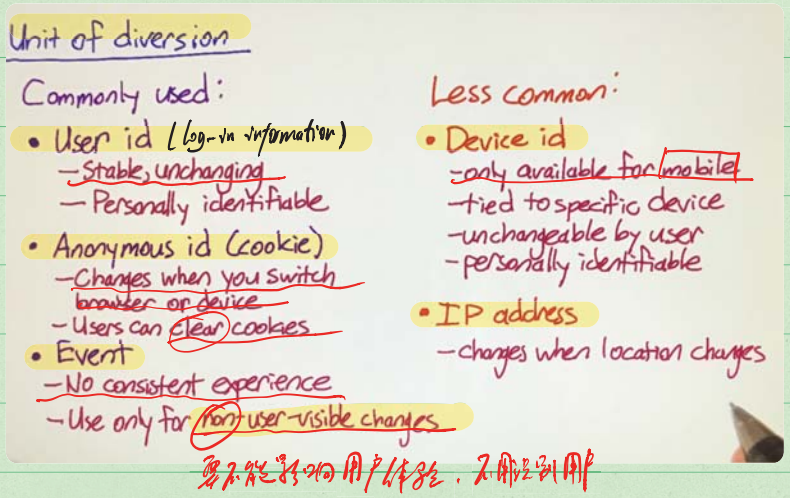
1. choose "subject"

1. Unit of diversion (Unit you are going to run test on and comparing)

need to choose a person, rather than choose event for our experiment



Eg: when can we start experiments for different unit of diversion

Only get user id after sign-in

User can clear the cookie

Each event can run different experiment

Don't have device id for non-mobile

Unit of diversion example

	desktop homepage	sign in	visit class	watch video	mobile: auto sign in	mobile: watch video
user-id	☒	☑	☐	☐	☐	☐
cookie	☑	?	?	?	☑	?
event	☑	☑	☑	☑	☑	☑
device id	☐	☐	☐	☐	☑	☐
IP address	☑	?	?	?	?	?

2. choose between different ways

① User consistency

(a) consistency for sign-in user (course layout)

If use user_id, user get a consistent experience even though they change device

(b) consistency for sign-in and out (start now)

Use cookie, but not able to accross device

User visible: user_id or cookie

{ If not visible : can use event (based on what we want to measure)

cookie : not clear (Only use when there are no other choice)

may not have a clear comparison

Which unit of diversion will give enough consistency?

Experiment	Event	Cookie	User-id
Change reducing video load time Users probably won't notice	✓	○	○
Change button color and size Distracting if button changes on reload Different look on different devices ok	○	✓	○
Change order of search results Users probably won't notice	✓	○	○
Add Instructor's Notes before quizzes Users will almost certainly notice Cross-device consistency important	○	○	✓

② Ethical consideration

Need to check security and confidentiality questions for identifiable data

Need to do informed consent (when use user_id)

Ethical considerations

Which experiments might require additional ethical review?

- ☐ Newsletter prompt after starting course — User id diversion
  - No new information being collected
  - Fine if original data collection was approved
- ☑ Newsletter prompt on course overview — Cookie diversion (email)
  - Depends: Are email addresses stored by cookie?
  - Potentially impacts other data collection
- ☐ Changes course overview page — Cookie diversion
  - Not a problem, and probably already being done

email address is identifiable which makes cookie identifiable

③ variability

(a) Unit of Analysis vs. Unit of Diversion

Empirically computed variability may much higher than analytically computed variability,

when unit of analysis (numerator) is different to unit of diversion (denominator)

eg:
- Event based diversion : $\frac{\#\text{ new page views}}{\#\text{ old page views}}$. two will be close
- User based diversion: $\frac{\#\text{ of cookie/user\_id}}{\#\text{ old page views}}$, variability of CTR computed analytically will be much higher, need to use empirically computed variability.

Reason: we make assumption of distribution of data, and assumption of what considered to be independent

- Event base diversion: every single event is a different random draw ∴ assumption valid
  User based diversion: diverting groups of events that are actually correlated ∴ assumption invalid
  ∵ correlated together ∴ will increase variability greatly

Eg: Use queries / cookies as diversion

Measure variability of a metric
Unit of diversion: query or cookie
Metric: Coverage = #queries with ad / #queries
Unit of analysis: query
Binomial: $SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{N}}$

Analytical results

Coverage

When unit of analysis = unit of diversion, variability tends to be lower and closer to analytical estimate

- Unit of Analysis: denominator of metric

When not match: analytical estimate will be an underestimate of variance

Unit of analysis and unit of diversion

When would you expect the analytic variance to match the empirical variance?

