L4.2

2020年4月28日 星期二 下午3:47

17 Population VS Cohort

population — whole group of users
within — def cohort
people enter exp same time — both side go forward

Ex: divert by (cookie / userid) — look at particular country

all kinds of problems — span of exp (come in — out. (time/period) / ip address — relocate 'location')

- start with initial group → use other info
  users — use site consistently 2 mon / userid = mobile - desktop.) — define cohort
- popu VS cohort
  cohort take more data ⇐ lose users
  ✓ user stability — learning effect eg. increased usage of site/mobile device
  see if change had real effect on behavior — history

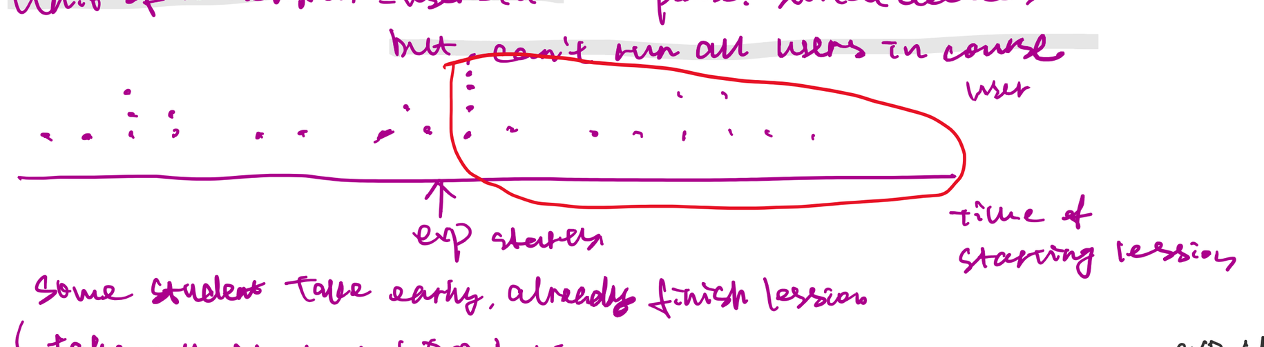
Unit of diversion: (cookie, userid, event) — Population: targets
consistent — period / location — userid) — cohort

Using cohort in exp

When to use a cohort instead of a population?
- look for learning effect
- Examing user retention 用户保留
- Anything requiring user to be established

Audacity Ex: Having existing course
change structure of lesson — if ↑ completion rate?

Unit of diversion = user-id — pause. switch devices
but can't run all users in course



Some student take early, already finish lession
(take all student (popu) ✗
✓ only include student take lession after exp start — subset of population

cohort — ↦ exp. start — time
cont. — (exp. / cont.

control: need to be a comparable cohort
✗ other system changes — affect new users
✓ split cohort (exp. / cont) same time start
choose base on time

cohort limit your exp to a subset of popu — can effect variability

Iterative process

size (practical significance, statistical significance, sensitivity)

decision — unit of diversion. popu
see application (size / duration)
✗ like — revisit — iterate

Size

Ex. Page load_time — 90th percentile latency
load-time — anonymous random identifier
page — load

diversion (event_based — each page load time / user-id — whether user use site more/less ↑ latency (learning effect!)

metric — large amount of user data — need → work.
globaly: variance of metric ✗ realistic — long time, lots of data) big investment
90 percentile — targeting
(people — slow connection
cohort: user use site regularly over past 2 mon — get data quickly
smaller scope. better sense — signal

How variability affects sizing

EX: Audacity includes promotions for coaching next to videos
Exp: change wording of message
Metric: click-through-rate = #click / #pageview
Unit of diversion: pageview or cookie — consistent experience

e.g. exp, metric, unit of diversion, variance

(Analytic variability ✗ change.
probably under-estimated for cookie diversion

Empirical estimate with 5000 pageview.
By pageview. 0.00515) variance
By cookie: 0.0119

	SE	N
pageview	0.00515	2600
cookie	0.0119	13,900

- To calculate size, assume $SE \sim \frac{1}{\sqrt{N}}$ $d_{min} = 0.02$
  Diversion by page view 2600
  by cookie 13,900
- Build intuition by getting practice

How to reduce the size of an experiment

Experiment: Change order of courses on course list
pageview — ~~order of courses~~ — clicks ✗ notice consistent.
Metric: CTR
Unit of diversion: consistent user, experience / Non-logged in traffic) cookie

$SE \sim \frac{1}{\sqrt{N}}$

$\alpha = 0.05$ $\beta = 0.2$
$d_{min} = 0.01$ — practical significance boundary
$SE = 0.0628$ for 1000 pageview. — empirically estimate SE

