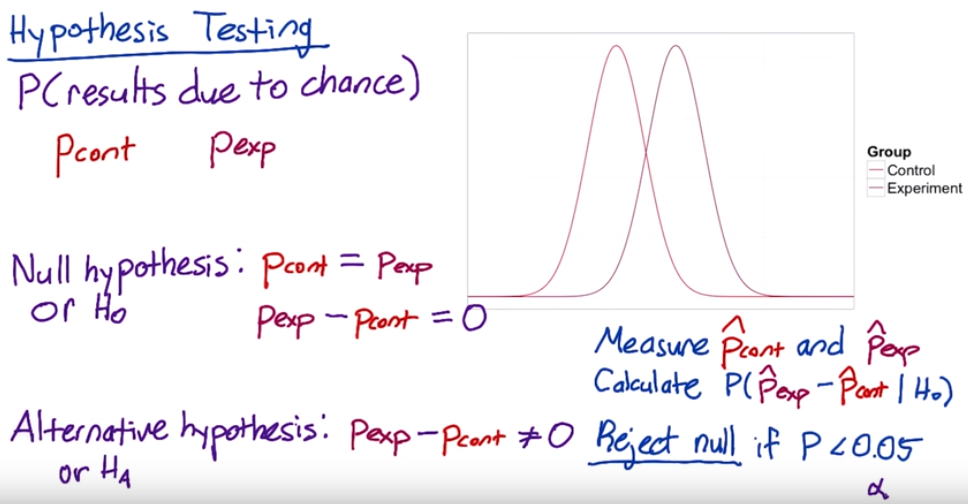
Lesson 1:

Metric

CTR

CTP = unique visitors who click/unique visits to page, to measure a total impact, to avoid the same customer click the same button multiple time in order to go to the second level of the funnel.

Video 17:



**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:

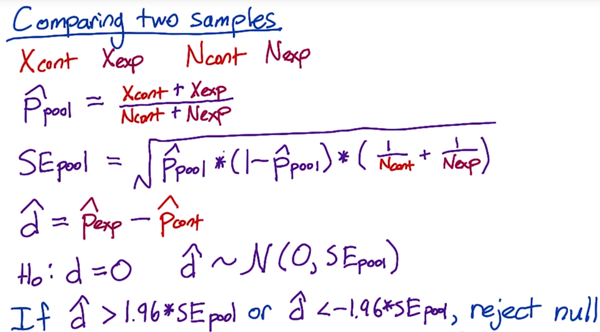
1. A statistically **significance** positive result
2. A statistically **significance** negative result
3. No statistically **significance** difference.

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 **significance** positive result
2. No statistically **significance** 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 **significance** 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.

Video 19



We have control how many pages views go into our control group and experiment group. Then we decide how many page views we need in order to get a statistically **significance** result.

Video 22 no fully understand



**Video 23 calculate number of Pages Views Needed**

Use a table <https://www.research-advisors.com/tools/SampleSize.htm>

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

How many page views will we need for each group?

We collect 1000 unique pages view, of which 100 resulted in a click, practical **significance** is 2% (2% change in click through rate is **significance**)

α=0.05 β = 0.2 (pretty standard) ,

Sensitivity is 1-β = 0.8, (people often choose 80%)

Video 24

Standard error depends on click through probability SE = , so if increase the conversion rate(click Through Probability) will need to increase sample size to keep SE constant.

**Video 25 Calculating Results**

= 10072 = 9886 = 0.02 (practical significance level)

= 974 = 1242 confident level = 95%

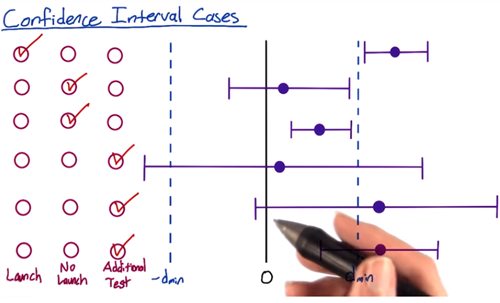
pool = = 0.111

SE pool = = 0.00445

Estimated difference: = 0.0289 margin error: m = SE pool \* 1.96

Confident internal : – m = 0.0202 between = 0.0376

We can conclude it’s highly probable that click-through probability changed at least 2%. We want to look for statistical and practical significance.



A/B test:

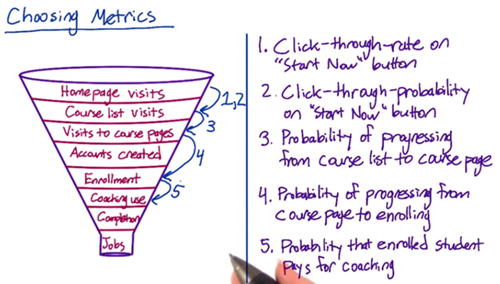
1. What risk is company being exposed to?
2. What benefit might be the outcome of the test?
3. What other choices do we have?
4. What expectation of privacy do they have?

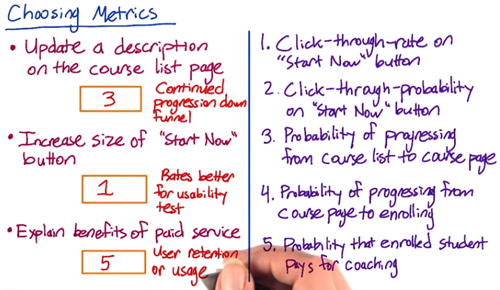
Lesson 3

Define metric

Video 6:

create a funnel: the reason this’s call funnel is that you have fewer and fewer user that get to each stage of the funnel.



