- Guideline: Define ; Build Intuition ; characterize

2. metrics Define

1. Overview: Before we define a single metric

① Some data are not tractable or take

② Need to define them accurately

Active user → How to define 'active' → check activity should take into account

③ Need to think all measurements and then summarize them into a single metric

④ One metric V.S multiple metrics

- One

(1) Every team towards same goal ; (2) good for PR and external report ;

(3) Can create a single composite metric (use objective function)

- Multiple

(1) Can know how other thing move and find the reason for the movment of important metrics, or composite metrics

(2) Won't over optimize one thing when don't look at how other thing move

⑤ Consider how generally applicable the metric is

If run a whole suite of AB tests, ideally you'd have one or more metrics that can use across the entire suite. It's much better to use a metric that's less optimal for you test if you can use it across the suites that you can do comparison.

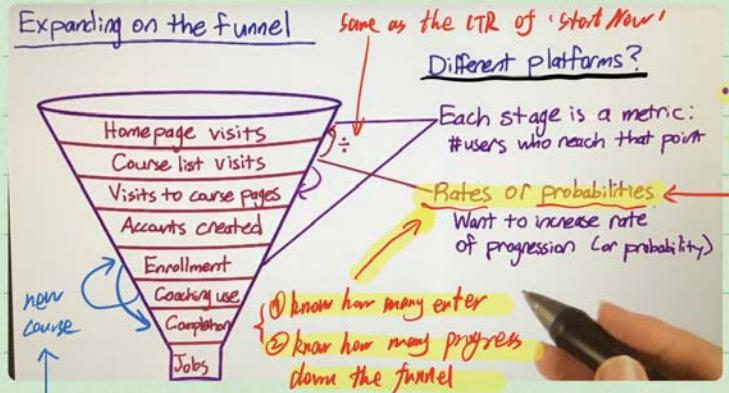
Custom metrics will increase the risk.

2. High level concept for metrics

Business Objective: { Helping students to find jobs
Financial sustainability

3. Expanding on the funnel

- Homepage visits
- Exploring the site
  - # users who view course list
  - # users who view course details
- Create an account
  - # users who enroll in a course
  - # users who finish Lesson 1, Lesson 2, etc
  - # users who sign up for coaching at various levels
- Completing a course
  - # users who enroll in a second class
  - # users who get jobs



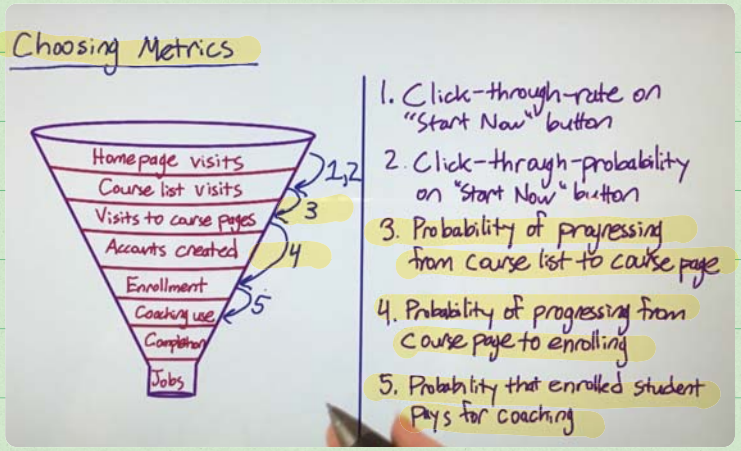
(count)
- only use number for a few key points

