# Lesson 1: Overview of A/B Testing

## Why are we doing A/B testing?

A/B tests allow you to determine scientifically how to optimize a website or a mobile app by trying out possible changes and see what performs better for you users.

## What is A/B testing?

A/B testing is a general methodology used online when you want to test a new product or feature. And what you’re doing is you’re going to take two sets of users and you’re going to show one set – a control set – your existing product or a feature, and then another set – your experiment set – the new version. Then analyze how these users respond differently, in order to determine which version of your feature is better.

## Should we A/B testing for everything?

No. An useful analogy: A/B testing is really useful for helping you climb to the peak of your current mountain, but if you want to figure out whether you want to be on this mountain or another mountain, A/B testing isn’t so useful.

## What are the things you can’t do with A/B testing?

A/B testing isn’t as useful testing out new experiences. Two issues involved:

* What is your baseline for comparison?
* How much time we need in order to actually have your users adapt to the new experience?

Because we want to determine what is going to be the plateaued experience so that I can actually make a robust decision.

Time can be a problem for some cases. For example, if you are running an apartment rental recommendation website, but as a matter of fact, people don’t rent an apartment that often. What you really want is return business or maybe you want to grow your business by referrals to other people who like your service. But the reality is, in the scope of an experiment, it’s going to be really hard to measure whether people actually come back to you from more referrals. (No data source and not easy to define metrices…)

## Some examples of when to use A/B testing

* Online shopping company: Is my site complete?

No. maybe we can try specific product but can’t answer in general. Just don’t have the missing information. We could ask users directly if there’s something missing.

* Add premium service?

No. Premium service requires users to opt in, so randomly assigning people to one group or another wouldn’t really work. It might help us to gather information, like how many users would choose to upgrade to this service but we cannot fully test out the change.

* Movie recommendation site: new ranking algorithm

Yes. There are clear control and experiment groups: the old algorithm and new algorithm. And the metrices are also clear to use, such as CTR or actually watching movie based on recommendations.

* Change backend: such as page load time, etc.

Maybe. Depend on whether you have the computing ability to run both versions of the backend at once. If we do, we can definitely test the results of the change. A/B testing requires run both experiments simultaneously.

* Website selling cars: will a change increase repeat customers or referrals?

No. Too long and don’t have data.

* A company wants to update brand, including their main logo.

No. might be helpful for collecting user experience information. Changing a logo can be surprisingly emotional, users might need some time to get used to the new logo, so you wouldn’t want to make a decision based on a short time window of data collected in an A/B test.

* Test layout of initial page.

Yes. Clear control and clear metrics.

## Other techniques beside A/B testing

* Analyze users’ logs retrospectively or observationally to see if a hypothesis can be developed about what’s causing changes in their behavior.
* Based on the observed hypothesis, design an experiment and run A/B test to compare results.

The above two options can be complementary to each other.

* User experience research
* Focus groups
* Surveys
* Human evaluation…

A/B testing can provide us broad quantitative data while the above options can give us qualitative data that are really complementary to A/B testing.

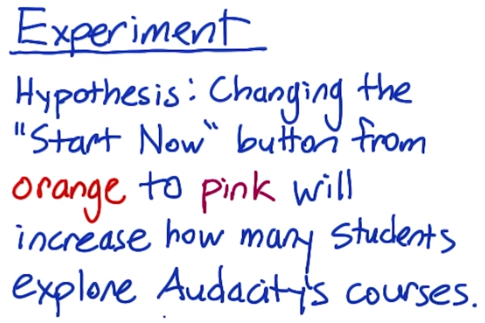
## Key for A/B testing:

We need to have a ***consistent response*** from both control and experiment groups, so that we can actually really determine and structure the experiment. Then we can determine whether there’s a significant behavior change in our experiment group as opposed to in our control group.

What is the difference between clinical trial and online A/B testing data?