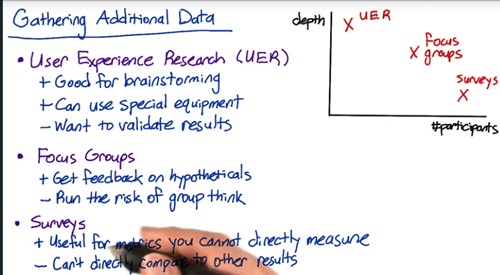


how many months were they active on the site? You want to get some baseline for that. And then given that they took a second course, then you might also want to trigger survey that happened within your site. Get a survey that says “ are you considering…

**video 7: Techniques to Gather Additional Data**

<https://s3-us-west-2.amazonaws.com/gae-supplemental-media/additional-techniquespdf/additional_techniques.pdf>

Eye-tracking cameras: Tobii is the most common manufacturer of cameras that can track eye movement to see what users are looking at



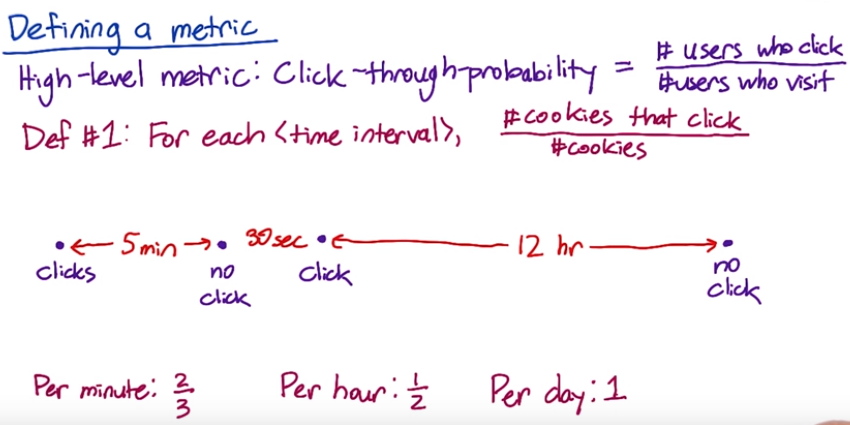
**Video 8: Other Techniques: Example**

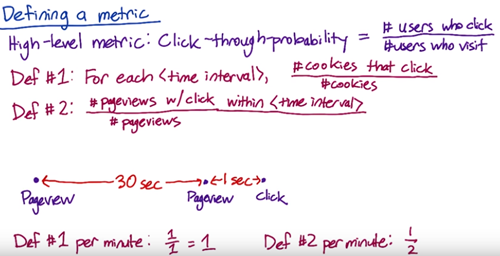
Buy external data from Comscore, hitwise, Nielsen

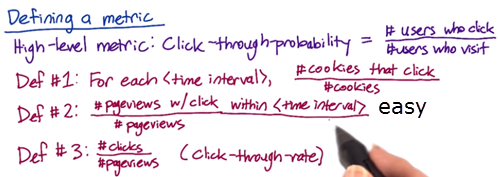
For example, watching the customers try to complete the checkout can help you figure out

* Do they understand where to click?
* Can they find everything on the screen?
* Tracking the latency

**Video 13 Metric Definition: Example**







Def #1 (Cookie probability): For each <time interval>, number of cookies that click divided by number of cookies

Def #2 (Pageview probability): Number of pageviews with a click within <time interval> divided by number of pageviews

Def #3 (Rate): Number of clicks divided by number of pageviews

There’s some technical issue(double click, back button caches page, click bug) that will affect the click/views accuracy. Check the video quiz

**Video 14**

There’s some fraud, spam, want to filter that out. You don’t want to dilute your result.

One way to figure out whether you’re biasing or de-biasing your data by applying these filters, is to slice your data, and what this means is that you’re computing your metric on a bunch of disjoint sets. So for example, by country, by language, or by platform. To identify spam and spot something that looks unusual.

You’re building intuition, you have to know what changes you’re going to be expected versus unexpected.

**Video 15 filtering and segmenting**

Looking at different segment of your data can be useful for evaluating metric definitions. Look at how the different definitions vary by segment.

e.g. There’ a weird spike that showed up sometime last week. *Analyze the spike by dividing current week’s data by last week and plot it, to see if the spike still there or not, if spike disappears, which means there’s a weekly variation.* If spike is there, one way to figure 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 trying to look 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 Canada, so that one country was causing the entire spike.

You can also see some weekly variation.

If it’s being artificially inflated by double clicks being recorded. It’s hard to know how much higher you would expect the rate to be.

**Video 16 Summary Metrics - establish a few characteristics for your metric**

**First one** is going to be the sensitivity and robustness, to be sensitive enough to detect a change

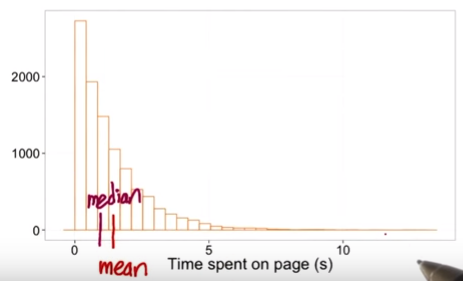
**Second one** is to characterize what the distribution of metric looks like. e.g. ideal way of doing this is to do a retrospective analysis, to compute a histogram. E.g. you’re going to have all the different value for the load time on the x axis, the y axis is going to be the frequency, so how often individual events have that particular load time. When you plot the histogram, you get a shape, a distribution, and what you’re looking at is what that shape is. If it’s a very normal shape, then a mean or median is going to make sense. As it becomes more one sides, or lopsided, you might want to go more for a 25th, or a 75th, or a 90th percentiles.