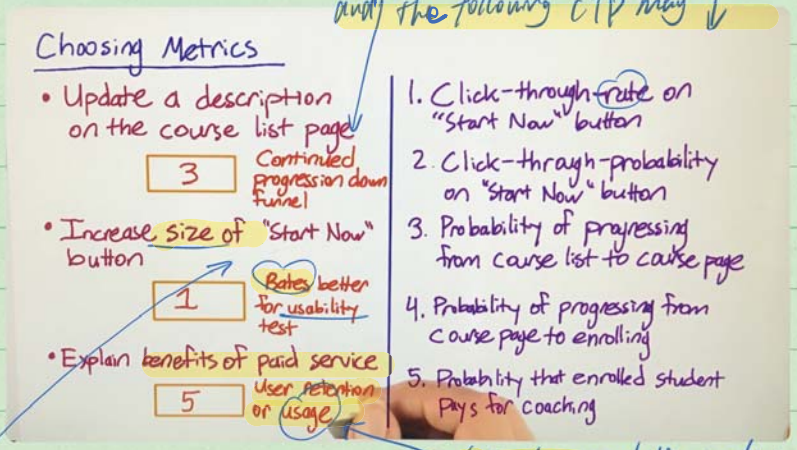
prob. that unique users progress down the funnel (Binary case)

Time line for a particular user is more like a swirl (might jump forward and back)

For viewing a list of courses { rate: how easy it is to do it from the Course Overview Page
prob: whether a student reach the course list at all (binary probability)

4. Choosing matrix





• If description is misleading, 3 ↑ and the following CTP may ↓

Size related to how easy to be found

for whom used this coaching, do they use more often now that it's more clear that coaches can do

∴ it's of usability ⇒ rates is better for it.

5. Difficult Metrics

① { Don't have access to data
Take too long

② Hard metrics example

- Rate of returning for 2nd course : Has data, but too long
- Average happiness of shoppers : Doesn't have data
- Probability of finding information via search : Doesn't have data

③ Techniques for approx of hard metrics

survey. retrospective analyses, focus groups :

Brainstorm new metrics and validating possible metrics

(a) external data

eg: { company collect market data such as market share, --
company run survey on users
Academic research

(b) internal data get baseline/theory based on past metrics and related changes

existing data: retrospective analyses, running experiments

gather new data: user experiment research, survey, focus group

problem:

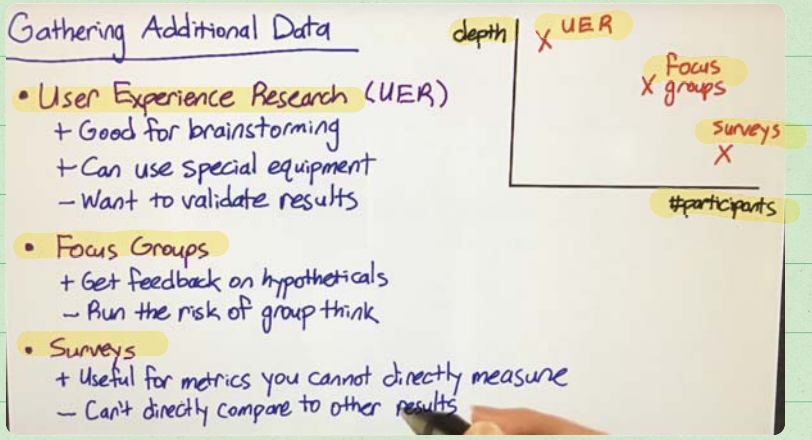
show correlation but not causation (actual controled experiment)

∴ The only way to validate is run experiment

④ • Run an experiment to determine whether or not we can use a metric for evaluating experiments

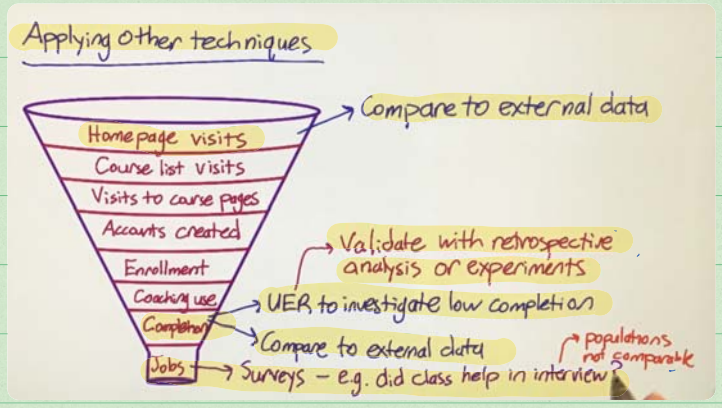
(Want to make sure the metrics will move as we make changes)

