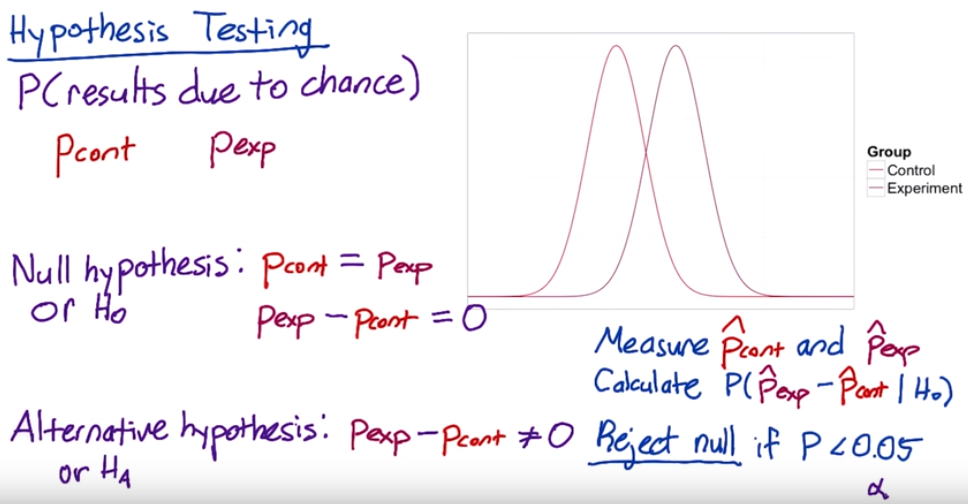
Lesson 1:

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:

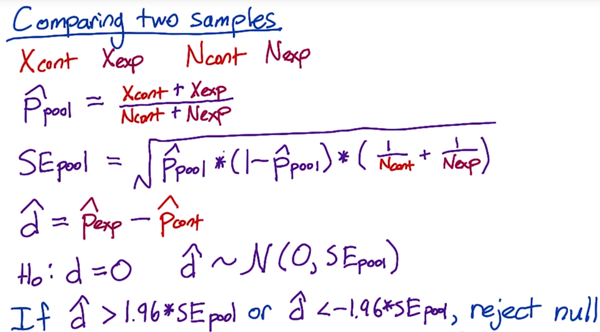
1. A statistically **significance** positive result
2. A statistically **significance** negative result
3. No statistically **significance** difference.

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

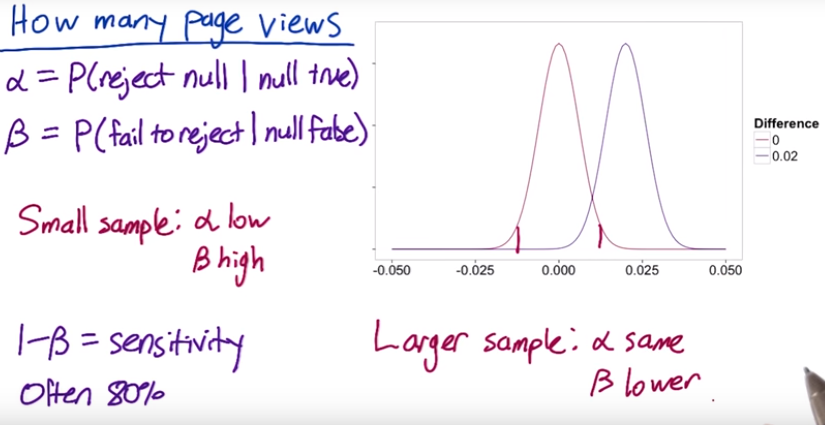
Which one you should use depends on what action you will take based on the results. If you're going to launch the experiment for a statistically **significance** positive change, and otherwise not, then you don't need to distinguish between a negative result and no result, so a one-tailed test is good 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%

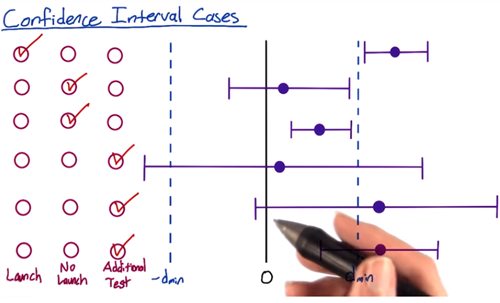
pool = = 0.111

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:

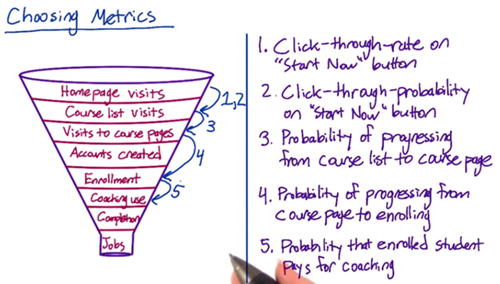
1. What risk is company being exposed to?
2. What benefit might be the outcome of the test?
3. What other choices do we have?
4. What expectation of privacy do they have?

