# Lesson 3 Choosing and Characterizing Metrics

## What can we do with our own internal logging data or data captured from website to choose the right metrics?

* Retrospective or observational analysis, which is good for establishing a correlation but not a causation.
* Running experiments to find the exact metric for evaluating experiments.

Example:

How to measure if students are happy about our website?

* Step 1: go back to our log. Given that somebody took a second course, which we’ll say is being happy, what did they do? How long did they spend? How many months were they active on the site? This will be our baseline.
* Step 2: Given that they took a second course, then we might also want to trigger surveys that happened within your site or do a focus group, where people who’ve done a single course get a survey about ‘will you take a second course?’ or ‘were you happy with your first course?’

## Techniques for measuring metrics:

* External data
* UER
* Focus group
* Survey
* Retrospective analysis
* Experiments

## Techniques to Gather Additional Data

* User Experience Research (UER): in-depth study for limited participants

+: good for brainstorming, can use special equipment

-: want to validate results, such as retrospective analysis

* Focus Groups: less deep study for broader participants.

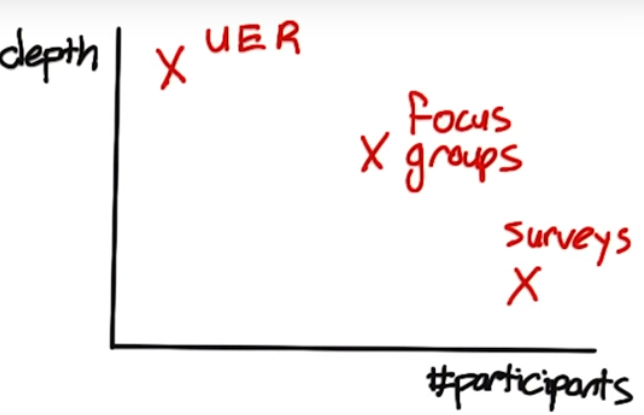
+: get feedback on hypothetical questions.

-: run the risk of group think and convergence on fewer opinions

* Survey: cheap to run on a large sample size, but not very deep or individually customized.

+: useful for metrics you cannot directly measure

-: can’t directly compare to other observational methods, because survey results are not 100% reliable and response depends on question design. (population not comparable)



## Some examples for applying different techniques:

* Rate of returning for 2nd course:
  + Survey: suitable for info that is hard to measure, but only for proxy
* Average happiness of shoppers
  + Survey: happiness is not long-term
  + UER
* Whether users can find useful information via search:
* External data
* UER
* Human grader

Possible proxies:

* Time spent
* Clicks on results
* Follow-up queries
* Measure user engagement

Metrics to use: course completion (but it’s long-term)

Techniques:

* Survey
* UER: maybe discover other short-term measurable metrics, such as the length of time spent on the page or clicking more links to extra material.
* Retrospective analysis: since course completion rate is a long-term metric, we can find data for users that had completed courses and check if there was something in common.
* Decide whether to extend inventory for a shopping site

Techniques:

* Focus group: get ideas from users
* External data: compare to other shopping sites, indicating what users have wanted on those sites.
* Which ads get most views

Techniques:

* External data: maybe beneficial for finding other proxies, such as mouse hover events or time spent on the page that we can use for our case
* UER: see which ads the user is viewing by an eye-tracking camera, and then try to find a metric that correlates with that.

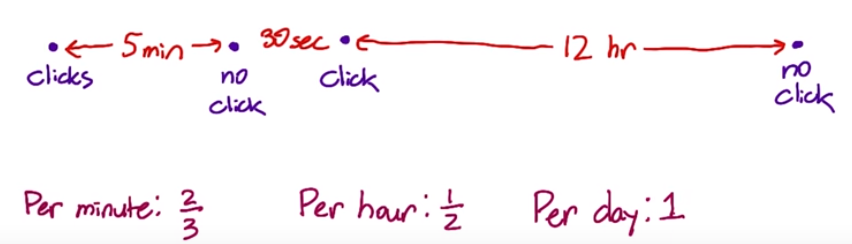
## Metric Definition

How to turn a high-level metric, such as the click-through-probability to a well-defined metric?