6. Gathering Additional data

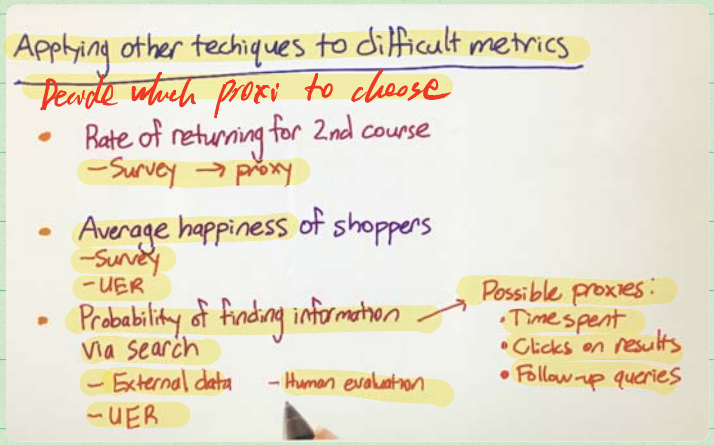


7. Use other techniques

①



②



③ Which techniques would you use?

- Measure user engagement
  Course completion too long-term
  [4] [2] 5
- Decide whether to extend inventory
  [3] [1]
- Which ads get most views
  [1] [2] 5 ↖ Use historical data to predict or set baseline

1. External data
2. UER
3. Focus group
4. Survey
5. Retrospective analysis
6. Experiments

II. Data capture and build intuition

1. Data capture

Note: CTR may different based on different browsers.

It's not because the actual CTR different, but because the technology you use in order to gather the clicks (different failure rate for JAVA commands)

2. Define metrics : from high level to detailed

① Need to decide details like time interval

Defining a metric

High-level metric: Click-through-probability = $\frac{\#\text{users who click}}{\#\text{users who visit}}$

Def #1: For each ⟨time interval⟩, $\frac{\#\text{cookies that click}}{\#\text{cookies}}$

Def #2: $\frac{\#\text{pageviews with click within ⟨time interval⟩}}{\#\text{pageviews}}$

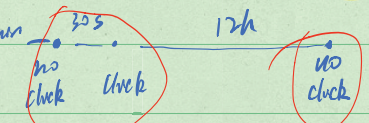
Def #3: $\frac{\#\text{clicks}}{\#\text{pageviews}}$ (click-through rate)

pageview — 30s — pageview — 1s — click

Def 1: per minute = 1 Def 2: per minute = $\frac{1}{2}$

Def 2 is easier to compute

• A user with same cookie

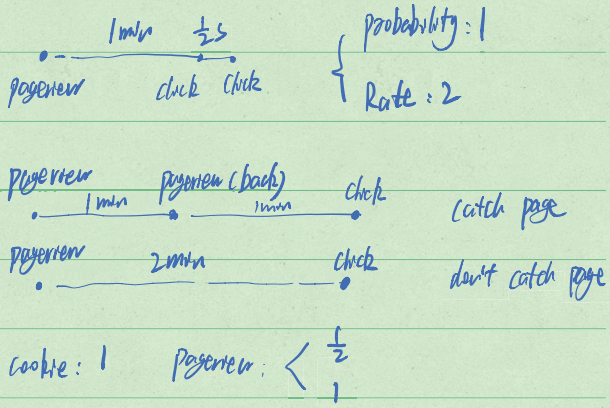


per minute = $\frac{2}{3}$, per hour = $\frac{1}{2}$, per day = 1

② Problem for metrics

Which metrics have which problems?

	1: Cookie prob	2: Pageview Prob	3: Rate
Double click	☑	☑	☐
Back button caches page	☑	☐	☐
Click-tracking bug	☐	☐	☐



↑ Just record pageview but don't record click

cookie: 1 Pageview: $< \frac{1}{2}$ 1

3. Filtering and Segmenting

Eg: Just want to make change for English users

Filter make sense because you don't want to dilute results.

