# Experimental optimization

## Lecture 5: Evaluating and presenting results

**David Sweet**

# Review

## A/B testing concepts

How and why do we randomize?

# Review

## A/B testing concepts

Why do we replicate
i.e., take many individual measurements?

# Review
## A/B test design

What is the goal of A/B test design?
(Why not just take a large number of randomized, individual measurements?)

# Review
## A/B test design

What are the two types of random measurement errors?

How do we limit them?

# Review

## A/B test design

N too small
Random errors (FP, FN) too frequent

N too large
Experimentation cost too high

N just right
FPR < 5%
FNR < 20%

$$N = \left(\frac{2.48\hat{\sigma}_\delta}{PS}\right)^2$$

# Review

## A/B test design

- [9 students in class] X [3 midterm experiments each] = 27 experiments
- P{false positive} = .05
- P{false negative} = .20
- If all A & B had equivalent BM, expect
  - .05 X 27 = 1 or 2   false positives on mid term
- If BM(B) = BM(A) + PS, expect
  - .20 X 27 = 5 or 6   false negatives on mid term

# Review

## A/B test design

How do you estimate $\hat{\sigma}_\delta$?

How do you choose PS?

# Review

## A/B test design

Run a pilot study until $\sigma_A$ and $\sigma_B$ stop changing.

$$\sigma_\delta = \sqrt{\sigma_A^2 + \sigma_B^2}$$

# Review
## A/B test analysis

What is the end goal of A/B test analysis?

# Review

## A/B test analysis

Under what two conditions should you switch your production system from version A to version B?

# Review

## A/B test analysis

$$z = \frac{\mu}{SE} > 1.64$$

$$\mu > PS$$

# HW 3, 1a

a) Design an A/B test -- i.e., report the number of individual measurements, $N$ -- required for a system where the standard deviation of the current system's business metric is $\sigma_A = 10$ and the desired practical significance level is $PS = 1$.

# HW 3, 1a

$$N = \left(\frac{2.48\hat{\sigma}_{\delta}}{PS}\right)^2$$

```
sigma_A = 10
sigma_delta = np.sqrt(2*sigma_A**2)
PS = 1
N = round((2.48 * sigma_delta / PS)**2 + .5)
print (f"N = {N}")
```

```
N = 1230
```

# HW 3, 1b

b) Imagine you're a quant who works on a trading system. The current system earns a pnl of 10,000 dollars/day with standard deviation of 15,000 dollars/day. You deploy a new returns-prediction model that you suspect will increase pnl. Your manager tells you that if it doesn't increase pnl by at least 500 dollars/day, they won't take the risk of deploying it. Design an experiment to test whether your new model should be deployed. If your trading system produces 1000 individual measurements/day, how many days will it take to run your experiment?

# HW 3, 1b

$$N = \left(\frac{2.48\hat{\sigma}_\delta}{PS}\right)^2$$

```
sigma_A = 15000
sigma_delta = np.sqrt(2*sigma_A**2)
PS = 500
N = round((2.48 * sigma_delta / PS)**2 + .5)
im_per_day = 1000
num_days = round(N / im_per_day + .5)
print (f"N = {N} num_days = {num_days}")
```

```
N = 11071 num_days = 12
```

# Evaluating results

## Present to stakeholders

- Stakeholders
  - You
  - Your team
  - Other affected teams (ex., dependencies, tradeoffs)
- Usually evaluating multiple metrics (ex., revenue, clicks, time spent)
- Stakeholders may value metrics differently

# Evaluating results

## Approval

- Create an approval process to follow for each experiment, ex:
  - Present to stakeholders
  - Discuss
  - Final decision: manager, designated committee, vote (?)
  - Document decision (people disagree, forget)
- Standardized process helps remove experimenter bias, reduce conflict

# A/B test presentation

## Ad serving system

- You work on an ad-serving team for a website
- Your pages all show a single ad, the one with the highest predicted probability of getting a click
- You earn revenue when users click on ads
- You just completed an A/B test …