* Traditional medical trial might only require 10, 20 or 50 participants. But we know a lot of information about patients.
* Online data is big in size-wise but kind of lower resolution. We might have millions of users, hundreds of thousands of clicks or respondents but not much specific info for individual end-user.

## What is typical user flow for an online study website?

This funnel is not total linear from top to bottom. Sometimes there’s a lot of back and forth swirl between the different states, and repeat visitors who skip over intermediate steps. This kind of behavior will determine what kind of metric choice we should make during the test.

## Which metric to use for the above hypothesis?

* Total # of courses completed?

No. It takes long time to finish a course. Not practical.

* # of clicks on the “Start Now” button?

Sounds good. We are assuming more clicks on the button will improve the bottom levels at customer funnel. But what if we have different sizes for two groups (different # of page views in two groups)?

* CLT?

Good.

* Click-through-probability?

Better. It’s more accurate compared to CLR.

We will use click-through-probability as our metric for this hypothesis. The original hypothesis is being refined as:

Changing the “Start Now” button from orange to pink will increase the click-through-probability of the button.

Which will also increase the final business metric – total number of courses completed.

## How to decide between click-through-rate and click-through-probability?

Generally speaking, we use a **rate** when we want to measure the usability of the site, and a **probability** when we want to measure the total impact. For example, if we want to measure the usability of a particular button, we use rate to describe how often users can find that button. But if we want to know how often users went to the second-level page of your site, we should use a probability. Because we don’t want to count if user just double-clicked, or did they reload, or all those types of issues.

In our example, we’re interested in whether users are progressing to the second level of the funnel, so we picked a probability.

## How to compute click-through-probability?

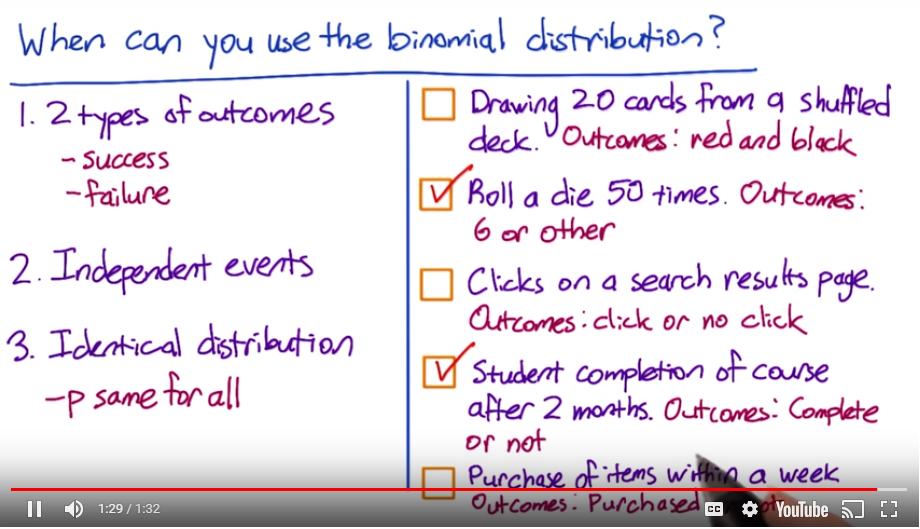
1. Work with engineers to modify the website and capture the event on every page view. And then whenever a user clicks, we also capture the click event.
2. To compute the **rate**, we just sum the page views and sum the clicks, then divide. For **probability**, match each page view with all of the child clicks, so that you count at most, one child click per page view.

## Which distribution should we use to describe click-through probability?

We can use binomial distribution if our results are any two exclusive outcomes. The metric that we will compute is the probability. We define Click as “Success” and No Click as “Failure”.

As we increase sample size, the binomial distribution starts to look more and more like normal distribution. If we know the probability of success for each toss and the number of total toss,

Assumptions for using binomial distributions:



* Sample without replacement. Each card reveals partial info for next card.
* Every rolling (each event) is independent and has identical distribution.
* If someone isn’t satisfied with search results, they will start searching again with different keywords. Therefore, these two events are correlated to each other, but it’s hard to determine whether or not two events come from the same user.
* Students are from all over the world and majority should not be correlated to each other.
* One person can add several items into shopping cart and purchase all of them.

