Experimental Design & Recommendations

Concepts on Experimental Design

↳ correlation != causation

↳ other factors may influence variables

How do we deduce causality?

↳ Experiments (control confounding vars)

Experiments

↳ comparing results of 2+ groups

↳ control by random assignment of to groups

- What if this isn't possible?

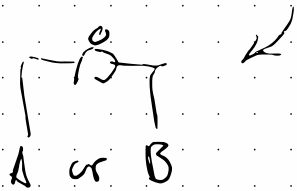
↳ observational studies

ex smoking, studies not possible due to ethics

↳ establishing causality becomes difficult

Types of experiments

A/B tests: compare treatments vs controls (Between subj. design)



↳ sometimes it's possible to test of A & B on the same person

Sampling

↳ working population data is difficult

↳ sample is used to est pop. sized parameters

(select at random)

Doesn't work well if people aren't distributed uniformly

Measuring Outcomes

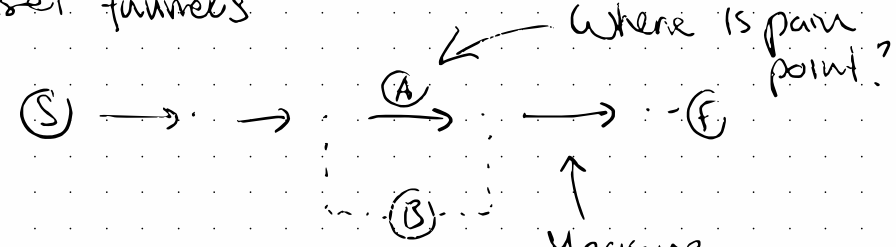
↳ what metrics are best to choose for experiments

depends on goals, keep in mind that your metrics can get hacked

Creating Metrics

↳ Think about user journey (have empathy)

↳ user funnels



Measure outcome(s)

↳ variations

↳ constants

controlling variables

↳ vars that crend (in)dependent should remain constant

↳ removes confounding variable

Checking validity (useful for handling data in general.)

- construct
  - ↳ Metrics are aligned w/ goals
- internal validity
  - ↳ "Degree to which an exp.'s claims of causality can be supported"
- external validity
  - ↳ Degree to which study can be generalised

Checking Bias

- Systematic errors that effect interpretation of results
  - ↳ Sampling
  - ↳ Methodological
  - ↳ Novelty
  - ↳ order (primacy, recency)

Ethics

- ↳ Participant risk ↓
- ↳ Clear benefits to experiment
- ↳ Informed consent (Difficult)
- ↳ Have strong privacy/security

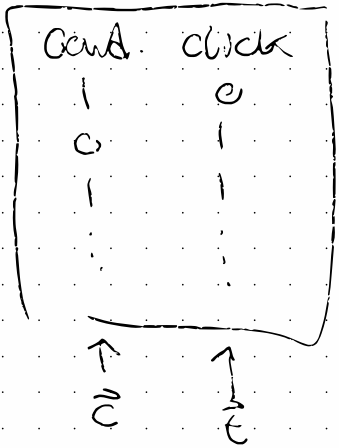
S - specific
M - measurable
A - Achievable
R - relevant
T - timely

Statistical Considerations in testing

↳ How much data?
↳ & other decisions

Stat. sig

↳ tests have tables (like so.



Steps,

① validate cond is even split

② check if clicks for sig diff

$obs = len(\vec{c})$

$n_{control} = obs - \sum \vec{c}_i = n(0 \in \vec{c})$

→ Check control

↳ $H_0 : p = 0.5$
$H_1 : p \neq 0.5$

$sd \mid p = 0.5 = \sqrt{p(1-p) * n}$

$$z = \frac{n_{control} + .5 - p\,n_{obs}}{sd}$$

observed correction

norm. const.

expect val

then... compare z w/ $1.96 * \text{norm.cdf}(z)$

p-value

p-value is prob. of data given parameter
ie prob of seeing data if null is true

↳ M evenly dist, p val should not be significantly different.

⟶ Check circles for sig diff.

for each c in $\vec{C}$, there is p of circle
{t,o?}

↳ $P(C_1) - P(C_1) = P_{diff}$

null hypothesis is overall avg

H_0: $P(C_1) = P(C_0)$

H_1: $P(C_1) \neq P(C_0)$ (?)

$$\text{std err} = \sqrt{P_{null}(1-P_{null}) * \left(\frac{1}{n_{cont}} + \frac{1}{n_{exp}}\right)}$$

$$z = \frac{P_{diff}}{\text{std err}} \Rightarrow 1 - \text{norm.cdf}(z)$$

p value

Note, to get p value you need to establish H_0 & H_1

↳ establish in terms of a diff (Δ)

↳ normalise by variance or std err

↳ find where H lies on a distribution depending on tests.

Simulation

↳ draw from Binomial(n_θ, p_{null}, many times) $\quad \theta \in \{exp, cont\}$

$$diffs = \frac{\text{exp clicks}}{n\ exp} - \frac{\text{cont clicks}}{n\text{-}control}$$

↖ List of normalised diffs

$$\text{if } [p_{diff} < diff_i \text{ then false else true}].mean()$$

↖ gives p-value

Practical Significance

- even if difference is significant, it doesn't mean that it should be implemented

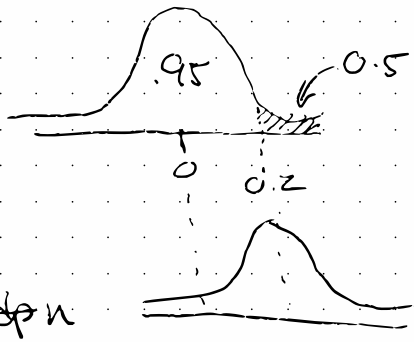
Experiment sizes

Statistical Power: given true mean, the prob. of rej H_0

ie.