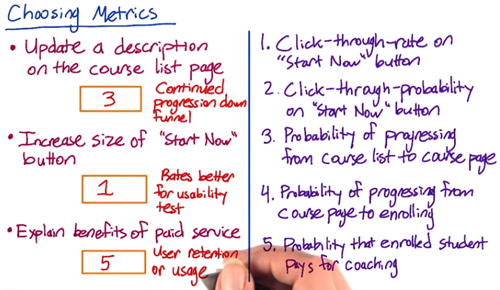
Lesson 3

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.



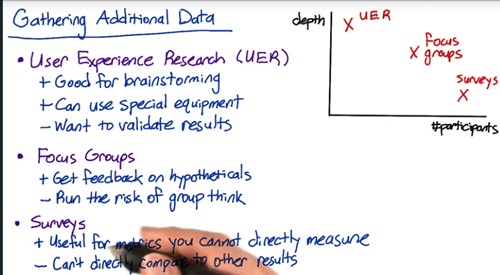


how many months were they active on the site? You want to get some baseline for that. And then given that they took a second course, then you might also want to trigger survey that happened within your site. Get a survey that says “ are you considering…

**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



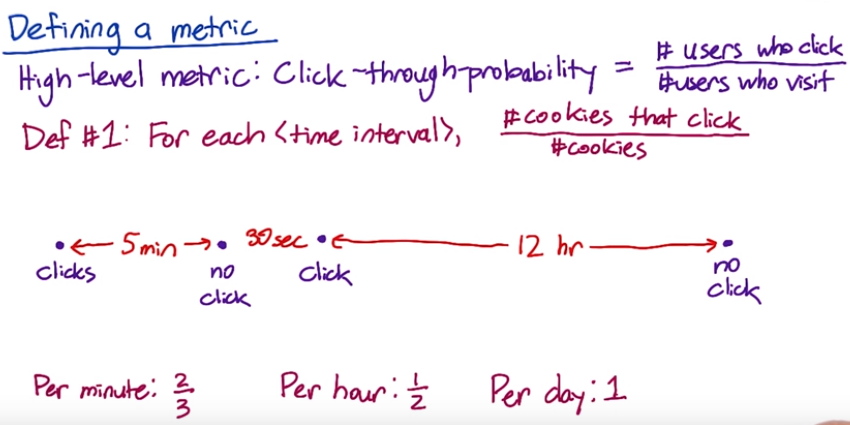
**Video 8: Other Techniques: Example**

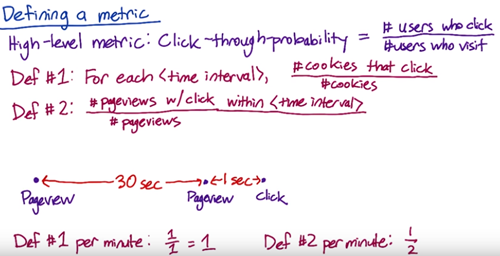
Buy external data from Comscore, hitwise, Nielsen

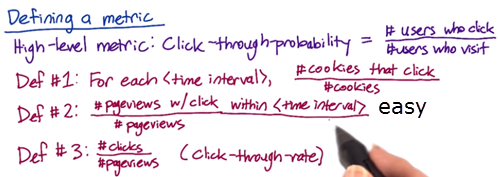
For example, watching the customers try to complete the checkout can help you figure out

* 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

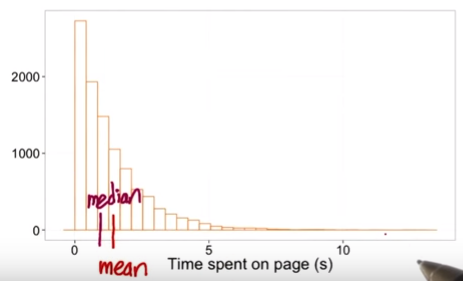
**Second one** is to characterize what the distribution of metric looks like. e.g. ideal way of doing this is to do a retrospective analysis, to compute a histogram. E.g. you’re going to have all the different value for the load time on the x axis, the y axis is going to be the frequency, so how often 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.