Filter to affected traffic, can increase power and sensitivity of experiment

① Goal: de-bias data

when filter out spam and fraud, don't introduce bias into your data

Eg: (a) metrics that can only be measured on logged-in user:

don't have data for non-logged users → biased

(b) Filtering out long or weird session of user behavior

Before do that, need to check and make sure it's not actually your website, metrics, or even your logging that's causing these sessions to come up.

② How to check: compute baseline value for metric

- slice data: compute metric on a bunch of disjoint sets. eg: country, language
- compute a metric on all the different slices:

  show if we move traffic disproportionately: indication of bias

- look at week over week or day over day traffic pattern changes: useful to spot something unusual.

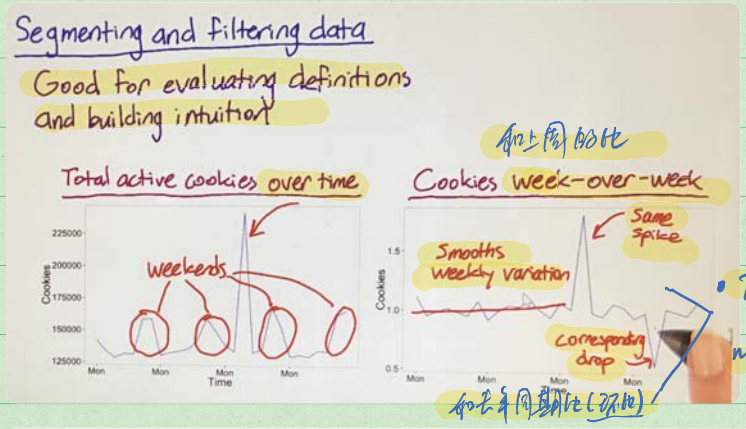
- key for testing: building intuition

Have to know what changes are going to be expected v.s. unexpected

∴ When see data for real experiment: can know if have a problem

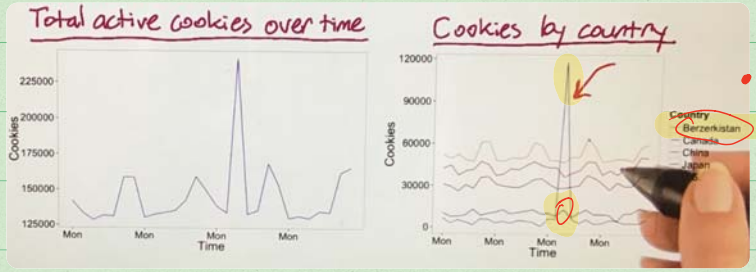
Do I believe this, what's really going on.

③ check if it's seasonality effect



- Is not caused by weekly or yearly variation

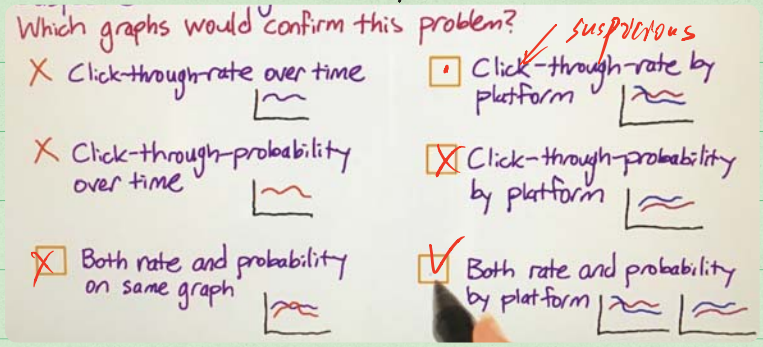
④ Dig into what cause of spike: look at different segments