- ☐ Metric: click-through-rate = #clicks / #pageviews
  Unit of diversion: cookie
  Unit of analysis: pageview
- ☐ Metric: #cookies that view homepage
  Unit of diversion: pageview cookie user-id
  Unit of analysis: cookie
  "larger" than unit of diversion!
  Metric not well-defined
- ☑ Metric: #users who sign up for coaching / #users enrolled in any course
  Unit of analysis: user-id

Unit of diversion: User-id

II. Choose population

1. Propose of diversion: proxies of users

- When User base diversion: you're going to have one group of users on A side of experiment and one group on the B side.
- When event based diversion: end up with the mix of the same people on both sides.

Be careful not inadvertently mismatch users.

2. Inter- v.s. Intra-User Experiments

① Def of Intra-User Experiments:

Expose the same user to this feature being on and off over time and analyze how they behave in different time windows

② Pitfall of Intra-User Experiments:

(a) Need to be careful to choose a comparable time window

(I don't want do this 2 weeks before Christmas and then have them behave differently in the second part)

(b) Have frustration or learning problem for some features

People learn to use the feature in first two weeks, and frustrated when you turn it off

- note: when related to ranked order list, can use interleaved experiment:

Expose same user to A and B at the same time

③ Typically, it's inter user: user in group A and B are different

- Can use cohort analysis to better control the experiment

3. Target

① Decide who you're targeting in your users (get user space)

- Some easy dimensions: browser, geo location, ~~
  How long they use website, age, ~~

② Reason for make decision in advance: restrict people who see the experiment

- High profile launch: restrict how many users actually seen it
- UI in different region: restrict by language
- Many experiment at same time: don't want overlap
  make user flow to different traffics
- Don't want to dilute effect of experiment across a global population.

③ Some time don't need to target:
apply for all users, large percentage of traffic

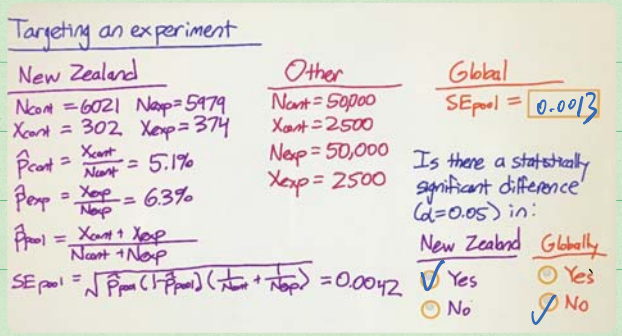
④ Involved work in real cases

(a) Talk to engineering team, tell are we sure that this is not going to trigger for particular browser; is our targeting exactly right?; Are we concerned about potential interactions so we might want to run a global experiment.

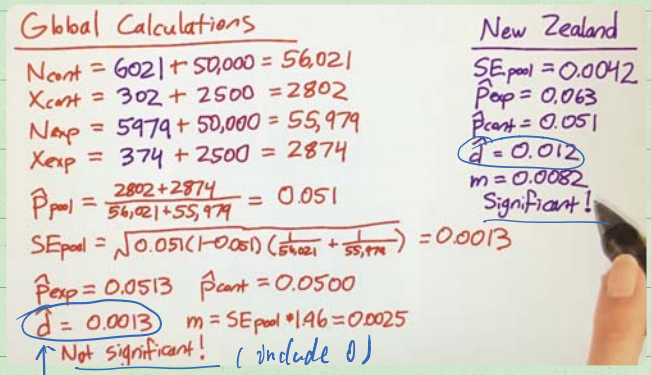
(b) Make sure have same filter on the targeted and untargeted part of experiment. Don't want to accidentally include only logged-in users on target bit, then when you compare it to your global population you realize that it's completely wrong. (Make sure everything is lined up)

(c) Before launch big changes:
Go back and run global experiment and make sure that you don't have any unintentional effects on traffic you weren't target.

⑤ Example



SE of global is much lower since there are more data

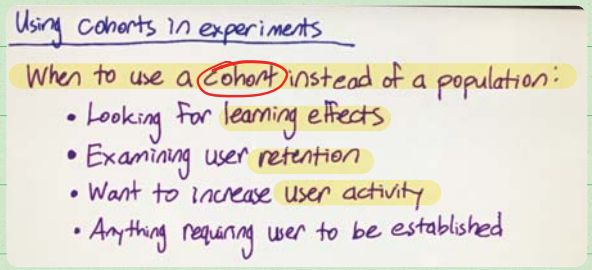


diluted by other traffic

4. Population v.s. cohort

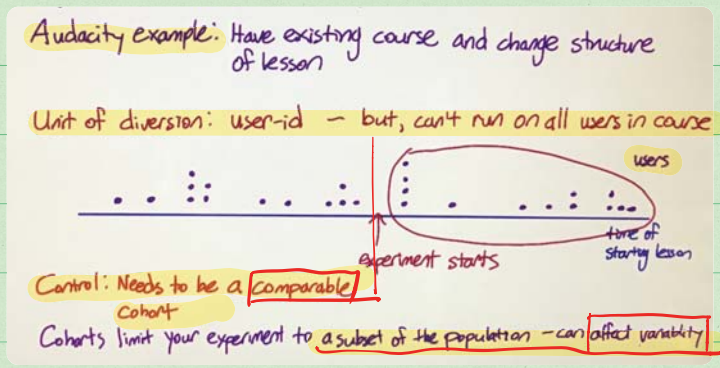
- population: a group of users
- cohort: Define an entering class and only look at users who: entered experiment on both sides around the same time / use for a long time