**4 categories:**

* sum or count
* Distributional metric: the means, medians, the 25th, 75th percentiles
* Probabilities (zero or one) and rates
* Ratios

For example, let’s measure the rate at which users click on a result on our search page, analogously, we could measure the average staytime on the results page before traveling to a result. In this case, you’d probably see what we call a [Poisson distribution](http://en.wikipedia.org/wiki/Poisson_distribution),

or that the stay times would be [exponentially distributed](http://en.wikipedia.org/wiki/Exponential_distribution).



How many users get information from the page? You might want to use something besides the mean or the median. Maybe the 75th percentile or the 90th percentile.

The load time mean can be heavily influenced by the network issue.

**Video 18 is the above**

<https://youtu.be/jmqZG6roVqU>

Another common distribution of user data is a “power-law,” [Zipfian or Pareto distribution](http://en.wikipedia.org/wiki/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 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.

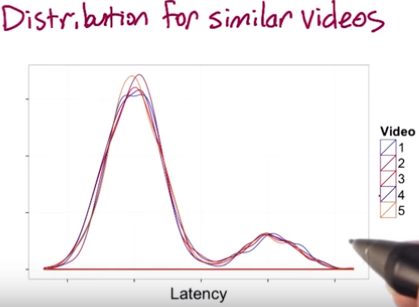
The key here is not to necessarily come up with a distribution to match if the answer isn’t clear - that can be helpful - but to choose summary statistics that make the most sense for what you do have. If you have a distribution that is lopsided with a very long tail, choosing the mean probably doesn’t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

Video 19 **measure Sensitivity and Robustness: Example**

e.g. choose summary metric for the latency of a video, need to look at sensitivity and robustness for each of 25th, 75th , 80th, 85th,90th,99th percentiles summary metric.

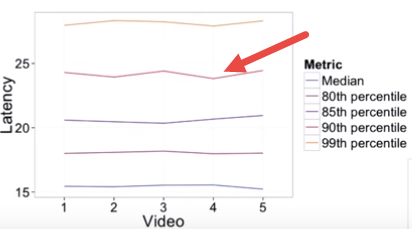
How? Way 1. Do the retrospective analysis.

1. Segment the data by different videos (in other word, look at the distribution of load times per video), if we look at the distribution of a single video, plot a histogram with a density line, to compare with multiple video.



I get similar distribution of load time for the different videos. You can see two peaks here, a fairly long load time, and more people with a shorter load time.

1. Plot a few different summary metrics by video, in theory, since these videos are all comparable, there should not be too much difference between the different videos for a good metric. WHY??? But the 90th and 99th % are zigzagging around a bit, which means not robust enough.

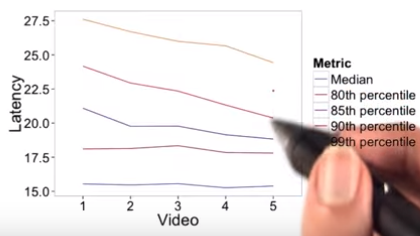


How? Way 2. Look at the actual experiment

1. Preferably, if we had experiments that changed the resolution. That should impact the latency, and if it doesn’t, then our metric isn’t sensitive enough.

 which means it should have highest low time. In fact, you do see that video one is off to the right a bit more.

1. We should see the latency going down as we increase the video number, that is , we have a lower resolution.

 in fact, for some of these metrics we do see that, but for the median and 80th %, they don’t really seem to be moving. This’s a good indication that the median and 80th are not sensitive enough. They don’t show a change when we do make a change that we care about. In this care, the 85th % might be a good choice of a metric that’s both robust and sensitive.

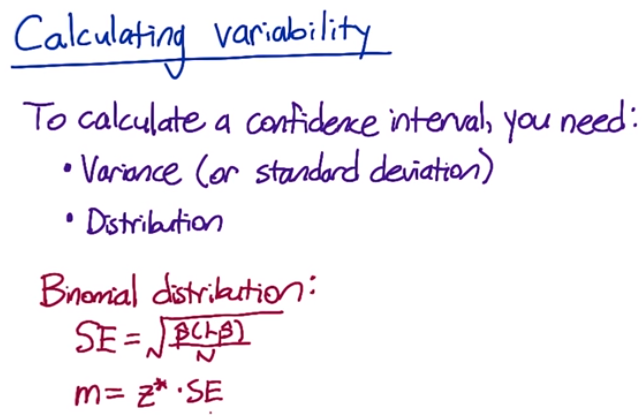
Video 20

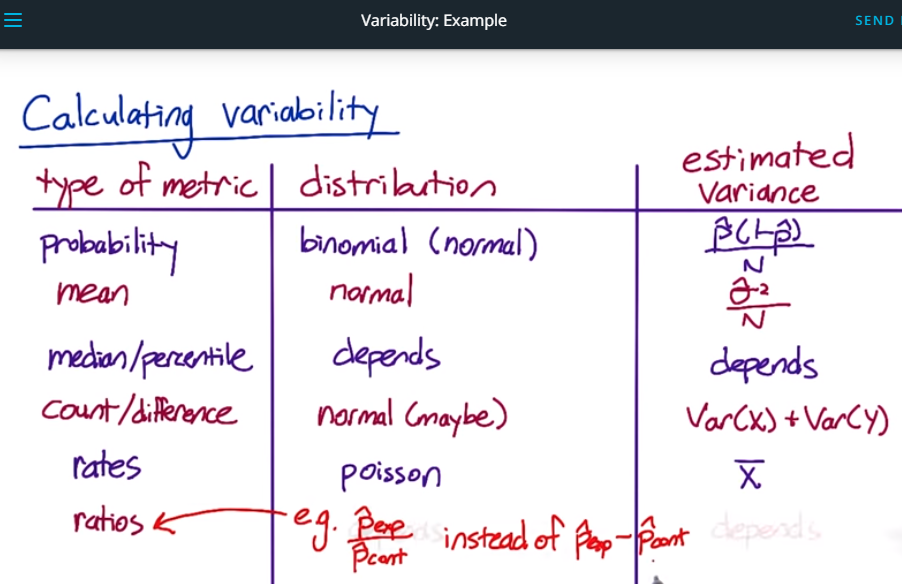
How to compare two group’s difference?