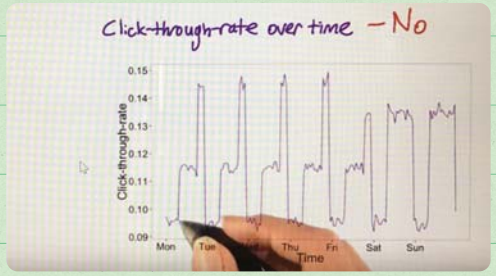


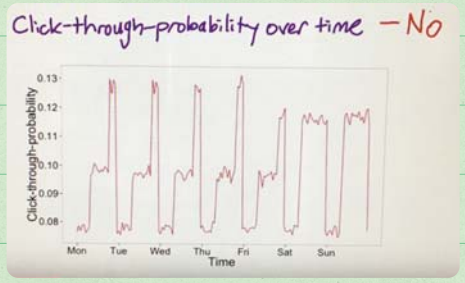
- Need to talk to engineering team

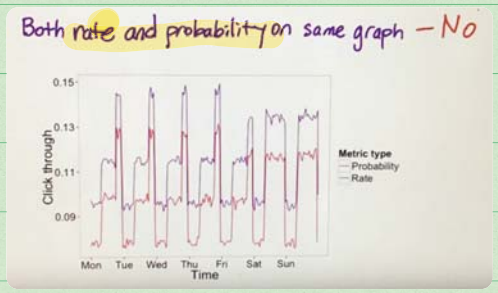
⑤ Detect suspect click tracking issue: clicks counted twice on mobile

which graphs would confirm this problem.

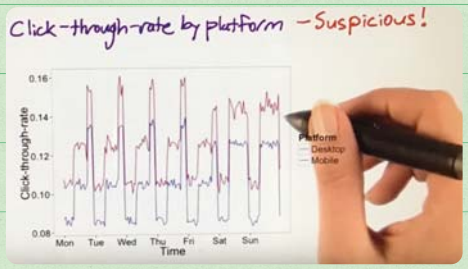




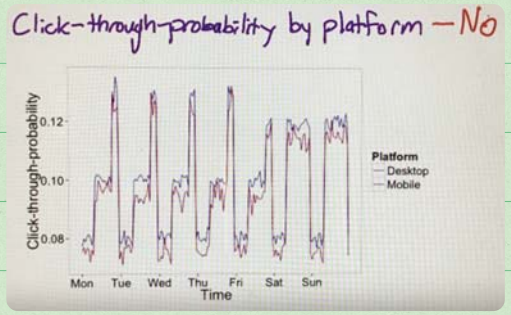




Don't know how much higher vs normal, and the difference vs consistent.



May have different behavior

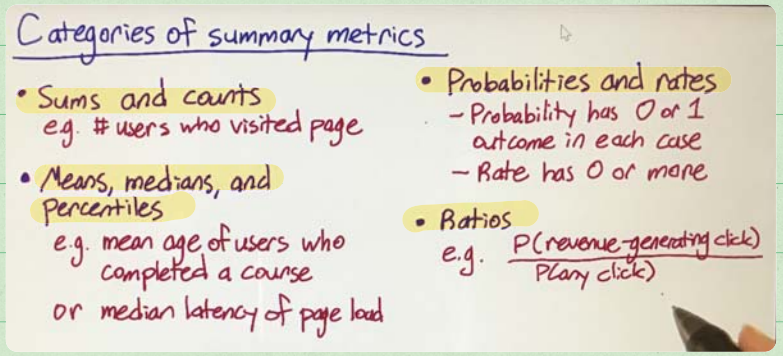


prob. eliminate effect of count twice

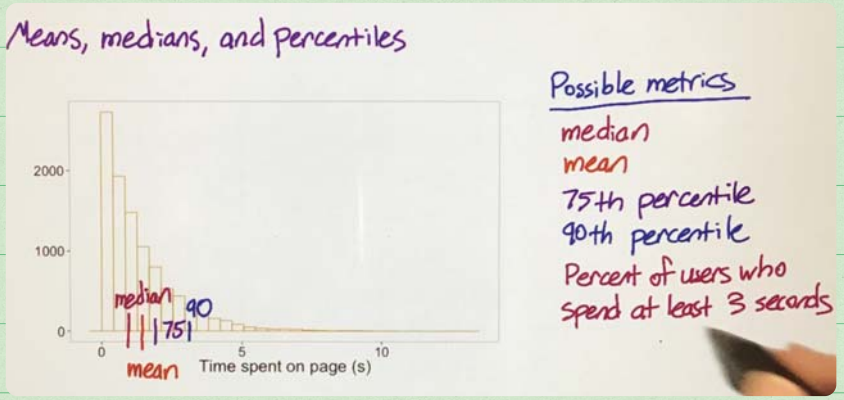
P: same but R: much higher

4. Summary Metrics

① Use metrics to summarize individual events



② Choose metric based on distribution



5. Sensitivity and robustness of metric (sensitivity)

① Idea: Choose a metric that picks up changes you care about but doesn't move a lot when nothing interesting happened (robustness)