α = 10% conv

β = 12% conv (desired)] — C.I. = 2%



increasing α will narrow curves & increase power of test

find n people we will need for experiment

note, power $= 1 - \beta$ ← rate of type 2 error
$P(\text{reject } H_0 \mid H_0 \text{ is true})$

$se_{null} = p_{null}(1 - p_{null})$

Dummy tests

↳ comparison between same exp. group

A/A test
- ↳ Make sure tests is working
- ↳ get more info about metrics

Analyzing Multiple Metrics

↳ multiple metrics → increases chance for false +ve

- ↳ variant 1
  - metric 1
  - metric 2
- ↳ variant 2
  - metric 1
  - metric 2

← both not sig $= .95^2 = .9025$

→ Apply bonferoni correction to account for this

Early stopping

↳ Stopping & viewing test results can lead to have false tives

↳ need to google how to do early stopping if necessary

Introduction to recommendation engines

↳ used in many ways on the web these days.

Knowledge based recommendations

↳ good recommendations?

↳ How about top song/movie/etc.

↳ most popular in genre?

knowledge based

↳ recommends top rated for certain criteria

How is this done?

↳ create df sorted by "top" criteria (highest avg, most rated, etc.)

↳ filter as necessary

Advantage of knowledge based?

↳ don't know a lot about users

⊗ not personalized to the user

Collaborative filtering

↳ "people who y x also y z."

→ two types

model based

Neighborhood based

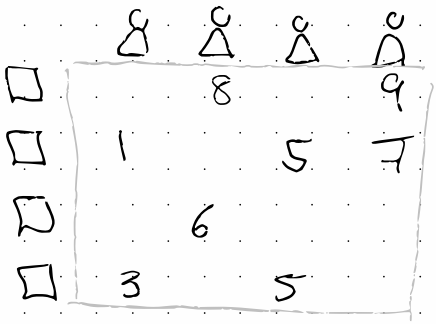
N. based collaborative filtering

Similarity vs distance

→ correlation coeff.

→ distance measurements

Movie, Person matrix



for recommendation to certain person remove:
- seen movies
- movies w/ low rat.

← creates sparse matrix

↳ for each user, calc list of movies seen

↳ recommend differently depending on number of movies seen.

<2 ↙

↘ ≥2

calc corr. between users

similar movies

ratings

↑ you can even use different similarity metrics

content based recommendations

↳ "cold start problem"

What if the number of users you have is small?

→ find similar items by: name
appearance
item type
etc.

↳ think of people w/ less than 9 ratings for a movie

get users & their top ranked movies

↳ if movies x attribute matrix is A then take $A A^T$ to get a dot prod matrix for movies

↳ $A^T A$ for which attributes are similar

→ for a top rated movie, find the most similar. (which haven't been seen already)

Types of ratings

- likes vs dislikes
- 3, 5, 10 star ratings

↓ higher granularity

→ include neutral? depends

Goals of recommendations

↳ relevance
↳ novelty
↳ serendipity

] exposing users to something new

Matrix Factorisation for Recommendations

How do we know if recommendations are good?

How to use MF for recommendations

How to recommend to new users

Quality of recommendations

- ↳ watch metrics over time
- ↳ AB test
- ↳ train/test cross validation
  - ← you can also do older vs newer data

SVD

$$A = U \Sigma V^T$$

- Σ: singular values
- U, V^T: orthonormal matrices with the basis of the row/col spaces.

Why use SVD? (Latent Factors)

↳ Factors that aren't in the data but are hidden underneath.

- think how PCA also reveals stuff like this.

$$A = \begin{matrix} \\ \text{user} \end{matrix} \overset{\text{item}}{\begin{bmatrix} \vdots & \vdots & \vdots \\ \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots \end{bmatrix}} = U \Sigma V^T$$

U ← orthonormal basis for user directions of most variance ($N \times K$)

V^T ← on basis for item dir of most var. ($K \times M$)

Σ ← connects U w/ V^T w/ strengths ($K \times K$) (latent factors) (Diagonal Matrix)

you can take a subset of PCs in order to recreate a full picture
↑ estimate

Σ ← will contain the variance found in each pc

↳ cumm of var explained = $\frac{\sum_i \text{exp } Pc_i^2}{\sum \text{all } Pc}$

Problem: SVD breakdown w/ missing values!

Funk SVD

↳ possible to do SVD w/ missing vals

① fill V^T & U w/ random numbers
② choose non missing var ($\hat{r}_{i,j}$)
③ get pred b w/ $U_{i,*} \cdot V^T_{*,j}$
④ get sq error
↑ minimize w/ GD

$$\frac{\delta}{\delta u_i}(y - uv)^2 = -2(y - uv)v_i$$

$$\frac{\delta}{\delta v_i}(y - uv)^2 = -2(y - uv)u_i$$

example

$y_{ij} = 9$

$u_i = [0.8, 1.2, -.2]$

$v_i = [-1.2, 1.1, -0.2]$

$u_i \cdot v_i^T = .4$

$\therefore$ error $= (9 - .4)^2$

$= \ldots$

update u_i, v_i

$u_{new} = u_i + 2\alpha(\text{error})\, v_i$ ($\alpha = 0.01$)

$v_{new} = v_i + 2\alpha(\text{error})\, u_i$

↳ rinse & repeat till convergence

Some quick notes on CV w/ old & new vols

1. take dataset
2. order by date
3. split data