realistic. It's going to take a very long time to get a lot of data. At this point, you might want to go back and choose the 90th percentile. Maybe in this case you may want to look at a cohort of users who have used my site fairly regularly over the past two months.

How to Size your Experiment

Suppose that Audacity sometimes includes promotions for their coaching next to the videos in their free courses. They run an experiment changing the wording of this

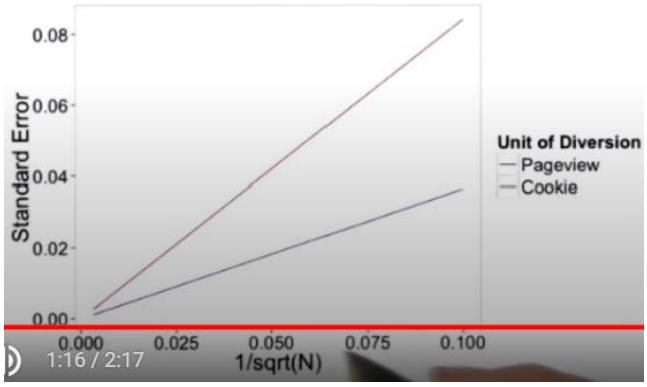
message and they use CTR as their metric (clicks/pageviews). Since the unit of analysis is pageviews, Audacity might want to make the unit of diversion pageviews. They might also want to consider using a cookie to get more consistent experience. Let's see how that would affect the size of the experiment.

If you calculate the variability analytically, it won't change between the two units of diversion. However, for the cookie-based diversion, the analytic estimate is likely to be under-estimate. For this reason, Audacity does an empirical estimate of the variability with 5000 pageviews in each group.

First, they randomly sample by pageview to simulate the unit of diversion being and they get a standard deviation of 0.00515. Then, they randomly sample by cookie to get a different set of pageviews and the standard deviation is 0.0119. That's quite a difference. However, how will it actually affect the size of the experiment? In order to calculate the size, we can assume that the standard error for the experiment is proportional to one over the square root of the sample size:

$$SE = \frac{1}{\sqrt{N}}$$

That means that as you vary the sample size, the standard error would look like this for each unit of diversion. We are using the 1/sqrt(N) so the standard error shows a straight line. Using this information about how the standard error varies with the sample size, you can calculate the size that would be needed.



Let's say the practical significance boundary or d_{min} = 0.02. If Audacity used pageview as the unit of diversion, they would need about 2,600 pageviews to get enough power. However, if they diverted by cookie, they would need about 13,900 pageviews, which is a huge difference.

How to Decrease Experiment Size

Let's say that Audacity does another experiment, this time changing the order courses appear on their course list page. The metric they use is the overall CTR to individual course pages. This is the total number of times the user clicks on any course divided by the number of pageviews. To give a consistent user experience of the list, while still no including non-logged in traffic in the experiment, Audacity chooses cookie as the unit of diversion.

They use their standard values of 0.05 and 0.2 for alpha and beta. Their practical significance boundary is 0.01 and they empirically estimate that their standard error to be 0.0628 for 1,000 pageviews, sampled using cookie-based diversion. The result is that Audacity would need at least 300k pageviews in each group in other to have enough power, which would correspond to all of their traffic for a month.

Audacity isn't willing to spend hat long getting results on this one experiment. Which of the following strategies could Audacity try to reduce the total number of page views needed to get a result?

Increase d_{min}, alpha or beta

We already saw in lesson one that they could increase the practical significance boundary (not try to detect a smaller change). Or alpha or beta, which means accepting a higher probability of a false positive or a false negative. So, this would work.

Change Unit of Diversion to Pageview

If each pageview was randomly assigned to the control or experiment group, even if two pageviews came from the same cookie, would this reduce the total number of pageviews needed? This change would almost certainly reduce the number of pageviews needed. By changing the unit of diversion to be the same as the unit of analysis, the variability of the metric will probably decrease and be closer to the analytical estimate.

By decreasing the variability of the metric, you decrease the number of pageviews you need to be confident in your results. The main question is whether the less consistent experience will be acceptable. In this case, if Audacity recalculated the empirical estimate of the standard error using the pageview as the unit of diversion,

they might find that the new standard error was 0.209 for the same sample size. Then, only 34k pageviews per group would be necessary.

Target the Experiment to Specific Traffic

Suppose that Audacity has classes in many languages, and this experiment only reorders the English classes. Would it help to restrict this experiment to only the English traffic? This will also reduce the total number of pageviews.

Since the non-English traffic is not affected, including it will dilute the results of the experiment, which would increase the number of pageviews needed. Of course, there are a fewer non-English pageviews available than total pageviews. So, this might not reduce the time frame of the experiment, but other experiments could be run on the non-English traffic in the meantime.