eg: median is more robust than mean

② Measure sensitivity and robustness

(a) run experiments / using experiments you already have

Data should move in a way that intuitively make sense

- eg: increase quality will leads to increase in load time:

we could see if the metrics actually response to that

(b) A v.s. A experiment:

to determine if they are too sensitive

Don't change anything, just compare people who saw the same thing to each other. See if metrics pick up any spurious difference between the two.

(Make sure not to call things significant that may not really meaningful)

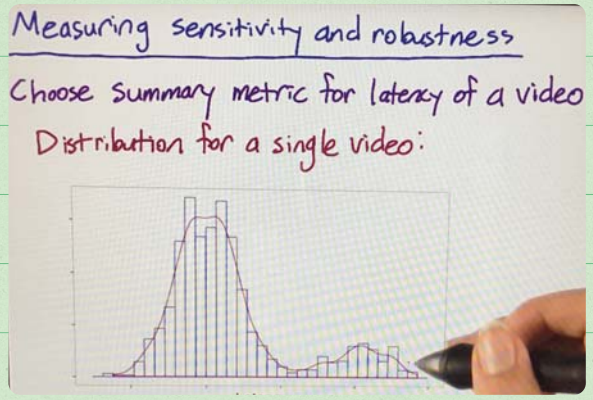
(c) Look back to experiments run earlier.

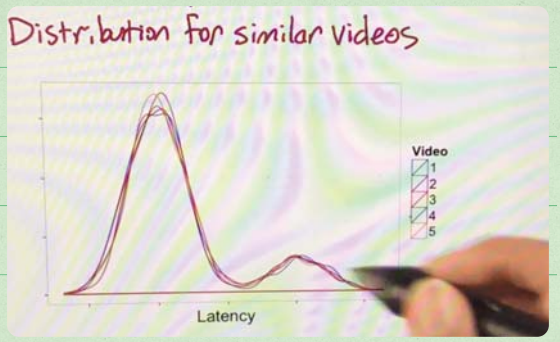
(how the metric change in past experiments)

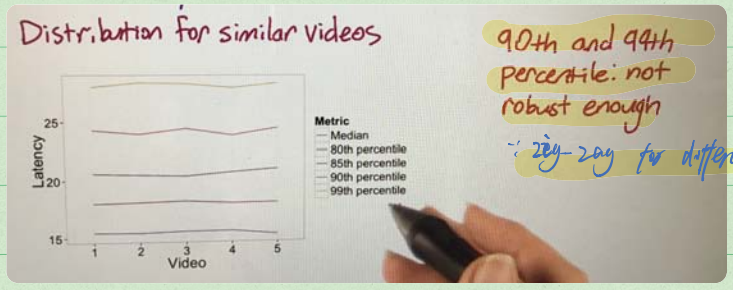
(d) Retrospective analysis

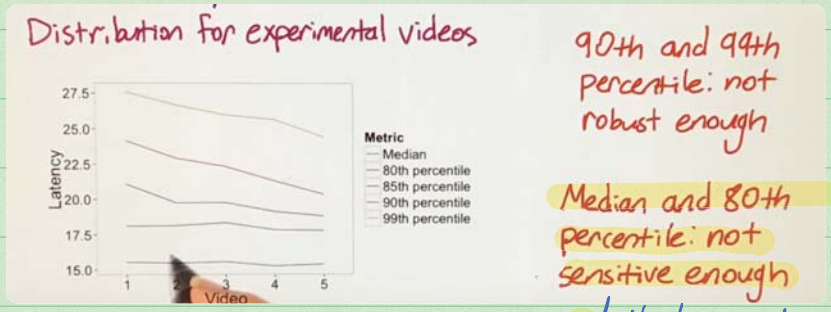
- look back at changes and see if metrics you're interested in actually moved in conjunction with those changes
- look at history of metrics to see if you can find a cause for any major changes that you see.