High-level metric: click-through-probability =

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

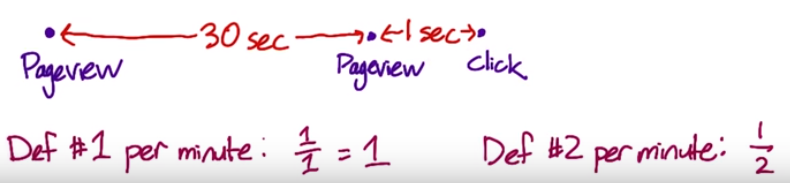
Depending on the time interval, we may have different click-through-probabilities for one cookie:



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

This method removes the idea of a unique user and instead creates a unique ID for each page view. When a user clicks, record the idea of the corresponding parent view. The definition is:

The <time interval> is the time we want to wait after each pageview to see if the user clicks.



* Def #3 (Rate): Number of clicks divided by number of pageviews. This is more like a click-through-rate instead of a click-through-probability (no consideration for double-click or reloading the page).

For same short <time interval>, def #1 and def #2 are almost indistinguishable. Since #2 is easier to calculate, we might want to go with this one.

Which metrics have which problem?

* Double click:
* Cookie prob: count cookies that clicked, so robust to this problem
* Pageview prob: count pageviews that clicked, so robust
* Rate: depends on the # of clicks, so not robust
* Back button caches page without generating a new pageview
* Cookie prob: robust to pageview
* Pageview prob: depends on # of pageviews, so not robust
* Rate: depends on # of pageviews, so not good
* Click-tracking bug: such as JavaScript fails to log some clicks
* If a click is completely missed, all three definitions will be zero.

## Filtering and segmenting

Segmenting and filtering data is good for evaluating definitions and building intuitions for our data.

External reasons:

* We would like to filter frauds and spams, or
* competitor clicks on everywhere of our website
* due to blog coverage for a big company’s website change, we could potentially get a lot of traffic

Internal reasons:

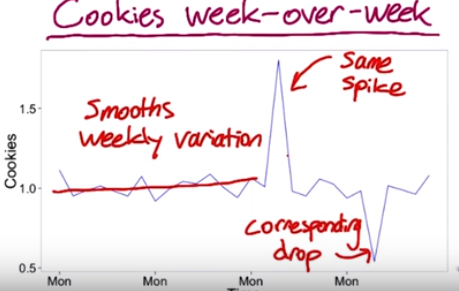
* the change only impacts partial of our products, such as changes only made to English traffic or mobile app without the web version. Therefore, we need to filtering the traffic.

## How to detect an anomaly?

Here is the time series for total active cookies over time,



We can see there is one huge spike and traffic is normally higher during the weekend. If we divide each data point by the point from one week ago, (cookies week-over-week)



The same spike still exists and one drop occurs a week later after the spike (divided by a huge value).

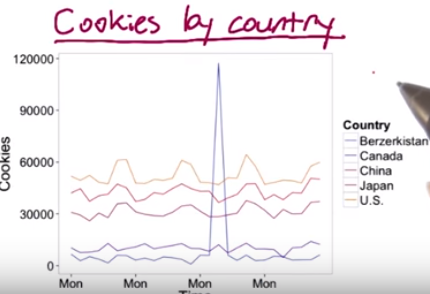
What about year-over-year data (divided by the data a year ago)? If there is an annual event, the spike should go away, but not in this case:



Weekend spikes exist because the data is not quite matched by the day of the week.