Filtering the traffic could also impact your choice of practical significance boundary. First, since you're only looking at a subset of your traffic, you might need a bigger change before it matters to the business. Or since our variability is probably lower, you might want to take advantage of that and detect smaller changes rather than decreasing the size of the experiment. Because the practical significance boundary could move in either direction, your size could really move in either direction. However, it's likely the variance will go down and the practical significance boundary will increase, so it's likely that the size will be smaller.

In this case, suppose that Audacity keeps pageviews as the unit of diversion and then targeting the experiment to English only traffic further reduces the standard error to 0.0188. They also decide to increase their practical significance boundary to 0.015 for the English traffic only. At this point, they would only need 12,000 pageviews per group.

Change Metric to Cookie-Based CTP

This change depends on changing the metric definition. However, it will often not make a significance difference to the variability, especially if you're using a short time window for the probability. If there is a difference, the probability will probably go down. Since the unit of analysis would be the same as the unit of diversion in this case. So, this could reduce the number of pageviews needed, but it also might not help much.

Duration vs. Exposure

Now, we need to translate our ideal size into a set of practical decisions. First of all, what's the duration of the experiment that I want to run? Secondly, when do I want to run the experiment? Third, what fraction of your traffic are you going to send through the experiment? Those are all interrelated as they get you to the ideal size but you need to think about them a little bit separately.

Let's deep-dive more in the third element. Imagine that your unit of diversion is a cookie. In this case, what we're asking is on any given day what proportion of the cookies are you sending to your experiment and your control.

We decide we need one million cookies in our experiment and control combined. If you only get 100,000 cookies visiting your site on any given day, that means that if you want to run 50% of your traffic through the experiment and 50% through the control, you need to run your experiment for ten days. Another choice is to run your experiment at 25% each. Then, you'll need to run your experiment for 20 days as opposed to 10.

This is how the duration of your experiment is related to the proportion of traffic that you are sending though your experiment.

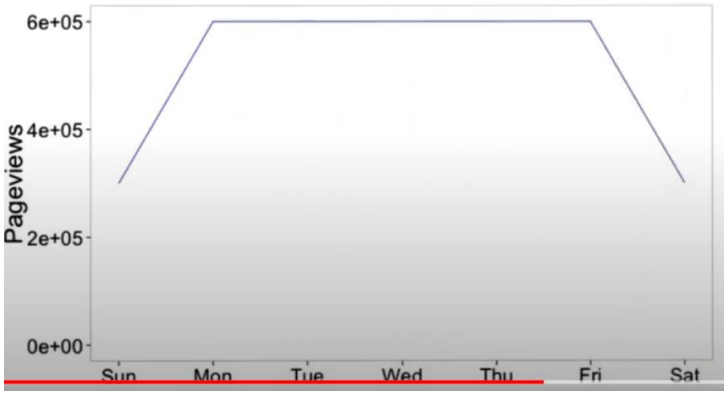
Is there any reason why you wouldn't run the experiment in all your traffic so you can get results quicker? There's a couple of different reasons. The first is safety. You may have a new UI feature, and you're not sure either how well it functions in all browsers. In this case, you might want to actually keep the site mostly the same and only expose a few people to it until you feel more comfortable with it.

The other reason would be something like press. Let's say it is a new feature and you're not sure you're going to keep it. Do you want a lot of people seeing this and potentially blogging about it if you're not sure it's even going to be the way you go with the site?

Another reason is that right now we're randomizing across diversion. The question is what other things are actually impacting the variability of your results? If you are running a 50%-50% experiment, then you can gather all data on a single day. Would you actually want to make a decision based on a single day if it was a holiday? A more common scenario is that you have to have a very different behavior on weekdays and weekends. So, you might actually prefer to run a smaller percentage across multiple days to get a sense of how they differences are by weekday and weekend, across holidays, by different times of the day.

Suppose you're considering another experiment where you've computed the size needed to be 1 million pageviews, split across control and experiment groups. If your average traffic per day is 500,000 pageviews, then even if you split all of your traffic evenly between the control and the experiment group, you'll need to run the experiment for two days.

However, is your traffic the same every day? It's more likely that you have some weekly variation such as in the following plot, where the traffic is lower on the weekends and higher on the weekdays.



If you look at your metric of interest, you might see a variation in that based on what day of the week it is also. In such a case, you should run on a mix of weekend and week days. You wouldn't want to just run the experiment on two weekdays. Then, you might need to run for three days rather than two to get enough traffic, since the weekends don't have much traffic.

If the change is risky enough that you don't want to expose such a large percentage of your traffic to it, you might run for longer, like 7 days, with a lower percentage of your traffic diverted.

Analyzing Results

In this lesson, we'll go over how to analyze the results from your experiment and determine what you can and what you cannot conclude.