# A/B test review #28364

- A: Currently displaying the one, best ad on each page
- B: Try displaying the two best ads on each page
- BM: Increase clicks/page
  - How? P{click on either of two} > P{click on just one}
- Guardrails: sessions/day, pages/session, time/session

session = one site visit, potentially multiple pages

# A/B test review #28364

- Design:
  - $\sigma_\delta = 0.12$ (estimated from logs)
  - PS = 0.003 clicks/page (from data science group report, 2021Q4)
  - $N > (\frac{2.48 \times \sigma_\delta}{PS})^2 \sim 10{,}000$
- Need at least N = 10,000 pages

# A/B test review #28364

- Measurement:
  - Allocated 1% of users to A and 1% to B; randomly-chosen users
  - Ran for 5 days
  - Collected measurements from 5,452 sessions with A and 5,896 sessions with B
  - (!) Entire system was down for 1.5 hours on the second day

# A/B test review #28364

- Analysis:
  - A clicks/page = .017
  - B clicks/page = .021
  - $\mu = .004 \pm .0017$ clicks/page
  - $z = 2.35$

- Both criteria for switching to B are met
  - $\mu >$ PS $= 0.003$
  - $z > 1.64$

# A/B test review #28364

- Guardrails: no change

| | A | B |
|---|---|---|
| • sessions/day/user | 0.403 +/- .03 | 0.39 +/- .03 |
| • pages/session | 2.2 +/- .015 | 2.4 +/- .013 |
| • time/session | 24.1s +/- 5.7s | 22.1s +/- 5.9s |

# A/B test review #28364

- Summary:
    - Clicks/page increases by 0.004 when we show two ads/page
    - This number is both statistically and practically significant
    - No guardrail metrics are worsened
- **Recommendation: Show two ads/page**

# Presenting results

- Describe the system
  - ex., ad server, fraud detector, recommender system
- Describe the business metric
  - ex., revenue, fraud accuracy, user engagement
- What part of the system is being modified? ex., the ML predictor
- How was it modified? ex., a new feature was added
- How/why do you think your “version B” will improve the BM?

# Presenting results

## A/B test design

- How did you take an individual measurement?
  - One presentation of an ad, and Was it clicked?
  - One day’s revenue
  - Time spent on your app by a single user in a single session
  - One presentation of a post, and Was it liked?
  - One play of a song, and Was it skipped?

# Presenting results

## A/B test design

- The value of $N$, the number of individual measurements you took
- How long should did it take to collect all $N$ (ex., 1 week, 1 month)?
- How did you monitor the business metric(s)? (ex., a URL to a dashboard)
- What is PS? What was your rationale for choosing this value?
- How was $\hat{\sigma}_\delta$ estimated?
- Display $\hat{\sigma}_\delta$, $PS$, $N$

# Presenting results
## A/B test measurement

- How did you perform randomization?
  - Did you assign users (randomly) beforehand to “A” or “B”?
  - Did you randomly choose A or B on every event?
  - Did you randomly choose A or B at time intervals?
- Discuss possible confounders

# Presenting results

## A/B test measurement

- Were there any system problems during measurement?
  - System problems might introduce sampling or confounder bias
  - Ex: “West-cost system outage”, sampling bias
  - Ex: B code failed on Monday, but was fixed; confounder bias if measurements from A on Monday are included

# Presenting results

## A/B test measurement

- Were there any broad-scale, unusual events during measurement?
    - COVID-19 discovered, markets go nuts
    - Election day, Twitter very active with election-specific tweets
    - Taylor Swift releases new album on Spotify, activity is high and focused
    - Blackout on East Coast, activity is low for those users
- Measurement may not be a good predictor of “most of the time”
- May introduce sampling bias (in blackout case)

# Presenting results
## A/B test analysis

- Clearly define the business metric, BM, being used to evaluate this experiment
  - Ex: “pnl” not enough; “pnl measured daily at 4pm, net of exchange fees, marked to prices from Bloomberg” is better
  - Describe the in-house technology used to measure the business metric; “the Python function pnl_3a() in pnl_metrics.py”
- Display $\mu$, $z$ and conditions required to accept B

# Presenting results

## A/B test analysis

- Discuss other relevant business metrics even if not the one used to evaluate
- Would switching to B reduce other metrics, even if it increases BM?
  - Often the case
  - Ex: Users retweet more, but post less
  - Ex: Profit increases, but so does risk
- Stakeholders may value metrics differently
  - Ex: ad team wants more ads shown, but song-recommender team wants more songs played