useful to analyses user stability: learning effect, increase usage of site



- Use a subset of population, who have seen new lesson but not the old one
- Control: need get from same time period in case there is some system changes

Need to split the latter into two groups < experiment / control



III. Experiment design and sizing

1. Overview

- It's an iterative process. try out some decisions for our unit of diversion and our population, see what the implication is on both the size as well as duration of experiment.
  - If don't like the result, we'll need to revisit decision and iterate.

2. Sizing process

① Eg: latency of redis load time based on user id:

want to see whether that user uses our site more based on the lacency experience

need to get more data (time) to get data to check variance.

∴ Need to define metrics before experiment and sizing

Redefine: Affecting the 90th percentile which is what I'm trageting

∴ look at people with slow connecting and look at a cohort where users our site regularly over the past 2 months

⇒ get more data about them more quickly

restriction cause smaller scope but give a better sense

Get a signal of experiment before invest more time for larger scale experiment.

② Variability affects sizing

How variability affects sizing

Audacity includes promotions for coaching next to videos

Experiment: Change wording of message

Metric: click-through rate = $\frac{\#\text{clicks}}{\#\text{pageviews}}$

Unit of diversion: pageview, or cookie

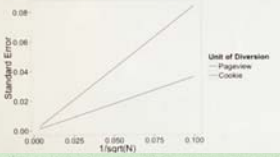
Analytic variability won't change, but probably under-estimate for cookie diversion

Empirical estimate with 5000 pageviews
- By pageview: 0.00515
- By cookie: 0.0119

How variability affects sizing

Empirical estimate with 5000 pageviews
- By pageview: 0.00515
- By cookie: 0.0119

To calculate size, assume SE ~ $\frac{1}{\sqrt{N}}$



$d_{min} = 0.02$

Diverting by Pageview: 2600

By cookie: 13,900

Build intuition by getting practice!

③ Reduce size of experiment

How to reduce the size of an experiment

Experiment: Change order of courses on course list

Metric: Click-through-rate

Unit-of-diversion: cookie

$\alpha = 0.05$ $\beta = 0.2$

$d_{min} = 0.01$ SE = 0.0628 for 1000 pageviews

Result: Need 300,000 pageviews per group!

Which strategies could reduce the number of pageviews?

- [x] Increase d_{min}, α, or β
- [x] Change unit of diversion to page view.
- [x] Target experiment to specific traffic
- [ ] Change metric to cookie-based click-through-probability

originally. we use cookies to count clicks; now, we use pageviews

How to reduce the size of an experiment

- [x] Change unit of diversion to page view — SE will decrease
  - Makes unit of diversion same as unit of analysis
  - But will less consistent experience be okay?
  - If SE changes to 0.0209 → only 34,000 pageviews per group
- [x] Target experiment to specific traffic
  - Non-English traffic will dilute the results
  - Could impact choice of practical significance boundary
  - SE changes to 0.0188, d_{min} to 0.015 → only 12,000 pageviews per group
- [ ] Change metric to cookie-based click-through-probability
  - Often doesn't make significant difference
  - If there is a difference, variability would probably go down

(d_{min} ↑)

① Since it's a subset, might need a bigger change to matter for business

② Variability lower, want to take advantage to detect smaller changes rather than decreasing the size

: unit of analysis is same as unit of diversion

Especially for short time, they are similar.

④ Sizing triggering

(a) Things we don't know

- which browser benefit most from change
- hard to predict language translation will affect how many users
- How to detect impact if have a feature triggers in a certain way

(back end. do you trigger the feature and generating recommendations)

(b) affect for size

Don't know what fraction of your population is going to be affected

∴ Need to be conservative when plan how much time and users

(c) what should do

Run a pilot or use first few days to get a better guess at what fraction of population you're really looking at.