Sanity Checks

Before we can take a look to the metric we chose for our experiment, first we have to do all those sanity checks in order to ensure that your experiment was run

properly. There are different things than can invalid your results. For example, maybe something went wrong in the experiment diversion and now your experiment and your control aren't comparable. Or maybe you used some filters and you set up the filter differently in the experiment and the control. It's possible that you set up data capture and your data capture isn't actually capturing the events that you want to look for in your experiment. We need to test before we can look at the results of your experiment.

How to test it? We are going to check the invariant metrics to perform to different type of tests. The first one is going to be the population sizing metrics, based in your unit of diversion. What you are really checking there is that your experiment population and control population are actually comparable. The other check is realized over the invariant metrics themselves, those metrics that shouldn't change when you run your experiment. You need to test that they didn't change.

Choosing Invariant Metrics

The first step to sanity checking our results is choosing a set of invariant metrics. There are two types of invariant metrics: population sizing metrics and any other metrics you don't expect to change.

We are going to go over two experiments and for each one we will choose which metrics should be invariant. In other words, what would be the same in the control and experiment group?

Case 1

In the first experiment, Audacity changes the order of courses in the course list to see how much this affects which courses users eventually enroll in. The unit of diversion is a user ID since users may visit the site multiple times before finally enrolling in one course. Now, which of the following metrics would be good invariants?

The number of signed in users is a good population sizing metric. Since the unit of diversion is the user ID, the signed in users are being randomly assigned to the experiment and control groups. You should definitively have roughly the same number of users in each group.

The number of cookies and number of events are a little trickier. They're not being directly randomized, but they should still be split evenly between the two groups.

Unless user in the experiment tend to clear their cookies more often or visit fewer pages. If that's happening, it's probably not a good thing, and it's something you need to know about. So, these are also good population sizing invariants.

The CTR on the 'Start Now' button is a good invariant metric. Since users click the button before they see the course list, its CTR shouldn't be affected by this change.

On the other hand, the time taken to complete a course could be affected by this change. If ordering the courses differently does case users to enroll in different courses, then that could change how long it takes users to complete them. Maybe putting easier courses first causes more users to start with the easier courses and then they finish them faster.

Case 2

In the second experiment, Audacity changes the infrastructure serving videos hoping to reduce the video load time. This time the unit of diversion is an event. Now, which of the following metrics would be good invariants?

The first three metrics will be all good population sizing invariants. The number of events is good since this is the unit of diversion. This is being randomly assigned between both groups. Signed in users and cookies are both larger than the unit of diversion in the sense that one user or one cookie could correspond to multiple events. So, since the events are being randomly assigned, the number of signed in users and cookies shouldn’t be different between the two groups either.

The CTR of the 'Start Now' button would be a good invariant metric since users click the button before viewing any videos. There could be a learning effect, but you won't catch learning effects if you're diverting by event anyway.

The time to get through a class can't be tracked if you are using event-based diversion. Since by the time the user gets through a course, they could have been assigned to both the experiment and the control group multiple times.

Case 3

Now let's consider invariant metrics for another experiment. This time, Audacity changes the location of the sign in button to see if they can get users to sign in sooner.

For the control group, the sign in button currently appears on the course list page and if a user who isn't signed in tries to enroll in a course, they are prompted to sign in. However, in the experiment group, the sign in button is added to every page, including the homepage. The unit of diversion is a cookie, since Audacity wants a consistent experience across the sign in / sign out boundary.

The number of events would be a good population sizing invariant. The number of cookies and number of signed in users would be good as well, the same reasons as before. Cookies are being explicitly randomized over and user ID are typically larger than cookies, in the sense that one user ID can correspond to multiple cookies. So, user ID should be evenly split as well.

The CTR in the 'Start Now' button wouldn't be a good invariant. By adding a sign in button to the homepage, the experiment could affect how many people click on the 'Start now' button.

The probability that a user enrolls in a course would not be a good invariant either. Users often enroll in courses after signing in, so changing sign in rates could affect enrollment.

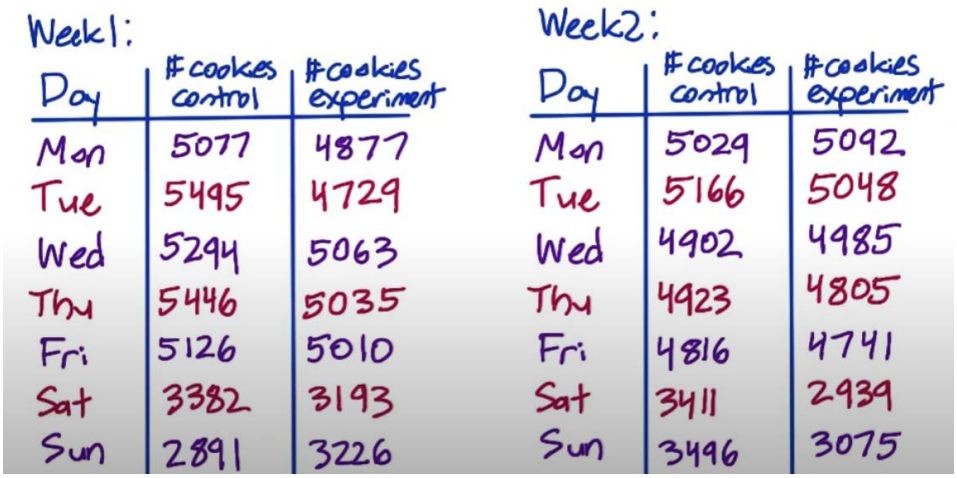
The sign in rate would not be a good invariant. This is the exact metric Audacity is trying to change in this experiment.