③ Measure sensitivity and robustness









6. Absolute / relative difference

start with absolute and then use relative (better for comparison)

III. Characterize

1. variability { use distribution / compute empirically

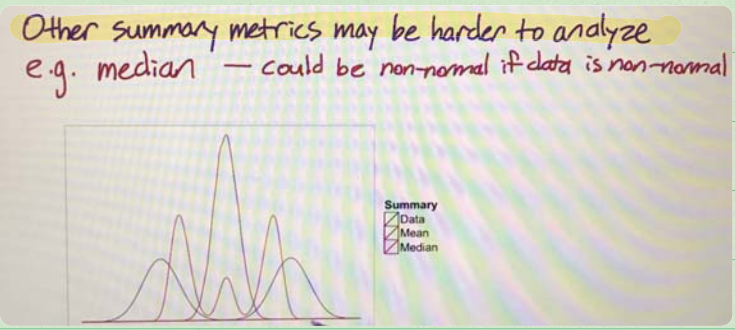
2. Calculating variability

① parametric way: get a nice distribution of data

To calculate CI, you need { variance / distribution

eg: For binomial: $SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, $m = z^* \cdot SE$

Other summary metrics may be harder to analyze
e.g. median — could be non-normal if data is non-normal



type of metric	distribution	estimated variance
Probability	binomial (normal)	$\frac{\hat{p}(1-\hat{p})}{n}$
mean	normal	$\frac{\sigma^2}{n}$
median/percentile	depends	depends
count/difference	normal (maybe)	Var(X) + Var(Y)
rates	poisson	$\bar{X}$
ratios ←	e.g. $\frac{\hat{p}_{exp}}{\hat{p}_{cont}}$ instead of $\hat{p}_{exp} - \hat{p}_{cont}$	

Confidence interval for a mean

Measure: Mean number of homepage visits per week

$m = z^* \cdot SE$
$= 1.96 \cdot SE$
$= 12,605$

$N_1 = 87,029$ $N_6 = 92,052$
$N_2 = 113,407$ $N_7 = 60,684$
$N_3 = 84,843$ $\bar{N} = \frac{N_1 + \cdots + N_7}{7} = 91,762$
$N_4 = 104,994$ $\sigma = SD(N_1, \ldots, N_7) = 17,015$
$N_5 = 99,327$ $SE = \frac{\sigma}{\sqrt{7}} = 6430$ 95% confidence interval

② Non-parametric way : use empirical variability

∵ for complicated metrics, the distribution may be very weird

Even for simple metrics, analytical estimate of the variance ended up underestimate

∴ use A/A experiments across the board to estimate empirical variability

- difference in measures are due to underlying variability: system, population, user behavior
- When don't have enough sample for A/A test: bootstrap
- { If bootstrap estimate is similar to analytical estimates, more on
  — — — not — —, do AA test to see what's going on.

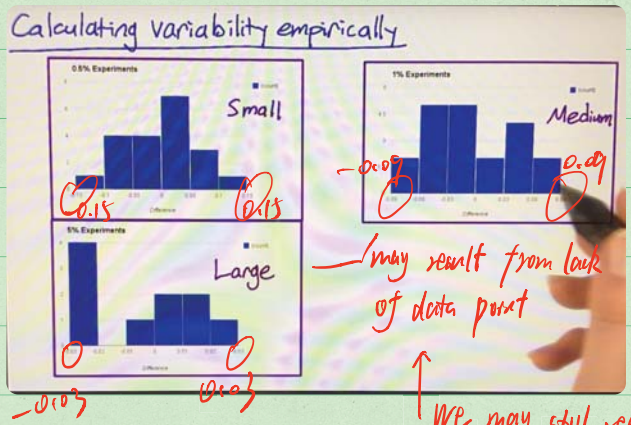
③ Calculating variability empirically

☆: Use of A/A test

(a) Compare results to what you expect using A-A test

{ 20 experiments, each on 0.5% of traffic : 10 users in each group
  20 more, each on 1% : 100 users —
  10 more, each on 5% : 500 users