IV. Duration v.s. Exposure

1. Overall

①
- what is the duration
- when do I want to run experiment (holidays? overlap something important?)
- what fraction of traffic is sent through your experiment

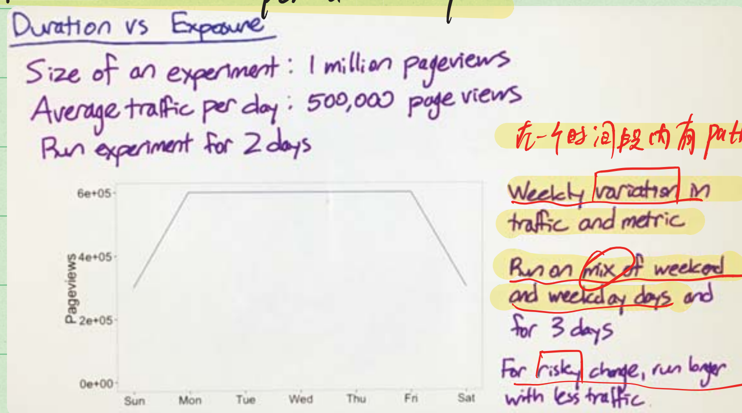
eg: 50% traffic + 10 days ⇔ 25% traffic + 20 days

② why don't run on all traffic

- Safety: new feature, not sure how it will function
- Press: new feature, not sure to keep ⇒ don't want many people to see it.
- Behave different in different days ∴ prefer small percentage for long time.
- Need to run different experiments for different levels of feature at the same time.

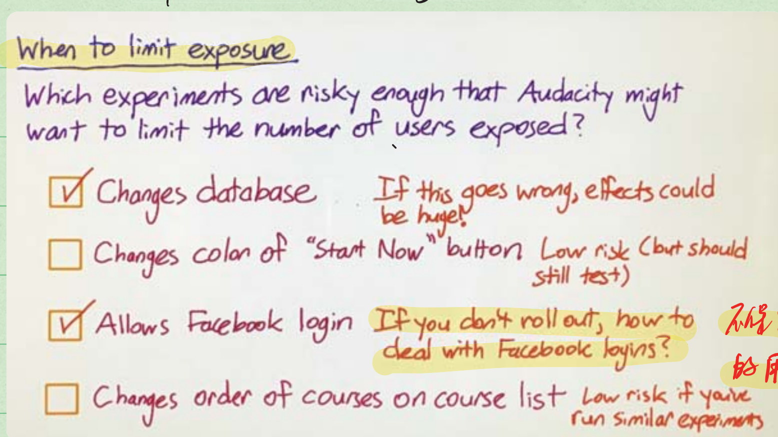
Run small traffic ∴ can be comparable at the same time

2、Pattern in a time period: example



在一个时间段内有pattern，考虑拉长时间

3、Limit exposure for risky changes



不保留这个feature，那么现在登陆的用户该怎么办 ∴ risky

4. Learning effect

① Intro

when change of version:

- Negative for change: What's that, I don't like anything
- Positive for change: This is new, what's going on. Let me try everything around.

② Definition

When user first see changes, they tend to react in one of the two ways.

But over time they're going to probably plateau a very different behavior.

③ key factor: time

It takes time for users to adopt the change and often we don't have the luxury of taking that much time to make a decision

④ Important facts when spend a lot of time for learning effects

- chose correct unit of diversion:
  Need a stateful unit of diversion to test user learning
- Many learnings are not just based on time but how often (dosage) they see the change:
  Use a cohort on both experiment and control based on how long they are exposed to the change or how many times they seen it.
- Risk and duration: only on small popotion for a relativly long time
  ∵ Need long duration ∴ Don't want many users use it. (may end up other changes)
  ∵ Uncertain of the effect ∴ tend to be high risk

⑤ pre-periods and post-periods check (Advanced ways)

A-A test that's specific to your experiment and control

(a) pre-period A/A testing (for sanity check)

- Before you run A/B test on experiment control and you have the population. you are on pre-period at same population but they're receiving the exact same frequency. (A-A test of same set of users)
  If get difference, it's due to something else: system/user variability
- Make sure the difference in experiment control is due to the experiment, but not the preexisting and inherent differences.

(b) post-period (for detection of learning effect)

- After the experiment, run another AA test.

If there is any differences in experiment and control population after I've run experiment ⇒ Attribute to user learning that happened in the experiment period

V. Lesson learned

1. Interaction between choose of metrics with unit of diversion

variability can change a lot for metrics based on units of diversion

2. Always an iterative process

3. Build intuition in the iterative process, use the intuition to better make guess about what's going to work (speed up the process)

4. Use invariant metrics (sanity check) to make sure your experiments run properly

Eg: choose a unit of diversion of cookie and a population of English,

Count how many cookies are in English for both my control and experiments.

If the invariant metric is the same, have a idea that it's properly

5. Keep an eye on the data to figure out 'surprise' points

VI. Conclusion

1. What are the different possible choices for a unit of diversion is and why we might have a preference

2. How to choose a population

3. Walk through iterative process of sizing experiment, and choose how long to run

Put these decisions with what we discussed in lesson one (statistical power, practical significance) I have all the decisions necessary to design your experiment.