The video load time would be a good invariant. No backend changes were made in the experiment and they user has no control over load time, so they can't affect it.

Checking Invariant Metrics

Let's go through an example of how to actually check an invariant and see whether it's reasonably close between the control and the experiment groups.

Let's say you run an experiment for two weeks and your unit of diversion is a cookie. So, the first sanity check you want to do is to see whether the number of cookies in the control if roughly the same as the number in the experiment group. So, you get the number of cookies in each group in each and the results look like this:



The first thing to do is to compute the total number of cookies for each group and see if the overall division looks even. If so, great. If not, then I'll look at the day-by-day breakdown. It turns out that the total number of cookies in the control group is 64,454 and in the experiment group is 61,818. There are more cookies in the control group than in the experiment group.

The question is: is this within what we expect? Ignoring the day-by-day data for a minute, how would you figure out whether the difference between the total number of cookies in the control and the experiment groups is within what you expect? Remember that each cookie is randomly assigned to the control group with a probability of a 0.5 and to the experiment group with a probability of 0.5.

As it's a binomial distribution, we can build a binomial confidence interval like in lesson one. The total sample size here is about 120,000 cookies, which is definitively enough to assume a normal distribution.

First, we will compute the standard deviation of a binomial distribution with probability 0.5 of success, which I'll assign as control group. In step 1, the "standard deviation" is the standard deviation of the sampling distribution for the proportion, or standard error. The abbreviation SE should be used in computations instead of SD.

Then, we'll multiply the standard deviation by the z-score to get the margin of error. After that, we'll compute a confidence interval around 0.5. If the experiment is set up properly, it's very likely that this observed fraction of successes or cookies in the control group will fall within this confidence interval.

Finally, we'll check this previous value falls within this confidence interval.

This is a little different approach than in lesson one, where we computed the confidence interval around the observed CTP. In that case, we didn't know the true CTP probability. However, in this example, we know that if the experiment is set up properly, the true probability is 0.5. So, I can compute the confidence interval.

The standard error comes to 0.0014, as we can see in the formula below:

$$SE = \sqrt{\left(\frac{0.5 \cdot 0.5}{64{,}454 + 61{,}818}\right)} = 0.0014$$

We'll use the 95% confidence level as usual, so the z-score is 1.96.

$$m = SE \cdot 1.96 = 0.0027$$

That means that the confidence interval goes from 0.4973 to 0.5027. So, 95% of the time the observed fraction of cookies should fall within this range. In fact, the fraction of cookies in the control group (p') is 0.5104.

$$p' = \left(\frac{64{,}454}{64{,}454 + 61{,}818}\right) = 0.5104$$

This is significantly greater than 0.5027, so something about the setup is incorrect.

To get a better idea of what could be going wrong, it is a good idea to look at the day-by-day data again. One good thing to check is whether any particular day stands out as causing the problem or whether it seems to be an overall pattern.

First, we'll look at which specific days had more cookies in the control group. It turns out that 11 out of 14 days, there were more cookies in the control than in the experiment group, which is quite high. Then, we'll compute the fraction of the cookies that were in the control group each day, there were a few days with 0.53 or higher. This points to an overall problem rather than a problem on a specific day.

Week1:

Day	# cookies control	# cookies experiment	$\hat{p}$
Mon	5077	4877	0.510
Tue	5495	4729	0.537
Wed	5294	5063	0.511
Thu	5446	5035	0.520
Fri	5126	5010	0.506
Sat	3382	3193	0.514
Sun	2891	3226	0.473

Week2:

Day	# cookies control	# cookies experiment	$\hat{p}$
Mon	5029	5092	0.497
Tue	5166	5048	0.506
Wed	4902	4985	0.496
Thu	4923	4805	0.506
Fri	4816	4741	0.504
Sat	3411	2939	0.537
Sun	3496	3075	0.532

At this point, it's a good idea to talk to the engineers and figure out if something was wrong with the experiment setup. If that doesn't work, you could try slicing. For example, by country, language or platform to see if one particular slice looks like it's causing the problem or you could check the age of the cookies in each group. Does one group tend to have more new cookies while the other group has older ones?