## How to identify the reason behind the anomaly?

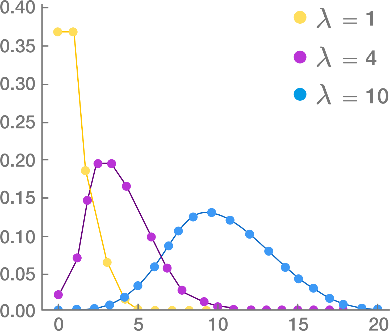
First we can look at different segments across our population and to see if one segment is causing the spike:



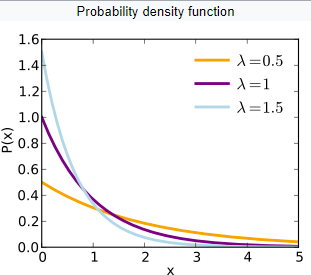
The spike happens only in Berzerkistan. At this moment, we should talk to engineering teams and see if it’s caused by rogue IP addresses, maybe spams or robots.

## Common distributions in online data

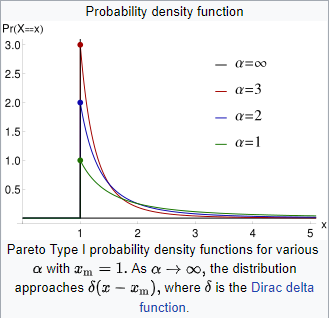
* Poisson distribution, such as the average staytime on the result page before traveling to a result.



* Exponential distribution:



* Pareto distribution, which will be heavy-tailed distribution, such as the frequency of words in a text (the most common word is eally really common compared to the next word in the list.)



One thing to notice is that overall data can be 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.

## Categories of summary metrics

* Sums and counts

e.g. # of users who visited page

* Means, medians and percentiles

e.g. mean age of users who completed a course or median latency of page load

* Probabilities and rates

e.g. probability has 0 or 1 outcome in each case or rate has 0 or more

* Ratios

Business metrics often makes sense with ratios, e.g.

## Sensitivity and Robustness

A good metric should be sensitive to the changes that we make, and robust enough for the changes that we don’t make. For example, depending on what we are trying to measure, median/mean/90th percentile/95th percentile will be selected.

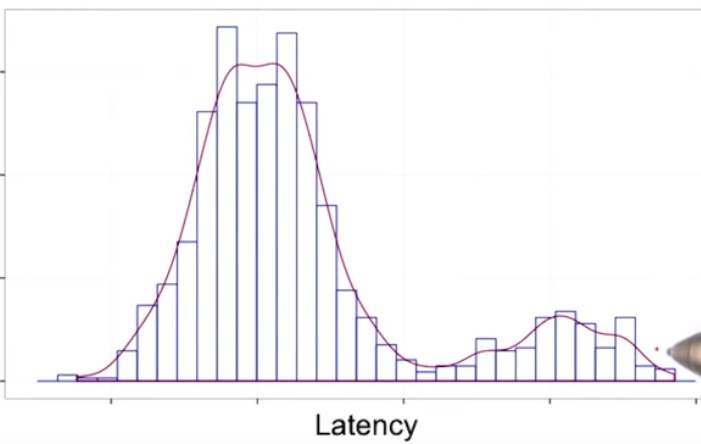
How to measure sensitivity or robustness?

* Look at experiment data. A/A test is a good to way to test metric sensitivity. We can see if our metrics pick up any spurious differences between the two groups. Historical experiment data is also valuable for measuring metric variance.
* Retrospective analysis. If we don’t have historical experiment data or don’t plan to run an experiment, we can look back at changes we know we made to our site and see if the metrics we’re interested in actually moved in conjunction with those changes.