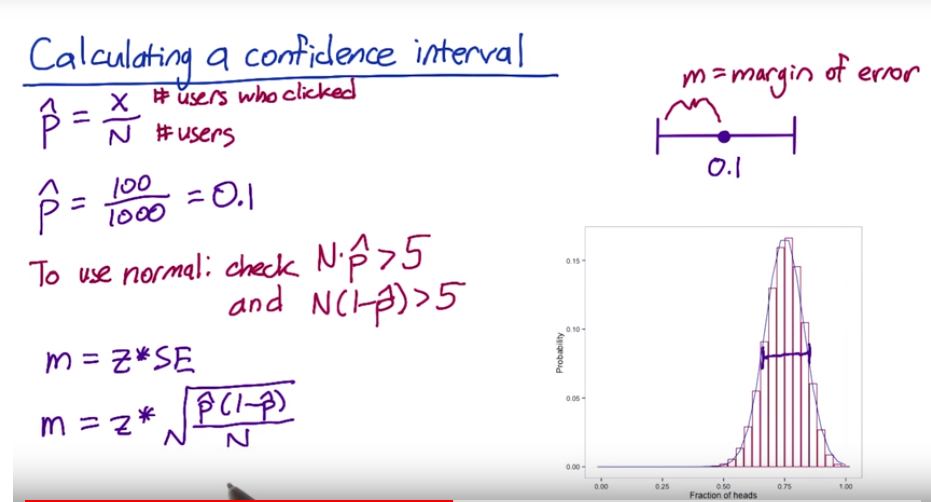
## How do we use binomial distribution for our click-through probability?

The benefit of knowing click-through probability should follow 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 does 95% confidence interval mean?

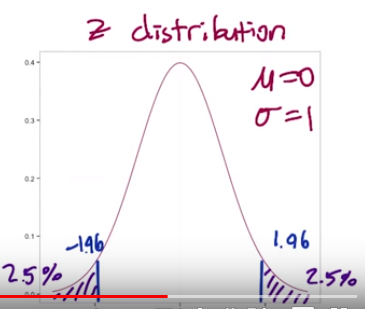
If we theoretically repeated the experiment over and over again, we would expect the interval we construct around our sample mean to cover the true value in the population around 95% of the time.

## How to compute a confidence interval for a binomial distribution?

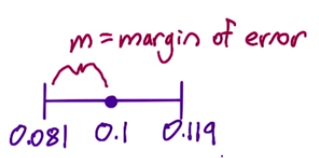


Rule of thumb: if , then we can use normal distribution to approximate the binomial distribution.

Next is to calculate Z-score for two-sided test with 95% confidence interval,



Therefore, , the CI will be:



Margin of error (the half width of the confidence interval)

Therefore, m is related to the proportion of success and the sample size, which means we need to consider the proportion of successes when we decide how many samples we should collect.

* When the success probability is farther from 0.5, then SE will be smaller, m will be smaller, CI will be tighter.
* If the sample size is larger, SE and m will be smaller, then the CI will be tighter.

## Why using hypothesis testing or inference?

Hypothesis testing is a quantitative way to establish how likely it is that our results occurred by chance.