How many experiments will show a statistically significant difference at 95% level ? 1/20



Range become tighter when size ↑

(b) Estimate variance and CI using A-A test

Calculating variability empirically

Estimate variance and calculate confidence interval:

Since we expect a normal distribution:

$m = SD \cdot z^*$ get SD empirically

$= 0.059 \cdot 1.96 = 0.116$ empirically

Analytically: $SE = \sqrt{\bar{p}_{pool}(1-\bar{p}_{pool})\left(\frac{1}{N_{cont}} + \frac{1}{N_{exp}}\right)}$

Slightly different margin of error for each experiment

多个那有一点点不一样

(c) Directly estimate confidence interval use A-A test

Calculating variability empirically

Directly estimate confidence interval:

2.5% -0.15 95% 2.5% +0.15

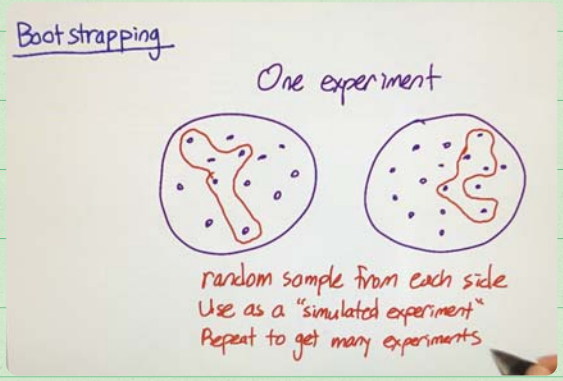
Since we have 20 data points, dropping the highest and the lowest gives a 90% confidence level: -0.1 to 0.06

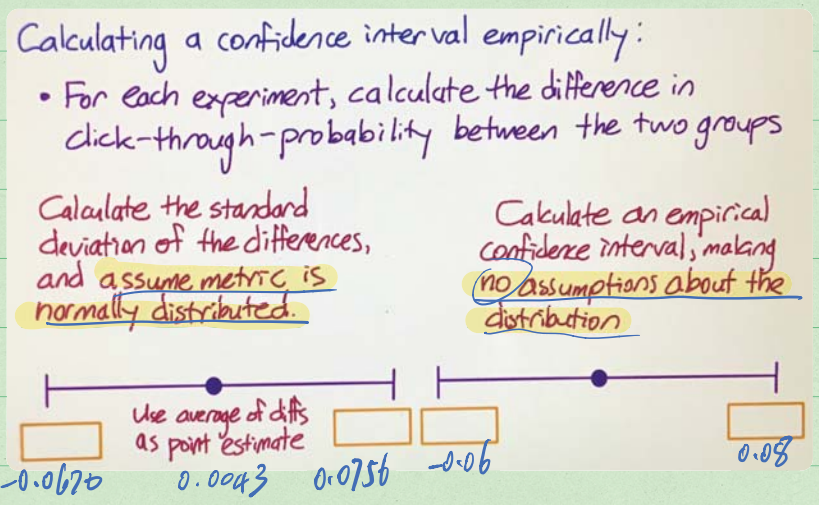
Not accurate for this case because number of A-A test to small

Empirical standard deviation: $0.059 * 1.65 = 0.097$

1.65 ← z-score for 90% confidence

(d) bootstrap for A/A test (when no enough sample)





$m = 0.036 * 1.96 = 0.0713$

SD z-score

find first/last two after sort

3. Variability Summary

① If too high, may not practical to use in evaluating experiments, even if the metric makes a lot of business or product sense

② To compute variability, need to understand distribution of underlying data

Can use both analytical and empirical techniques for compute variability

IV. Lessons learned

① There are many ways to define and aggregate metrics (related to sensitivity)
∴ need to build intuition, understand data/system and work with engineer to understand how data is captured.

② Variability: start with analytical way
(if distribution normal (revenue per query): easier to compute empirically

③ About metrics
Necessity of invariance / sanity checking

V. Conclusion:

① Establishing high level concept → ② go to detail to compute the metrics
→ ③ How to build intuition about metrics, how appropriate they'll be for different experiments.
→ ④ characterize variability of metrics