Now, suppose you run another experiment, this time for seven days with event-based diversion. This table shows the number of pageviews in the control and the experiment group each day. The total number of events in the control is 15,348 and the total events in the experiment is 15,312.

Day	# events control	# events experiment
Mon	2451	2404
Tue	2475	2507
Wed	2394	2376
Thu	2482	2444
Fri	2374	2504
Sat	1704	1612
Sun	1468	1465

Calculate the 95% confidence interval for the fraction of total pageviews in the control group. The center of your confidence interval should be 0.5. Then, calculate the observed fraction of pageviews in the control group and the experiment group.

The standard deviation in this case is 0.0029 using the following formula:

$$SE = \sqrt{\left(\frac{0.5 \cdot 0.5}{15{,}348 + 15{,}312}\right)} = 0.0029$$

The margin of error is 0.0056:

$$m = SE \cdot 1.96 = 0.0056$$

The confidence interval balance is then 0.4944 and 0.5056.

The actual fraction in the control group is 0.5006, which is within what is expected. So, this does pass the sanity check.

$$p' = \frac{N_{con}}{N_{total}} = 0.5006$$

However, what happens if one of your sanity checks fails? If your sanity checks fail, don't continue. Go straight to analyzing why your sanity checks fail. If not, your conclusions are almost certainly wrong.

How can you figure out what went wrong? First, something might have gone wrong technically and you want to work with your engineers to understand, if there was something going on with the experiment infrastructure or if the experiment was set up correctly.

Secondly, you could perform a retrospective analysis. Recreate the experiment diversion from the data capture to understand if there is something endemic to what you're trying to do that may be causing the situation.

The third thing is to use those pre and post periods we talked about in lesson four. If you're in a pre-period, you can wonder if you see the same changes in those invariances. If you saw them in the pre period and the experiment, that points to a problem with the infrastructure, the setup, etc.

However, if you see the change only in your experiment but not in the pre-period, that points to something with the experiment itself, maybe the data capture.

Single Metric

Let's suppose we have a single evaluation metric, what do you do then? The goal is to make a business decision about whether your experiment was favorably impacted your metrics. Analytically, that means you want to decide if you're observed a statistically significant result of your experiment. Typically, we also want to estimate the magnitude and the direction of the change.

Then, once you have all the information, you can make a decision about whether you want to recommend that your business actually to perform the change.

To estimate if the changes statistically significant, we will try to estimate the variability we need to analyze the A/B experiment as we talked about at the beginning of the course.

If our results are not statistically significant, it's a good time to take a much longer look at your results, especially if you were expecting a really noticeable difference.

For example, you might want to break it down into different platforms or different days of the week. This can not only help you find bugs in your experiment setup, but it might give you a new hypothesis about how people are reacting to the experiment.

What *Not* to Do if Your Results aren't Significant

There are some ideas of what you can do if your results aren't significant, but you were expecting they would be. One tempting idea is to run the experiment for a few more days and see if the extra data helps get you a significant result. However, this can lead to a much higher false positive rate than you expecting! See **this post** for more details. Instead of running for longer when you don't like the results, you should be sizing your experiment in advance to ensure that you will have enough power the first time you look at your results.

Let's suppose Audacity runs an experiment where they test changing the color and placement of the 'Start Now' button. Their metric is the CTR and they divert by cookie. Their practical significance boundary is 0.01 (d_{min} = 0.01). They use alpha of 0.05 and beta of 0.2. Here are the results of the experiment which was run over seven days. The total numbers are on the last row. If you do a sanity check, you'll find that the number of pageviews is comparable between the two groups.

	Control clicks	control pageviews	experiment clicks	experiment pageviews
Day 1	51	1242	115	1305
Day 2	39	853	73	835
Day 3	64	1129	91	1133
Day 4	43	873	60	871
Day 5	55	1197	78	1134
Day 6	44	1023	72	1015
Day 7	56	1003	76	977
Total	352	7370	565	7270

In lesson one, we analyzed a similar experiment, but we measure CTP instead of CTR, so we could assume a binomial distribution. With CTR the distribution is more likely to be Poisson, which is harder to deal with analytically than the binomial. So, this time, in order to calculate a confidence interval, it's a good idea to estimate the variance empirically. In fact, Audacity already calculated the empirical standard error when they decided what size to use for their experiment.

With a sample size of 10,000 pageviews in each group, they had measured the standard error to be 0.0035. We can assume that the standard error is proportional to one over the square root of N:

$$SE = \frac{1}{\sqrt{N}}$$

However, N here was the sample size of one group, which work well if there's the same size in both groups, as in the estimation of the standard error. However, in our results, there's not the same number of pageviews in both groups. In this case, we can assume that the standard error follows the following formula. This should work well if the ends are fairly close.

$$SE = \sqrt{\frac{1}{N_1} + \frac{1}{N_2}}$$

So, our standard error for our experiment can be determined using this equation, where the empirical standard error divided by the empirical scaling is equal to the standard error of our experiment, divided by the scaling factor for our experiment. This involves the number of pageviews in both control and experiment group.

$$\frac{0.0035}{\sqrt{\frac{1}{10{,}000} + \frac{1}{10{,}000}}} = \frac{SE}{\sqrt{\frac{1}{7{,}370} + \frac{1}{7270}}}$$

Then, the standard error for our experiment comes out to 0.0041.