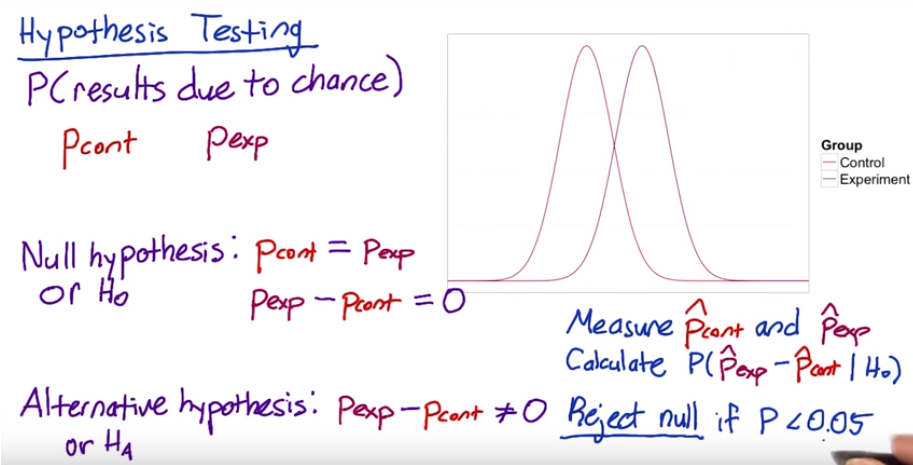
The null hypothesis: there is no difference between control and experiment groups.

Alternative hypothesis: higher or lower or any kind of difference?

## How to perform hypothesis testing for two groups?

Control group: the orange button with click-thru probability

Experiment group: the pink button with click-thru probability



## 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:

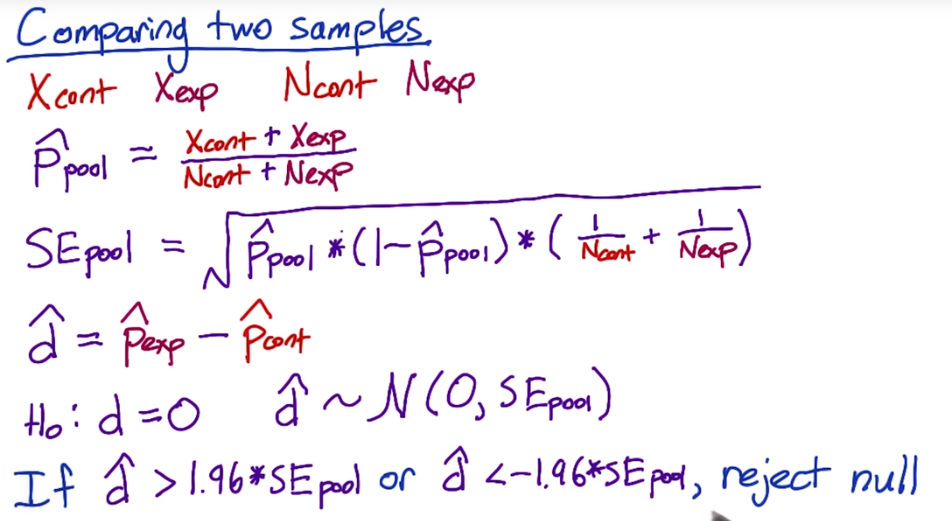
1. A statistically significant positive result
2. No statistically significant result

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.

## How to compare two samples with different sample size?

Use Pooled Standard Error.

X is the # of users who click on the button, and N is the total # of users in each group.



## Practically significant vs. statistically significant

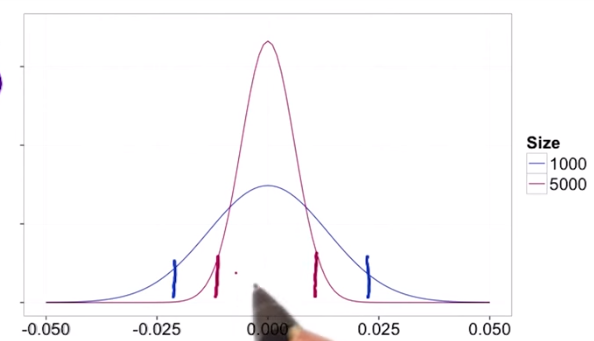
Practically significant (**substantive**) can be quite different among online websites, depending on the business perspective of view. It’s decided from a business perspective based on what change in the metric is practically significant. In other words, what size of changes matters to us?