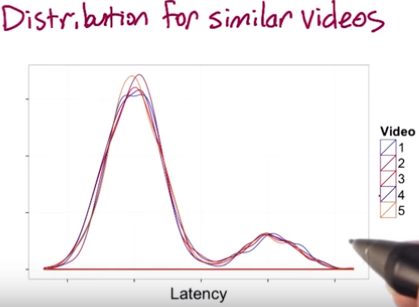
The key here is not to necessarily come up with a distribution to match if the answer isn’t clear - that can be helpful - but to choose summary statistics that make the most sense for what you do have. If you have a distribution that is lopsided with a very long tail, choosing the mean probably doesn’t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

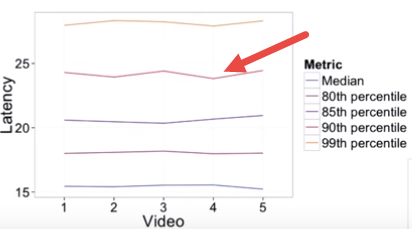
How? Way 1. Do the retrospective analysis.

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.

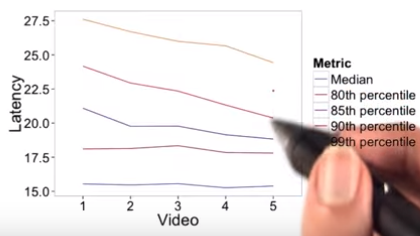


How? Way 2. Look at the actual experiment

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.

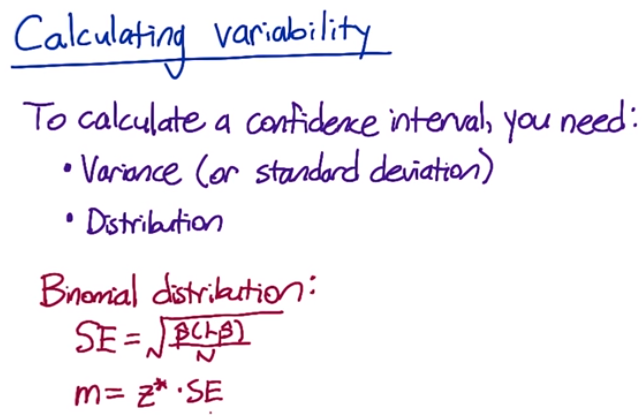
Video 20

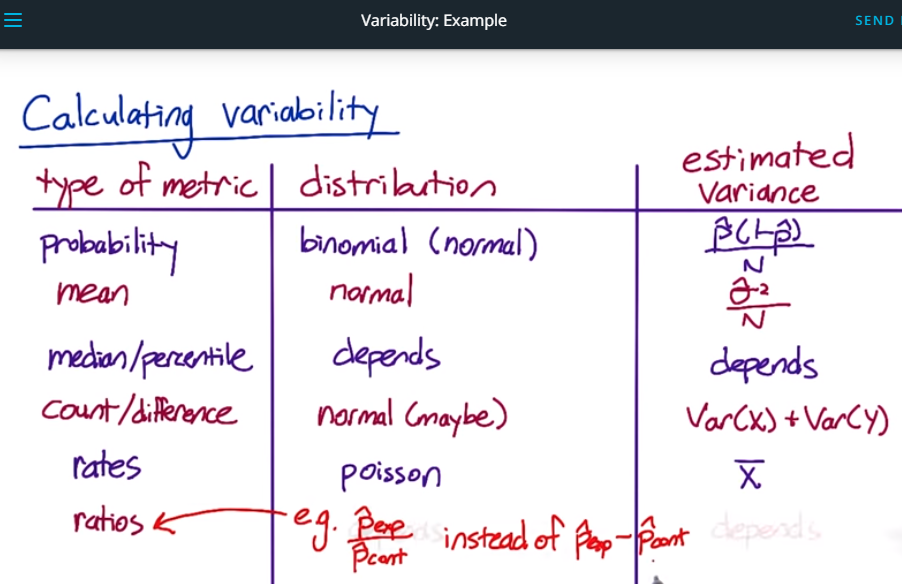
How to compare two group’s difference?