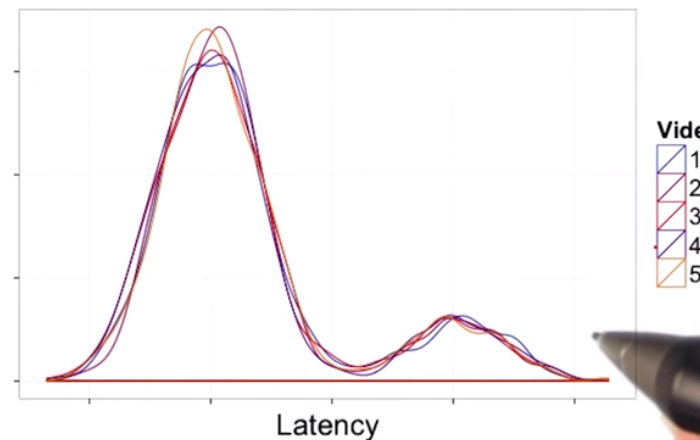
For example, we would like to choose summary metric for latency of a video (how long it takes to load the video).

* Retrospective analysis first:

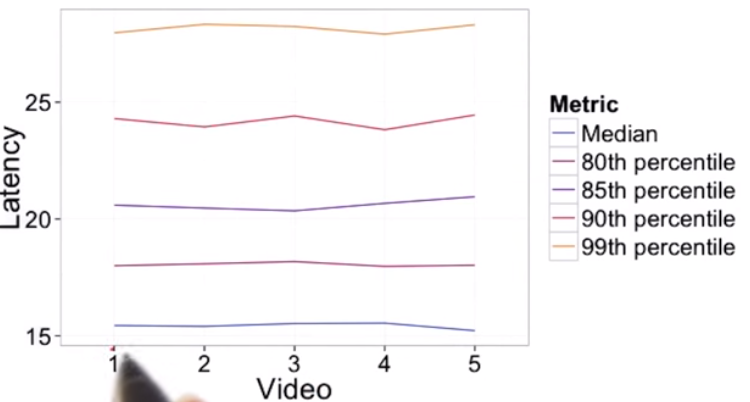
Assume we have the distribution for a single video as follows:



If we plot for multiple videos of the same size:



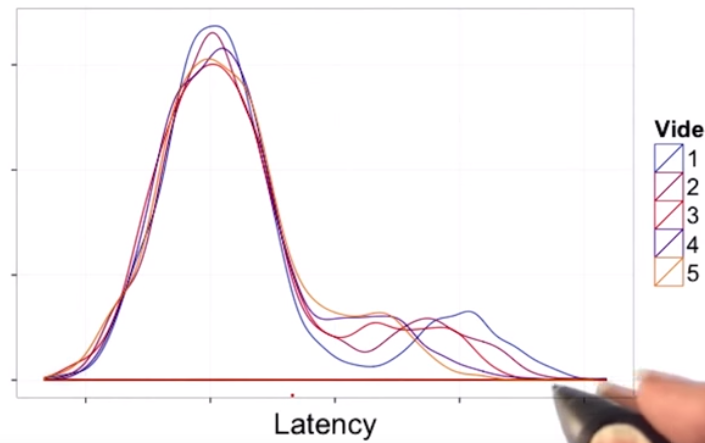
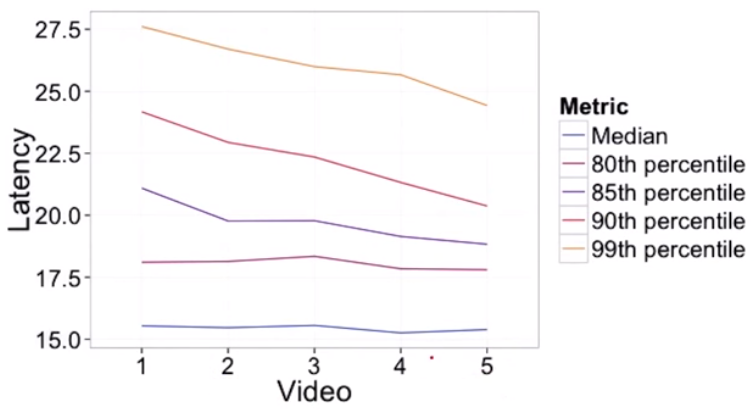
There are two modals here, people with long loading time and short loading time, which could happen because people have different speed internet. If we plot the medians and certain percentiles for these 5 videos:



Compared to median, 80th and 85th percentile, 90th and 99th percentile zig zag more, which means they are not robust enough as summary metrics.