Statistical significance is about repeatability. We want to make sure we setup our experiment right and results are repeatable, so it’s statistically significant. We also want to make sure that if we see a change that we’re interested in from a business standpoint, so it will be both practically significant and statistically significant. Therefore, we need to size your experiment appropriately, such that the statistical significance bar is actually lower than the practical significance bar.

## What is statistical power?

In our experiment, we need to decide how many page views we need in order to get a significant a statically significant result, which is called as statistical power.

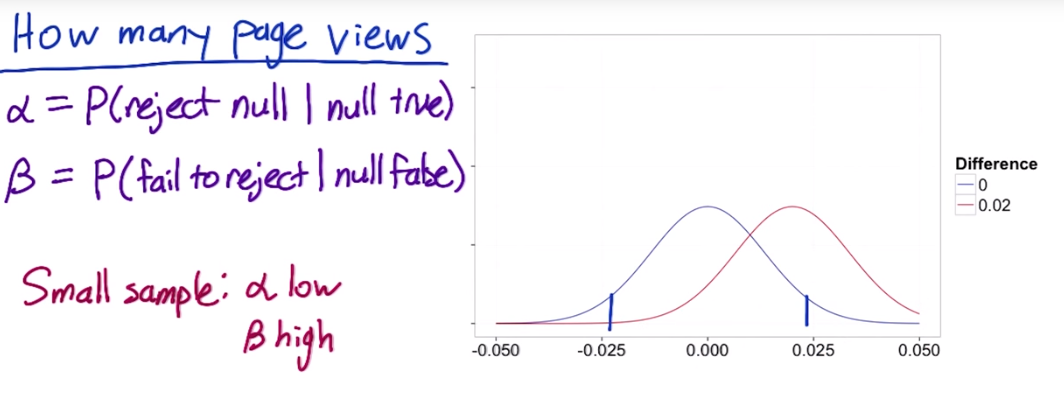
The key thing to keep in mind is that power has an inverse trade-off with size. The smaller change we want to detect, or the increased confidence that you basically want to have in the result, means that you have to run a larger experiment, so more page views in your control and experiment groups.



In order to keep confidence level same for both big and small sample cases, the cut-offs for larger sample to reject the null will be closer to zero, therefore it will be easier to reject H0 compared to small sample problem.

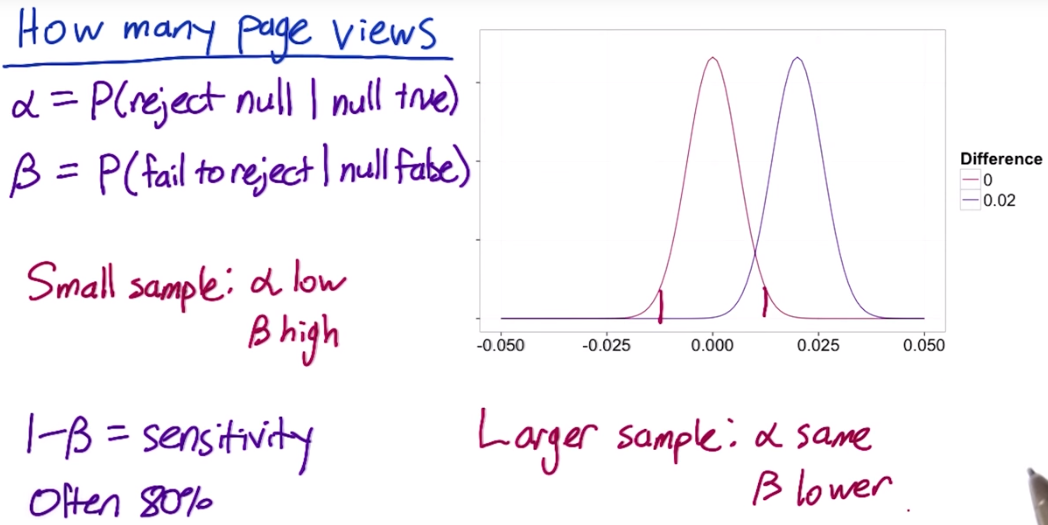
## How page views affect sensitivity?