**Relative differences in probabilities**

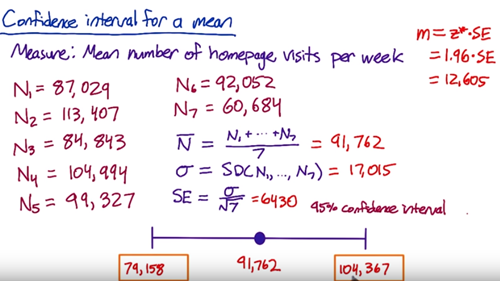
For probability metrics, people often use percentage points to refer to absolute differences and percentages to refer to relative differences. For example, if your control click-through-probability were 5%, and your experiment click-through-probability were 7%, the absolute difference would be 2 percentage 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!

Video 21 , 22 **Variability**





**Variability**



Video 23 **Nonparametric Answers**

Compute the Nonparametric confidence internal

Video 24 Empirical Variability (**check the variability within the same data group**)

For more complicated metrics, you might need to estimate the variance empirically instead of computing it analytically.

When computing a metric, you’re making an assumption about the underlying distribution of the data. For simple metric, it makes sense absolutely. But for complex metric, the distribution can be very weird, you might want to shift to an empirical estimate.

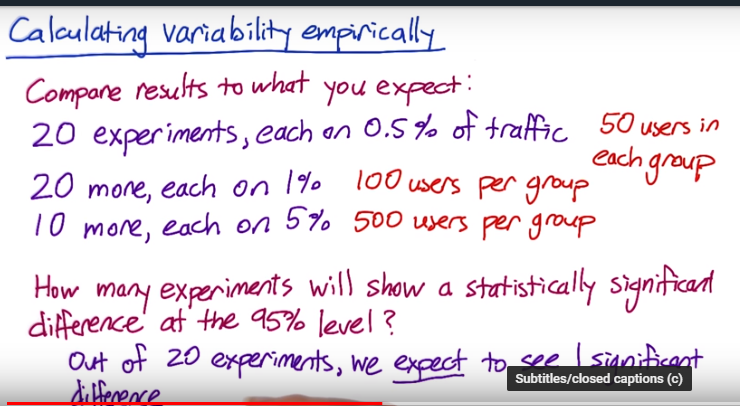
The underlying variability

We use A vs A to test for the sensitivity and robustness of a metrics. For example, if you see a lot of variability in a metric in an A vs A test, it’s probably too sensitive to be useful in evaluating a real experiment. So how can you pin down the variability with these A/A test. How many do you usually need to get a good sense? One of the popular methods in statistics is Boot strap, which takes a big sample and split it up into a bunch of small samples. (e.g do only 1 experiment, then randomly pick the data point and average them, repeat the process)

Video 25 A/A test data

[This spreadsheet](https://docs.google.com/spreadsheets/d/17wWNY2jkDlG9BDMYQq2l-ku_8HGajXuF2Zvy__dBEL4/edit?usp=sharing) contains the data, calculations, and graphs shown in the video.

Video 26 empirical variability sanity check



Video 27 how to calculate boot strap?

Video 29 summarize what you learn

Latency is the challenge, which tends to be lumpy. Coz you have user who have very different connection speed, which cause lumpiness in the distribution, you don’t want to use the MEAN, start looking at the higher percentile metric.

--Learn 4 Design A/B test

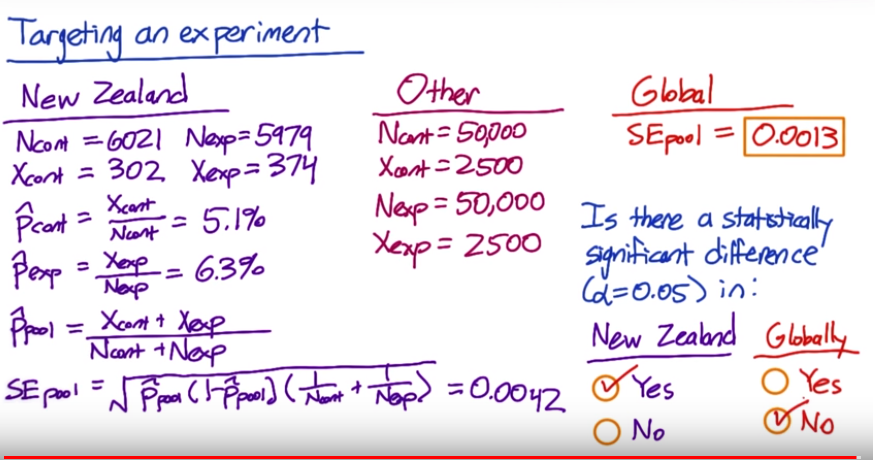
1. choose population, only US?

Unit of Diversion:

* user id, like user log in account
* cookie
* mobile id, tie to specific mobile device, it’s unchangeable by the user, but it doesn’t have cross device consistency.
* Ip address

If you change the layout of the pag, you use cookie, coz user might not sign in.