Now, we'll estimate the difference between the CTR in the control and experiment groups by subtracting the control estimated CTR from the experiment estimated CTR:

$$d' = \frac{565}{7270} - \frac{352}{7370} = 0.030$$

Then, the margin of error is the standard error times z-score, that is 1.96 in this case:

$$m = 0.0041 \cdot 1.96 = 0.0080$$

So, the confidence interval is from 0.0020 to 0.0380. Based on this, we can recommend launching the experiment. The confidence interval does not include the practical significance boundary of 0.01, meaning I can be confident at a 95% confidence level that the true change is large enough to be worth launching.

However, just to double check, let's look at the results of the sign test. To do this, we'll need to look at the day-by-day data. In this case, we have added the CTR for each date:

	Control clicks (CTR)	control pageviews	Experiment clicks (CTR)	Experiment pageviews
Day 1	51 (.039)	1242	115 (.088)	1305
Day 2	39 (.046)	853	73 (.087)	835
Day 3	64 (.057)	1129	91 (.080)	1133
Day 4	43 (.049)	873	60 (.069)	871
Day 5	55 (.046)	1197	78 (.069)	1134
Day 6	44 (.043)	1023	72 (.071)	1015
Day 7	56 (.056)	1003	76 (.078)	977
Total	352 (.048)	7370	565 (.078)	7270

To do the sign test, we need to know the number of days, which in this case is seven. Also, we need to know the number of days with a positive change. So, comparing the CTR in each row, we can see that the experiment group actually had a higher CTR on all seven days. This certainly looks good for the experiment, but what will be the chance of this happening randomly?

If there was no difference, then there would be a 50% change of a positive change on each day. So, the question is if you flip a fair coin seven times, what is the chance it comes up heads seven times? The main difference with the cases where we assumed a 50% chance, for example, when assigning to the control or experiment group, we can't assume a normal distribution, since seven days it not enough for the binomial to closely approximate a normal. We can do this calculation using the probability density of the binomial distribution. Also, you can do a sign test calculation using **this online calculator**.

If we use the calculator, we will see we are interested in the two-tail p-value, which is 0.0156. That's the probability of observing a result at least this extreme by chance.

Since this is less than the chosen alpha of 0.05, the sign test agrees with the hypothesis test: this result was unlikely to come about by chance. Give this, the recommendation of implementing the chance would not change.

Sometimes it's possible the hypothesis test on the effect size showed statistically significant results, but the sign test didn't. Why could this happen? First, the sign test has lower power than the effect size test, which is frequently the case for non-parametric tests. That's the price you pay for not making any assumptions. So, this isn't necessarily a red flag, but it's worth digging deeper and see if we can figure out what's going on.

Gotchas

What do you do when the sign test and the hypothesis test don't agree? Or what if you have significant results on weekends but not on weekdays?

One of the reasons can be the Simpson's paradox. It means that there's a bunch of different subgroups in your data like user populations. Within each subgroup, the results are stable but when you aggregate them, it's the mix of subgroups that actually drives your result.

For example, in the admission process in a famous USA university, if you looked at the number of people who accepted divided by the number of people that applied, the rate of women being accepted was statistically significantly lower than men being accepted, which seemed bad. However, when you look at it by department, there were actually departments where women were accepted at higher rate than men. So, how can that be?

The answer turns out to be that you had to look at by department because the acceptance rates by department were variable, from 6% as much as 60%. What was happening is that more women were applying to the smaller departments that had a very low acceptance rate. So, when you aggregated the numbers, ignoring department, you saw women accepted at a lower rate. But if you looked at each department individually, the rates were really comparable between men and women.

You can get this happening on our experiment because you have these subgroups like people who use it more on weekdays or weekends. You may find that, for example, new users are correlated with weekend use and experienced users who react differently are correlated with weekday use. What you sometimes find in these

cases is that what drives the result of your experiment are how many people from each group you get. Within each group, their behavior is stable, you can get statistically significant result or an insignificant result. However, when we add all them together the changes in your traffic mix are driving the results.

Simpson's Paradox

Let's take a look to the simplified version of the Berkley paradox with only two departments. More men applied to the department A than department B and vice versa for women. We get the following table:

	Men applied	Women applied	Men accepted	Women accepted
Department A	825	108	512 (62%)	89 (82%)
Department B	417	375	137 (33%)	132 (35%)
Total	1242	483	649 (52%)	221 (46%)

In department A, women acceptance is fairly high. However, for department B, we get the overall acceptance rate is lower but still more women were accepted than men as a percentage of applicants. Let's look at the total number of men and woman who applied and were accepted across departments. Overall, a greater percentage of men have been accepted than women. How can this happen? The explanation is that more women applied to department B, which has a lower acceptance rate. Because of this, there is a lower percentage of women accepted overall.