Result: Need 300,000 pageviews per group!
all traffic in a month
Which ~~strategies~~ could reduce the no. of pageviews?
- ✓ increase d_{min}, α, or β.
- ✓ change unit of diversion to page view. unit of diversion = unit of analysis
  ↓ variability of metric — ↓ no. of pageviews
  But less consistent experience be okay?
  If SE change to 0.0209 → only 34,000 page views / group
- ✓ Target exp to specific traffic
  - Targe to English traffic — ↓ tot pageview
    Non-English traffic dilute the results — ↑ no. of pages
    could impact choice of practical significance boundary ↑
    SE change to 0.0188, d_{min} to 0.015 → only 12,000 pageviews/group
- ✓ change metric to cookie based CTR
  - often ✗ make significant difference
    If there is a difference, variability would probability ↓
    Unit =

19 Strong trigger

① ✗ beforehand. target exp — ✗ know
Ex: Design an improvement — web page
✗ know browser benefit most
① feature on site. e.g. language detection
classif it as a language.
no. of search result higher: lang 1 > lang 2
hard to predict — how many users — affect.
② detect impact
feature trigger way
UI feature, — know people expose to it
back end = feature → recommendation — metric is for.
③ proportion of popu. effected — ✗ know
should exposed to feature ⇒ actually exposed to
conservative — plan (how long / who use — (turn on exp. for a while. — see who's affected / use 1st day/week — data — guess: what fract. of popu.

20 Duration VS Exposure

ideal size ⇒ a set of practical decision
1. Duration of exp
2. When run exp.
3. fraction of traffic — send through exp.

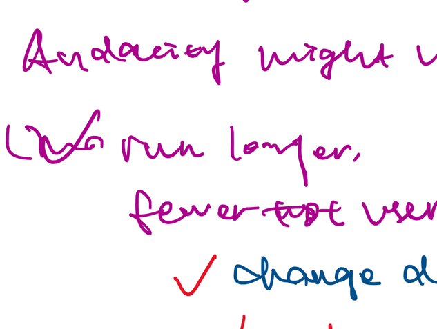
% cookie — (exp / cont

Ex 1 million cookies
100,000 cookies — visit site — given day
(50% traffic — exp. / 50% — cont — 10 days
(25% / 25% — 20 days.
Duration — traffic %

✗ run all traffic
- Safety Eg. new UI feature ? (how function in browsers / user react) — (frustrated / ✗ work well
  keep site same — expose to a few people. — comfortable
  Ex2: press — new feature keep ? = site.
  lot of people see — blog.
- randomize across unit of diversion
  other thing ⇒ variability
  Ex: (holiday = (50 / 50 single day — make decision
  weekday — weekends
  → small % — multiple days — source of variability
- 2: multiple tests on company
  diff. levels — param / types — feature) directly comparable, run same time — smaller % of traf, compare relative order of tests

Duration VS Exposure

size of an experiment = 1 mill pageview
Avg traf / day = 500,000 page views
Run exp. 2 days



- weekly variation in traf, & metric
- Run on mix of weekdays & weekends and for 3 days
- for risky change, run longer (7 days) — less traf. (low %)

When to limit exposure

Which exp risky enought that
Audacity might want to limit no. of users exposed?
(run longer, fewer ~~test~~ users
- ✓ change database
  (If this goes wrong. effect could be huge.
  test all changes → users
  change — traf. — new bugs
  change to a small % of users → every one
  wrong — effect huge. test all → users 少量多次
- change color of 'start now' button
  low risk (but should still test → users)
- ✓ Allow Facebook login
  If U ✗ roll out — ? deal with FB login? ✗ support
  affected users — small no.
- change order of courses on course list
  Low risk if u've run similar exp

23 learning effect

user learning — ? use adapt to change
(changing: change — like? / novelty = like everything.) — over time: plateau stat. behav — measure.
→ change time — start
measure: time = adapt to change
✗ much time decid

① stateful unit of diversion = cookie / userid
② how often see change — dosage — cohort ✗ popu (exp / cont
(how long time / how many times ~~exp~~ change
③ Duration — time = see effect
✗ put lot of users — change . long time] small proportion of user long time
Uncertain — effect — higher risk change
④ (pre-periods / post). uniformity trials

pre — same popu. — same freq A/A test.
diff. problems ⇒ (exp / cont — (system variability / user.
difference — exp,
✗ preexisting inherent difference in popu
post — period
after (exp / cont → A/A — differences happen in exp. period
 lot exp. already

- choose diversion — what need to be consistent
  population — who — affected
  sizing
  run exp. — put user through — see reaction
  go through work
- analytics work ✗
  ✓ good framework — how many users need
  type of variability
- ~~choice of metrics~~ — unit of diversion
  ✗ compute
  change
- iterative process
  metrics — exp.
  like results — see sth.
  go back — settle
  build intuition — use — better guess what going to work — speed up process
- invariant choice — unit of diversion / population — Ex: (cookies / English — (exp / cont

— feature ~~trigger~~ 5% time.
exp 1st day 50% traffic