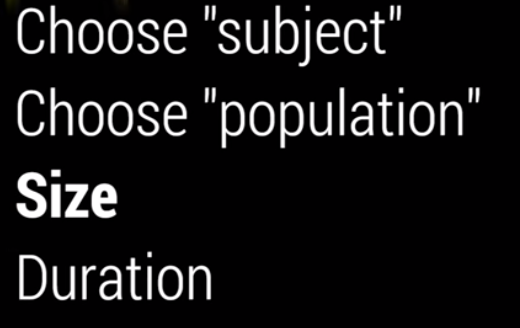
Video 12 Target population



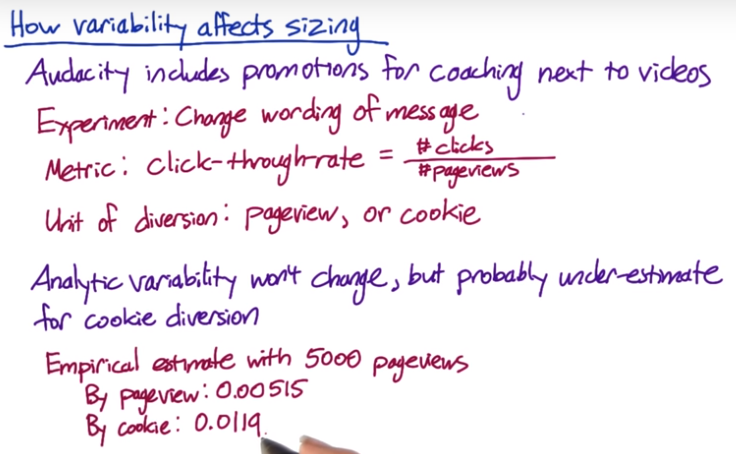


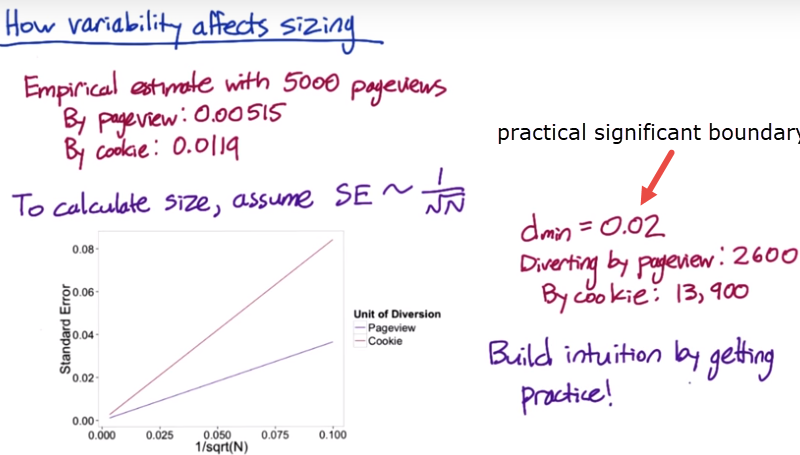
Video 13 population vs cohort

Track increased usage of a website, you want a cohort



Video 17 Sizing





VIDEO 18 how to decrease the experiment size? 难

**Way 1 increase dmin, α, β**

Increase dmin significance boundary, not try to detect a smaller change

Increase α, β accept a higher probability of a false positive or false negative

**Way 2 Change unit of diversion from cookie to pageview**

The variability of the metric will probably decrease and be closed 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.

Makes unit of diversion same as unit of analysis, the question here is whether the less consistent experience will be acceptable. In the case, if we recalculated the empirical estimate of the standard error suing the pageview as the unit of diversion, they might find that the new standard error was 0.0209 for the same sample size. Only 34000 pageview per group would be necessary, which is quite an improvement.

**Way 3 Target experiment to specific traffic**

Non-english traffic will dilute the result, which would increase the number of pageviews needed. (Hugo understanding is other language has different variability, affect the total variability, Also, only focus on English traffic will reduce pageviews needed), Filtering the traffic could impact your choice of practical significance boundary. Variability decrease and the significance boundary will increase, so size will be smaller.

Video 21. learning effects

We’ve developed methodology where we use both pre-periods and post-periods. pre-periods is before we run the experiment, we do AA test, on the same set of users. To measure system variability. post-periods is after we do the experiment, we do AA test again.

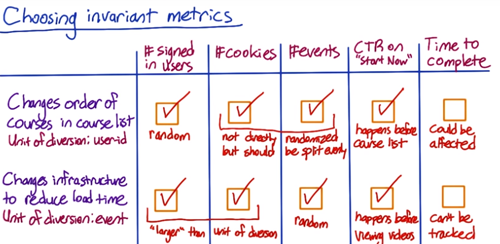
**Lesson 5**

Video 2 Sanity check

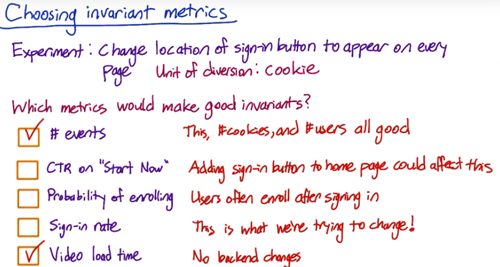
Choose good invariant (unchanged) metrics to verify, e.g. your experiment is for CTR, we need to make sure the load time is invariant (unchanged).