* Experiment data:

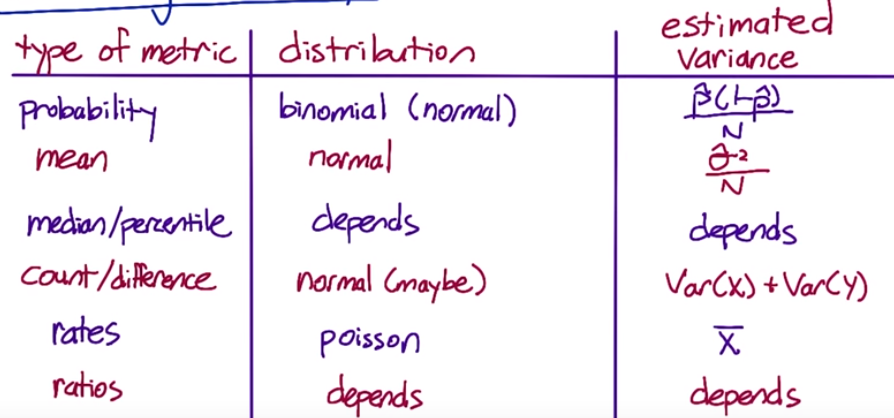
If we change the resolution of our videos, the load time should increase. If our metric doesn’t change, it probably won’t be a sensitive enough summary metric.

Video 1 has the highest loading time. The latency affects people with slow internet more than with the fast internet. Median and 80th percentile barely changes across different types of videos, which is an indication that median and 80th percentile are not sensitive enough. They are not showing a change when we do make a change that we care about.

**Overall, 85th percentile might be a good metric that’s both robust and sensitive.**

## Calculating Variability



## Empirical variability

Sometimes we have no knowledge of the underlying distribution for our metrics, or the underlying distributions are not typical, we can either run a large A/A test or bootstrap on small samples to decide what the empirical variability for our metric.

**Why wouldn’t we just use bootstrap instead of running large A/A test?**

If the experiment system is complicated, A/A test is a good for testing on the system. It helps to answer questions such as ‘is our randomization function truly random’ or ‘do we have any other issues with regards to bias or weird population effects?’ The difference we’re observing is due to the variability of underlying system, such as the user population or user behaviors.

## What’s the uses of A/A tests?

* Compare results to what we expect (sanity check)

If there is discrepancy, maybe we made wrong assumptions about our data distribution.

* Estimate variance empirically and then use our assumptions about the distribution to calculate the confidence interval.
* If we do not want to make any assumptions about our data, we can directly estimate a confidence interval from the results of the A/A tests.

For example, if we want to look at A/A tests on click-through-probability and compare results to what we expect.

* 20 experiments, each on 0.5% of the traffic. 50 users in each group.
* 20 more, each on 1%, which has 100 users per group.
* 10 more, each on 5%, which has 500 users per group.

**So how many experiments will show a statistically significant difference at the 95% level?**

95% confidence level means that out of 20 experiments, we expect to see 1 significant difference.

Analysis data:

<https://docs.google.com/spreadsheets/d/17wWNY2jkDlG9BDMYQq2l-ku_8HGajXuF2Zvy__dBEL4/edit#gid=0>

The distributions get tighter as we increase sample size, which means the variance is decreasing with larger sample size.

## Estimate variance and calculate confidence interval

1. Calculate the difference between two groups for each pair.
2. Calculate the standard deviation of the differences.
3. If we expect our metric follow a normal distribution,

The margin of error is computed as the (**empirical**) standard deviation times z score of our confidence level. (95% CI -> z = 1.96, 90% -> 1.65)

The CI will be,

We can check the histogram of our metric and see if it follows a normal distribution.

**Analytically**,

Therefore, we have slightly different margin of error for each experiment. For empirical situation, we only calculate one margin of error across all the experiments.