For small sample problem,



The overlapped area from two distribution is large, therefore it’s not easy to reject H0, which means alpha is low and beta is high.

For large sample problem, it’s the opposite:



The overlapping area between two curves are small. It’s easier to reject H0 compared to small sample problem.

## How to calculate number of pages views needed?

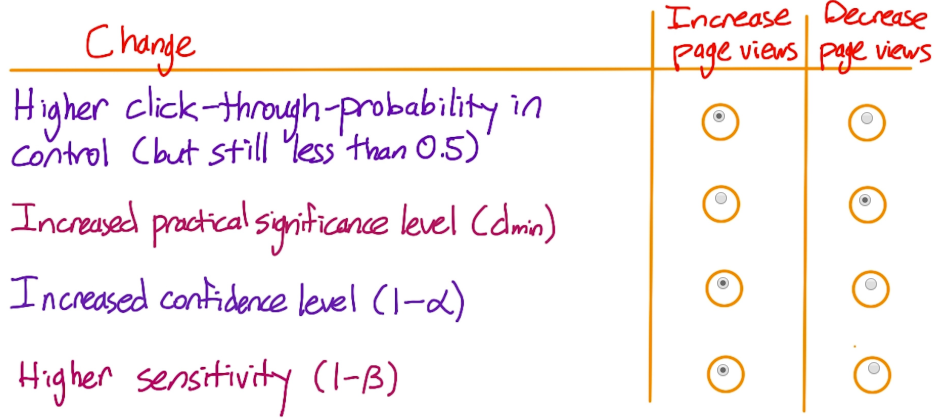
Use this calculator to determine how many page views we’ll need to collect in our experiment.

<http://www.evanmiller.org/ab-testing/sample-size.html>

A few key points for this online calculator:

* Minimum Detectable Effect: Practical significance level.
* Check “Absolute”
* Standard values for alpha and beta: alpha = 0.05, beta = 0.2

## How number of page views varies with parameters?



* Higher click-through-probability in control? (but still less than 0.5)

The upper term is and its maximum happens when , therefore as p gets closer to 0.5, we also need to increase N to reduce SE back to its original level.

* Increased practical significance level (d\_min) aka Minimal Detectable Effect?

This means we no longer care about detecting a 2% change and we would need the change to be larger than 2% before you cared about detecting it.

Larger changes are easier to detect, so we shouldn’t need as many page views.

* Increased confidence level (1-alpha)?

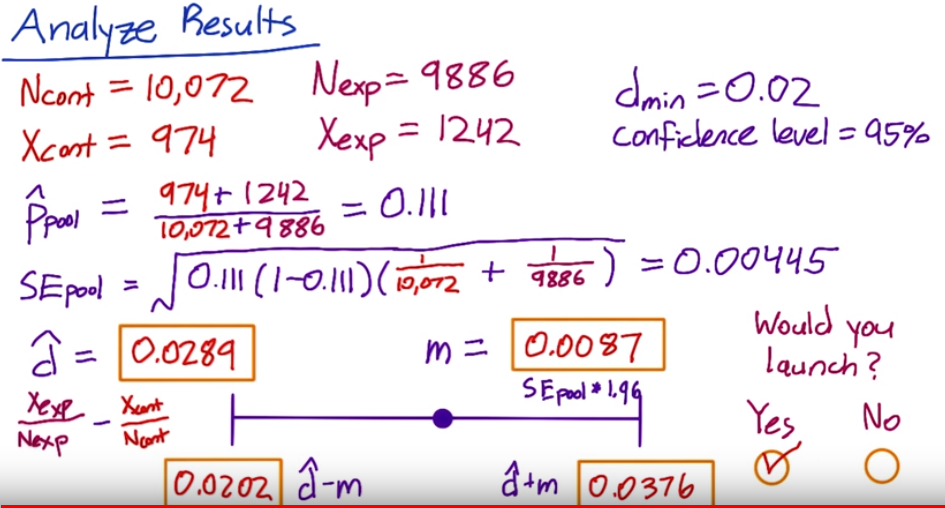
This means we want to be more certain that a change has occurred before you reject the null hypothesis. In essence, we are being more conservative.