1. Population sizing metrics check based on your unit of diversion.
2. Actual invariants check, those metrics shouldn’t change when you run your experiment

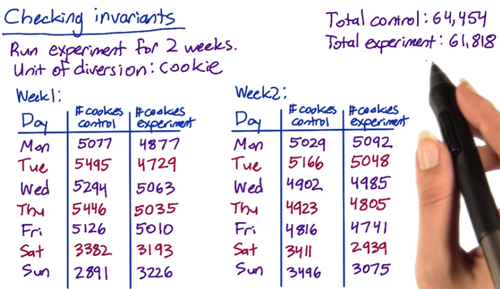
Video 3

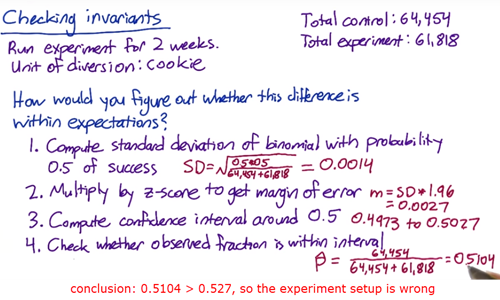


**Video 4**

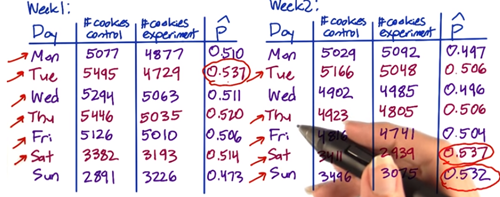


**Video 5, 6: Real example on how to check invariants**





So, you have to go back to check the day by day data,



This points to an overall problem rather that a problem on a specific day. Talk to engineer if any setup is wrong. Try slicing to see if one particular slice is weird.

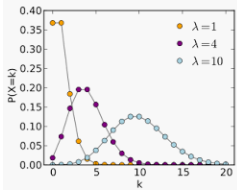
**Video 8 What not to do if your results aren't significant**

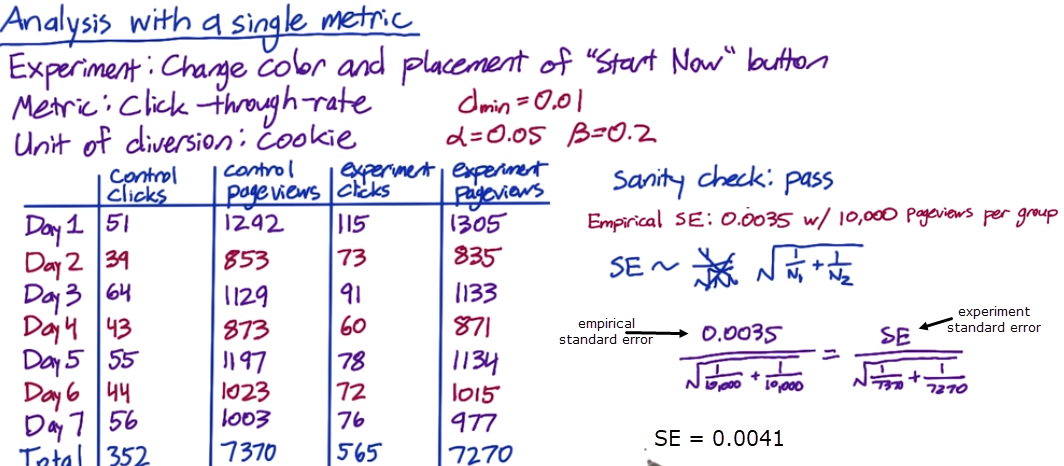
Carrie gave some ideas of what you can do if your results aren't significant, but you were expecting they would be. One tempting idea is to run the experiment for a few more days and see if the extra data helps get you a significant result. We also estimate the magnitude and direction of the change.

<https://www.evanmiller.org/how-not-to-run-an-ab-test.html>

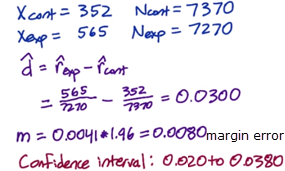
video 9 Quiz Single Metric example **the most useful**

click through probability is binomial distribution

click through rate is poisson distribution  it’s a good idea to estimate the variance empirically.



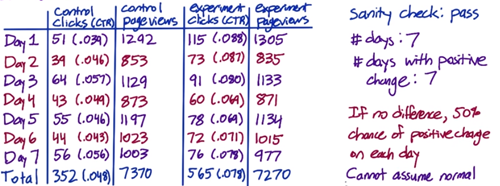
Then estimate the difference between the click through rate and the control/experiment group

 if d hat is a negative #, take absolute

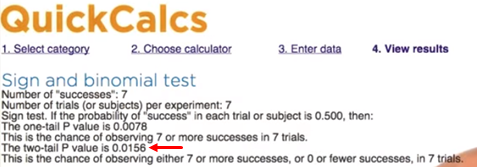
**Recommend launching the next design:** coz the confident interval doesn't include the practical significant boundary 0.01, meaning I can be confident at a 95% confidence level that the true change is large enough to be worth launching.

the confident interval doesn’t include zero, the result is statistically significant. It’s unlikely that there was no real difference.

Finally, do a sign check,



Since 7 days is not enough for binomial to closely normal, use online calculator due to only 7 days <https://www.graphpad.com/quickcalcs/binomial1.cfm>



P value is the probability of observing a result at least this extreme by chance. Since this less than the chosen alpha of 0.05, the sign test agrees with the hypothesis test.

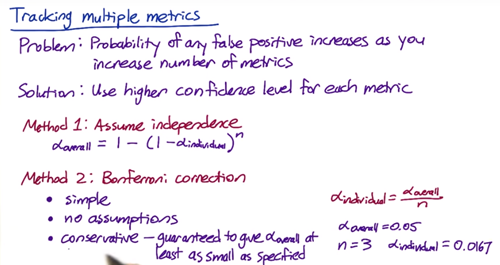
Hey what if the p value is greater than 0.05, meaning this test is not significant at alpha = 0.05 level. What happen? The sign test has lower power than the effect size test, which is frequently the case for nonparametric tests.