**Relative differences in probabilities**

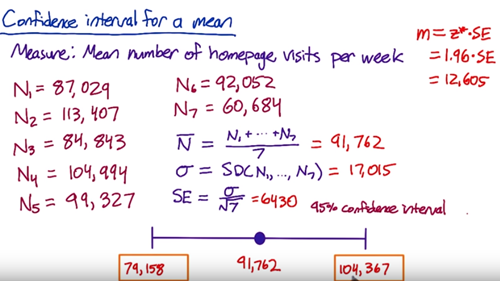
For probability metrics, people often use percentage points to refer to absolute differences and percentages to refer to relative differences. For example, if your control click-through-probability were 5%, and your experiment click-through-probability were 7%, the absolute difference would be 2 percentage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

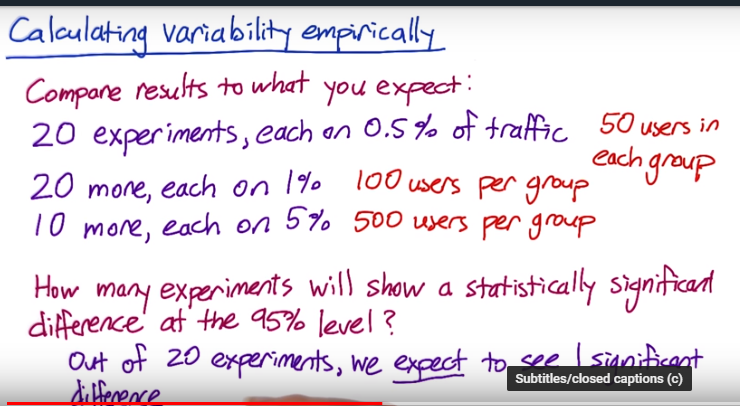
The underlying variability

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

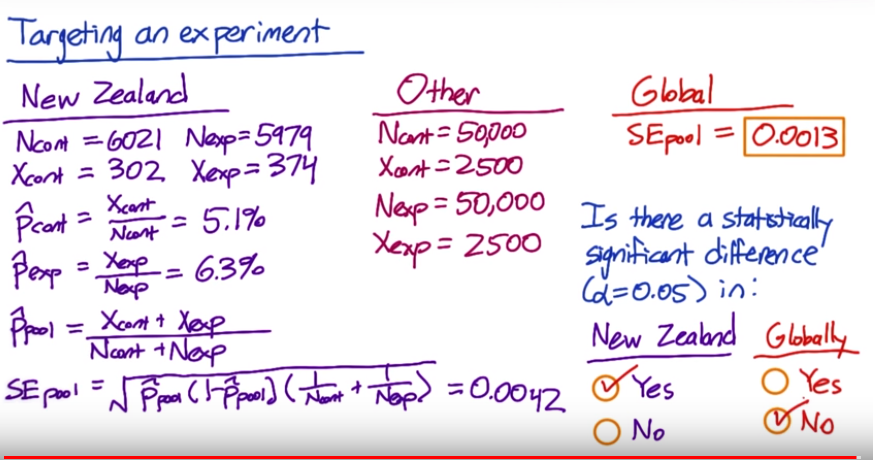
1. choose population, only US?

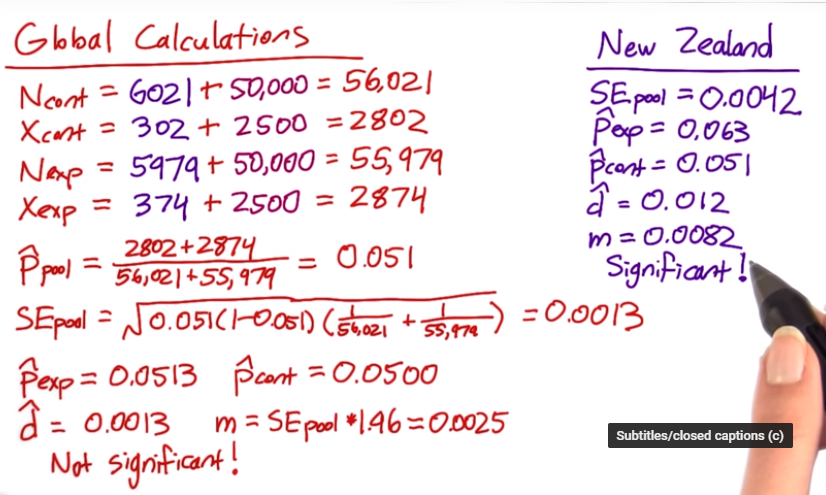
Unit of Diversion:

* 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.

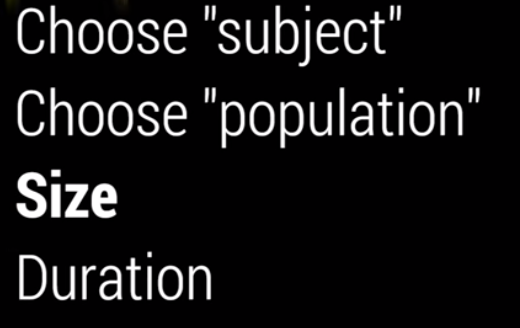
Video 12 Target population





Video 13 population vs cohort

Track increased usage of a website, you want a cohort



Video 17 Sizing