We could accomplish that by rejecting the null less often, but our sensitivity will be going down. Therefore, in order to keep sensitivity the same, we will need to increase the number of page views we collect.

* Higher sensitivity (1-beta)?

If we want to increase the sensitivity for our experiment, we need to collect more page views to narrow the distribution.

## Analyzing results?



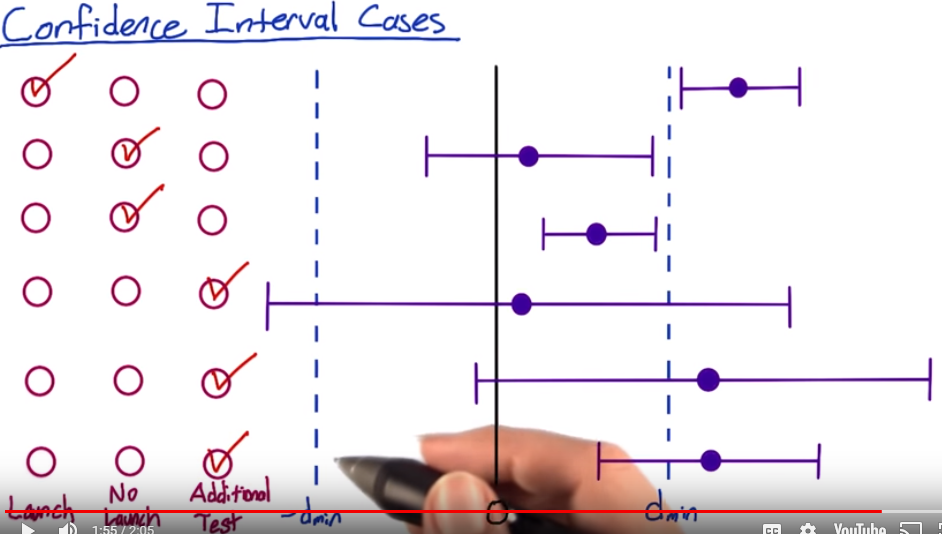
* is the difference between click-through rates from control and experiment groups.
* is the margin of error, which equals to .
* The confidence interval is

For statistical significance,  , therefore we can reject the null.

Based on the confidence interval, we can conclude that it is highly probable that click-through probability changed by at least 2.02%, which is higher than dmin (2%).

Therefore, we have both statistical and practical significance and we should launch this new version.

## How to make decisions based on Confidence Interval?



Assume that a positive change in click-through probability is desirable, and a negative change is undesirable.

* The lowest CI is still higher than practical significance, therefore we should definitely launch this new feature.
* This result is often called neutral. There is no statistically significant change from 0 since the confidence interval includes 0 and we are also confident that there’s not a practically significant change because the left boundary of CI is still inside dmin. So no launch.
* Our result is significant because 0 is outside out CI, and we are confident there will be a positive effect. But it’s still smaller than dmin. Therefore, there is a statistically significant change but no practically significant change. So no launch.
* This CI covers all possibilities, which means the new feature call cause 10% increase and also can cause 10% decrease. Conclusion is we do not have enough power to draw a strong conclusion. We should run an additional test with greater power.
* The point estimate is beyond the practical significance level, but CI overlaps 0, so there might not even be a change at all. So run with another test with greater power.
* Best guess is that there is a positive change but it’s also possible our change is not practically significant. Therefore, additional test.