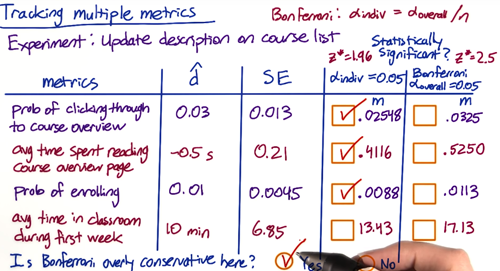
I would dig deeper to see if the change has a small effect during the weekend, larger effect on weekend. Then calculate weekend weekday separate.

WHY**, Simpson’s paradox**: meaning there’s a bunch of different subgroups in your data. And within each subgroup, the result is stable, but when you aggregate them all together, it’s a mix of subgroups that drive your result.

we have some subgroup like people use it more on weekday or weekend, you might find that, for example, new users are correlated with weekend use. You need to find out how many users from each group.

**Video 14: Multiple comparison for multiple metric, 0:39 why z =2.5**

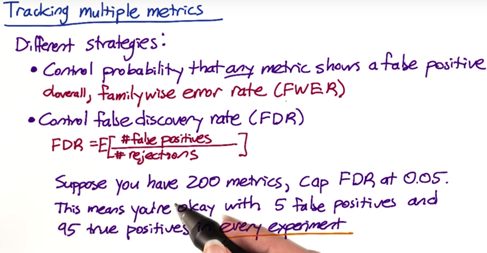




Video 15

**Less conservative multiple comparison methods**

The [Bonferroni correction](http://en.wikipedia.org/wiki/Bonferroni_correction) is a very simple method, but there are many other methods, including the [closed testing procedure](http://en.wikipedia.org/wiki/Closed_testing_procedure), the [Boole-Bonferroni bound](http://en.wikipedia.org/wiki/Bonferroni_bound), and the [Holm-Bonferroni method](http://en.wikipedia.org/wiki/Holm%E2%80%93Bonferroni_method). [This article](http://en.wikipedia.org/wiki/Multiple_comparisons_problem) on multiple comparisons contains more information, and [this article](http://en.wikipedia.org/wiki/False_discovery_rate) contains more information about the false discovery rate (FDR), and methods for controlling that instead of the **familywise error rate** (FWER).



You have to understand how user are reacting to the change?

**Video 18: Ramp up your experiment (increase the % of the traffic )**

in some case, during the ramp up, the effect will be flattened out.

Holdback: a way to capture Seasonal or event driven impact. A set of unchanged users for the control group.

Any other reason that flattens the effect? Novelty effect and change aversion. The humans keep changing their behaviors. Once they adopt the change, a cohort analysis will help.

# Final Project Instructions

## Experiment Overview: Free Trial Screener

At the time of this experiment, Udacity courses currently have two options on the course overview page: "start free trial", and "access course materials". If the student clicks "start free trial", they will be asked to enter their credit card information, and then they will be enrolled in a free trial for the paid version of the course. After 14 days, they will automatically be charged unless they cancel first. If the student clicks "access course materials", they will be able to view the videos and take the quizzes for free, but they will not receive coaching support or a verified certificate, and they will not submit their final project for feedback.

In the experiment, Udacity tested a change where if the student clicked "start free trial", they were asked how much time they had available to devote to the course. If the student indicated 5 or more hours per week, they would be taken through the checkout process as usual. If they indicated fewer than 5 hours per week, a message would appear indicating that Udacity courses usually require a greater time commitment for successful completion, and suggesting that the student might like to access the course materials for free. At this point, the student would have the option to continue enrolling in the free trial, or access the course materials for free instead. [This screenshot](https://www.google.com/url?q=https://drive.google.com/a/knowlabs.com/file/d/0ByAfiG8HpNUMakVrS0s4cGN2TjQ/view?usp%3Dsharing&sa=D&ust=1576023411916000) shows what the experiment looks like.

The hypothesis was that this might set clearer expectations for students upfront, thus reducing the number of frustrated students who left the free trial because they didn't have enough time—without significantly reducing the number of students to continue past the free trial and eventually complete the course. If this hypothesis held true, Udacity could improve the overall student experience and improve coaches' capacity to support students who are likely to complete the course.

The unit of diversion is a cookie, although if the student enrolls in the free trial, they are tracked by user-id from that point forward. The same user-id cannot enroll in the free trial twice. For users that do not enroll, their user-id is not tracked in the experiment, even if they were signed in when they visited the course overview page.

## Metric Choice

Which of the following metrics would you choose to measure for this experiment and why? For each metric you choose, indicate whether you would use it as an invariant metric or an evaluation metric. The practical significance boundary for each metric, that is, the difference that would have to be observed before that was a meaningful change for the business, is given in parentheses. All practical significance boundaries are given as absolute changes.

Any place "unique cookies" are mentioned, the uniqueness is determined by day. (That is, the same cookie visiting on different days would be counted twice.) User-ids are automatically unique since the site does not allow the same user-id to enroll twice.