How can this apply to A/B testing? Imagine that if you look at your total experiment results, then the experiment group has a lower CTR than the control group. But then, if you break it down by new users and experienced users, you see the reverse.

Multiple Metrics

What changes when you have multiple evaluation metrics instead of just one? One thing that comes up when you run evaluations for multiple metrics at the same time is that the more things you test the more likely you are to see significant differences just by chance. So, if you are testing 20 metrics and you have a 95% confidence level, you would expect to see one case at least that time where you got a result that says it's significant but it's only concurring by chance.

There is a technique for this called **multiple comparisons** that adjusts your significance level, so that it accounts for how many metrics or how many different tests you're doing. You can also find more information in **this article**.

Let's take a look at an experiment that tracks multiple metrics. Suppose that Audacity sometimes prompts students when they miss quizzes, asking if they'd like to contact a coach. They run an experiment that makes this message appear more frequently. There are a few different metrics they could track:

- Probability that students sign up for coaching at any point during the course.
- How early students sign up for coaching. This could be something like the average amount of progress a student makes before enrolling for coaching.
- Average price paid per students. Applicable if coaching is priced differently, depending on how early in the course the students sign up.

If Audacity tracks all three metrics and does three separate significant tests, which alpha equal to 0.05 for each. Then, what's the probability that at least one metric will show a significant difference, given there is no true difference? In other words, for three metrics, what is the chance of at least one false positive?

To make the problem easier, we'll first calculate the probability that there are no false positives. The chance that each individual metric does not show a false positive is 95%. Now, in order for none of the metric to have a false positive, the first one can´t show a false positive, which happens with probability 0.95. Neither can the second, so we need to multiply by 0.95. And the third can't have a false positive either, so we multiply by 0.95 again.

$$P(FP = 0) = 0.95 \cdot 0.95 \cdot 0.95 = 0.857$$

Then the probability that there is at least one false positive is:

$$P(FP \Rightarrow 1) = 1 - 0.857 = 0.143$$

We made an assumption in this calculation. When we multiplied the probabilities together, I was assuming that the metrics were independent. In fact, this isn't true here. These three metrics are all related and more likely to move together. So, 14.3% is an overestimate of the probability of a false positive. But assuming independence is an easy way to get a conservative estimate.

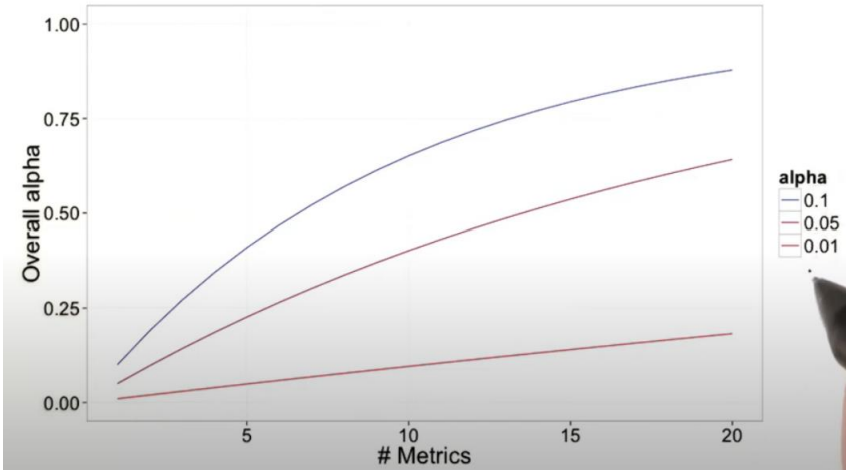
Now, suppose you ran an experiment with ten metrics and you used a 95% confidence level for each metric. What would be the probability that any of those metrics would show a false positive? What if you had those ten metrics but with a 99% confidence level?

In general, the overall probability of at least one false positive, which is called alpha overall, will be the following one:

$$\alpha_{overall} = (1 - \alpha_{individual})^n$$

In the first case, the answer will be 0.401 for a 95% confidence level and in the second case 0.096 with a 99% confidence level. For the first one, is quite large, getting close to half. Even the 99% confidence level isn't enough to limit the overall chance of an error to 5% or less. It's closer to 10%.

This plot shows the overall alpha. This is the chance that any metric shows a false first as the number of independent metrics being tested for three difference confidence levels. For alpha equals 0.1 or 0.05, the overall alpha blows up quickly. By the time you have five metrics with an individual alpha of 0.05, your overall alpha is almost a quarter (25%). When alpha is equals to 0.01, the overall alpha looks more manageable by comparison, but like you saw if you have ten metrics you have a total positive rate of almost 10% and that keeps steadily going up as you increase to 20 metrics.