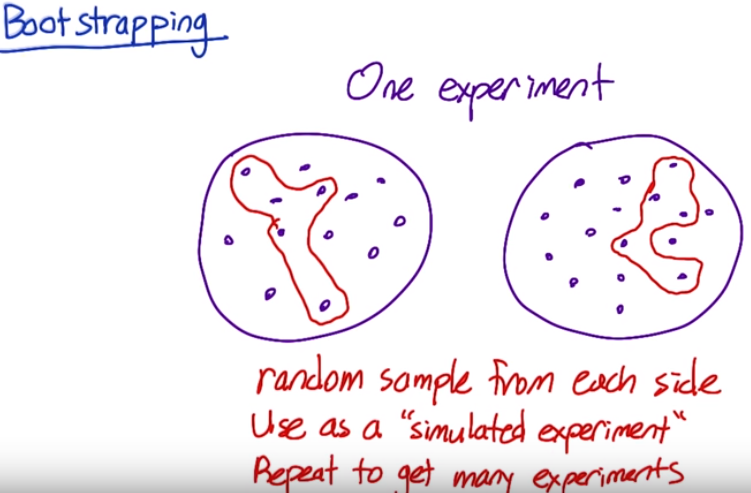
1. If no assumptions on the underlying distribution, we will directly estimate confidence interval. After running an A/A test, we sort all the differences, and select a box that includes only 95% of the values, which discards 2.5% data on each side.

For example, if there are 40 data points, we sort the differences, discard the maximum and minimum values, the rest will compose our 95% confidence interval. (40 \* 0.025 = 1)

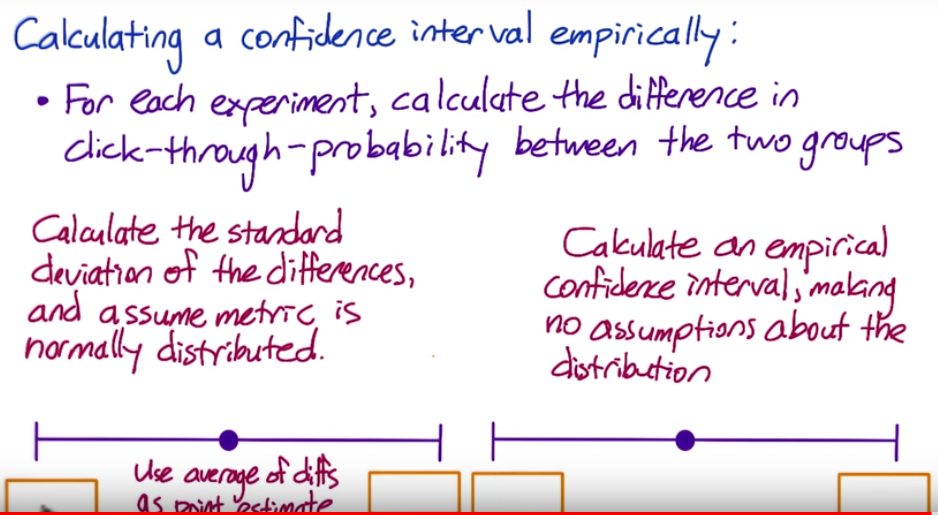
## Estimate variance by bootstrap

We only run 1 A/A test, but there are dozens of data points in each group.

* Random sample from each group and calculate the click-through-probability as it was a full experiment.
* Compute the difference in click-through-probabilities and use that as a simulated experiment.
* Repeat the above process over and over, and record the results.



* Draw the confidence interval based on bootstraps results.



## Variability summary

* Some metrics with high variability might not be suitable for experiment in practical purpose, even if the metric makes a lot of business or product sense.
* In order to compute the variability, we need to understand the distribution of the underlying data. There are both analytical and empirical techniques for computing variability.
* Normally analysts spend more time on validating and choosing metrics as opposed to evaluating the experiments themselves.