* **Number of cookies:** That is, number of unique cookies to view the course overview page. (dmin=3000)
* **Number of user-ids:** That is, number of users who enroll in the free trial. (dmin=50)
* **Number of clicks:**That is, number of unique cookies to click the "Start free trial" button (which happens before the free trial screener is trigger). (dmin=240)
* **Click-through-probability:** That is, number of unique cookies to click the "Start free trial" button divided by number of unique cookies to view the course overview page. (dmin=0.01)
* **Gross conversion:**That is, number of user-ids to complete checkout and enroll in the free trial divided by number of unique cookies to click the "Start free trial" button. (dmin= 0.01)
* **Retention:**That is, number of user-ids to remain enrolled past the 14-day boundary (and thus make at least one payment) divided by number of user-ids to complete checkout. (dmin=0.01)
* **Net conversion:**That is, number of user-ids to remain enrolled past the 14-day boundary (and thus make at least one payment) divided by the number of unique cookies to click the "Start free trial" button. (dmin= 0.0075)

You should also decide now what results you will be looking for in order to launch the experiment. Would a change in any one of your evaluation metrics be sufficient? Would you want to see multiple metrics all move or not move at the same time in order to launch? This decision will inform your choices while designing the experiment.

## Measuring Variability

[This spreadsheet](https://www.google.com/url?q=https://docs.google.com/a/knowlabs.com/spreadsheets/d/1MYNUtC47Pg8hdoCjOXaHqF-thheGpUshrFA21BAJnNc/edit%23gid%3D0&sa=D&ust=1576023411919000) contains rough estimates of the baseline values for these metrics (again, these numbers have been changed from Udacity's true numbers).

For each metric you selected as an evaluation metric, estimate its standard deviation analytically. Do you expect the analytic estimates to be accurate? That is, for which metrics, if any, would you want to collect an empirical estimate of the variability if you had time?

## Sizing

### Choosing Number of Samples given Power

Using the analytic estimates of variance, how many pageviews **total**(across both groups) would you need to collect to adequately power the experiment? Use an alpha of 0.05 and a beta of 0.2. Make sure you have enough power for **each** metric.

### Choosing Duration vs. Exposure

What percentage of Udacity's traffic would you divert to this experiment (assuming there were no other experiments you wanted to run simultaneously)? Is the change risky enough that you wouldn't want to run on all traffic?

Given the percentage you chose, how long would the experiment take to run, using the analytic estimates of variance? If the answer is longer than a few weeks, then this is unreasonably long, and you should reconsider an earlier decision.

## Analysis

The data for you to analyze is [here](https://www.google.com/url?q=https://docs.google.com/a/knowlabs.com/spreadsheets/d/1Mu5u9GrybDdska-ljPXyBjTpdZIUev_6i7t4LRDfXM8/edit%23gid%3D0&sa=D&ust=1576023411921000). This data contains the raw information needed to compute the above metrics, broken down day by day. Note that there are two sheets within the spreadsheet - one for the experiment group, and one for the control group.

The meaning of each column is:

* **Pageviews:** Number of unique cookies to view the course overview page that day.
* **Clicks:**Number of unique cookies to click the course overview page that day.
* **Enrollments:**Number of user-ids to enroll in the free trial that day.
* **Payments:** Number of user-ids who who enrolled on that day to remain enrolled for 14 days and thus make a payment. (Note that the date for this column is the start date, that is, the date of enrollment, rather than the date of the payment. The payment happened 14 days later. Because of this, the enrollments and payments are tracked for 14 fewer days than the other columns.)

### Sanity Checks

Start by checking whether your invariant metrics are equivalent between the two groups. If the invariant metric is a simple count that should be randomly split between the 2 groups, you can use a binomial test as demonstrated in Lesson 5. Otherwise, you will need to construct a confidence interval for a difference in proportions using a similar strategy as in Lesson 1, then check whether the difference between group values falls within that confidence level.

If your sanity checks fail, look at the day by day data and see if you can offer any insight into what is causing the problem.

### Check for Practical and Statistical Significance

Next, for your evaluation metrics, calculate a confidence interval for the difference between the experiment and control groups, and check whether each metric is statistically and/or practically significance. A metric is statistically significant if the confidence interval does not include 0 (that is, you can be confident there was a change), and it is practically significant if the confidence interval does not include the practical significance boundary (that is, you can be confident there is a change that matters to the business.)

If you have chosen multiple evaluation metrics, you will need to decide whether to use the Bonferroni correction. When deciding, keep in mind the results you are looking for in order to launch the experiment. Will the fact that you have multiple metrics make those results more likely to occur by chance than the alpha level of 0.05?

### Run Sign Tests

For each evaluation metric, do a sign test using the day-by-day breakdown. If the sign test does not agree with the confidence interval for the difference, see if you can figure out why.

### Make a Recommendation

Finally, make a recommendation. Would you launch this experiment, not launch it, dig deeper, run a follow-up experiment, or is it a judgment call? If you would dig deeper, explain what area you would investigate. If you would run follow-up experiments, briefIy describe that experiment. If it is a judgment call, explain what factors would be relevant to the decision.

## Follow-Up Experiment: How to Reduce Early Cancellations

If you wanted to reduce the number of frustrated students who cancel early in the course, what experiment would you try? Give a brief description of the change you would make, what your hypothesis would be about the effect of the change, what metrics you would want to measure, and what unit of diversion you would use. Include an explanation of each of your choices.

## Congratulations!

You've finished the quantitative parts of the final project. Don't forget to think about the answers to the qualitative questions as well. You can use [this template](https://docs.google.com/document/d/16OX2KDSHI9mSCriyGIATpRGscIW2JmByMd0ITqKYvNg/edit?usp=sharing) to answer the questions in the [project instructions](https://docs.google.com/document/u/1/d/1aCquhIqsUApgsxQ8-SQBAigFDcfWVVohLEXcV6jWbdI/pub?embedded=True) and the [project rubric](https://docs.google.com/document/u/1/d/1Hga00A4258wSJ9dir0Td_ZPLeAtJ079UGbM3C0ykUtE/pub?embedded=true) to evaluate your work.