The main problem with tracking multiple metrics is that a false positive, which should be a rare event, becomes more common as you increase the number of metrics you're measuring. The main way to fix is to use a higher confidence level for each

individual metric to bring down the overall probability of a false positive. What confidence level you use depends on how conservative you want to be.

The first method you might think of is assume that the metrics are independent of each other. Then, you can use the same equation we used earlier to calculate the overall alpha:

$$\alpha_{overall} = (1 - \alpha_{individual})^n$$

Instead, this time you would set the overall alpha to be what you wanted, maybe 0.5, and then solve for the individual alpha you needed in order to get the overall alpha you are happy with.

There's a difference method that people are more likely to use in practice, called the **Bonferroni correction**. It has two main advantages. It's very simple to calculate and it doesn't assume independence that our other method. This method is also very conservative. It's guaranteed to give an alpha overall at least as small as you specified. To calculate it, you calculate the individual alpha you should use for each metric by taking the overall alpha you want and dividing by the number of metrics.

$$\alpha_{individual} = \frac{\alpha_{overall}}{n}$$

So, if you have N equals three metrics and you want your overall probability of a false positive to be 0.05 or less than the individual alpha that you would use for the significance test of each metric.

The main problem with the Bonferroni correction is that often you'll be tracking metrics that are correlated and all tend to move at the same time. In which case, this method is too conservative. In our coaching example from earlier, if users are adopting coaching earlier in the course, it's likely that the probability of adoption is also going up and the price is probably more likely to change too. In this case, the Bonferroni correction would be too conservative.

Now, suppose Audacity runs an experiment where they update one of the descriptions on the course list. We talked about some metrics that might be good for this experiment in lesson three:

- Probability of clicking through to course overview
- Average time spent reading course overview page

- The probability of enrolling in that specific course
- Average time spend in classroom during the first week

Find here the measured difference and the standard error each metric. You can assume they are normally distributed. Which of these metrics showed a statistically significant change?

First determinate which metrics showed us a statistically significance difference, with the individual alpha for each metric of 0.05. Next, set your overall alpha and use the Bonferroni correction to calculate the individual alpha.

At a 95% confidence level or alpha of 0.05, the z-score is 1.96. We will multiply this z-score by each standard error to get the margin of error in each case. For the first metric, the margin of error is smaller than the observed difference, which means the confidence interval won't include zero. So, the difference is significant. The same happens with the following two metrics, but not with the fourth.

Using the Bonferroni method, we get an individual z-score of 2.5. To get this, we have divided the overall alpha of 0.05 by 4 and looked up the result in the z-score table. Multiplying this by the standard error, we get these margins of error. Each margin of error is now larger than the observed difference, so none of these metrics have a significant change using the Bonferroni method. In this case, this method is probably too conservative.

We just saw an example experiment where three out of four metrics had a statistically significant difference at the alpha = 0.05 level for each metric individually. However, when we combined the metrics and tested for significance using the Bonferroni correction, none of the metrics were significant.

So, how we do actually make a recommendation here? The Bonferroni method is likely to be too conservative, since I expect the metrics to be correlated, but my rigorous results say that the change isn't significant. In this case, we need to use a

more sophisticated method than Bonferroni, ideally one that takes into account the fact that the metrics are likely to be correlated.

The **Bonferroni correction** is a very simple method, but there are many other methods, including the **closed testing procedure**, the **Boole-Bonferroni bound**, and the **Holm-Bonferroni method**. **This article** on multiple comparisons contains more information, and **this article** contains more information about the false discovery rate (FDR), and methods for controlling that instead of the familywise error rate (FWER).

Drawing Conclusions

Once we have figured out which metrics have significant changes, what comes next? You have to decide what your results do and don't tell you. If you have statistically significant results, then that means that you're unlikely to have zero impact on the user experience. The question is: do you understand the change. Or do you want to launch the change?

What if I have a statistically significant change in some metrics but then in other I don't. Maybe you know that for some small changes, a change in one metric but no change at all in the other metrics is perfectly fine. But if you were to see those same results for a big change, that would probably indicate that there's something wrong.

What if your change has a positive impact on one slice of your users, but for another slice there's not impact or there's a negative impact. For example, we like to bold certain words to give them more emphasis. In the Latin alphabets, works really well for this matter. However, if you try to bold in Japanese or Korean, a bolded character is really hard to read. Maybe for those slices you want to use a different color as opposed to doing a bold.

How do you decide whether to launch your change or not? You really have to ask yourself a few questions:

- Do I have a statistically significant and practically significant results in order to justify the change?
- Do I understand what that change has actually done with regards the user experience?
- Is it worth it?