Week 4: Presenting Results

AIM-5014-1A: Experimental Optimization

David Sweet // 20230622

Review: LLN, CLT, A/B Testing

- As $N \to \infty$

  - LLN: $\bar{y} \to E[y_i]$, measured BM approaches "true" BM

  - CLT: $\bar{y} \sim \mathcal{N}(E[y_i], \sigma^2)$, measured BM is normal (gaussian)

- **Design**: $N \geq \left(\dfrac{2.5 \hat{\sigma}_\delta}{PS} \right)^2$

- **Measure**: Randomize, $\bar{\delta} = \bar{y}_B - \bar{y}_A, \quad se = \sigma_\delta / \sqrt{N}$

- **Analyze**: If $\bar{\delta} > PS$ and $\dfrac{\bar{\delta}}{se} \geq 1.64$ (and guardrails ok), then accept B.

Review: False Positive Traps

- **Don't stop early**, even if t-stat looks good

- **Beware multiple comparisons** in A/B/C/… tests

  - Use Bonferroni correction: p = 0.05 / (K-1)

  - Then accept if: $\bar{\delta} > PS$ and $t = \dfrac{\bar{\delta}}{se} \geq 1.64$

Where have we used the iid assumption so far in this class?

Standard Errors

- Poorly-estimated se will ruin an experiment

- Usually se gets underestimated:

- $t = \dfrac{\bar{\delta}}{se} \geq 1.64$, t is overestimated

- Generates false positives

Standard Errors

- iid - independent, identically distributed

- Ex, $\sigma^2 = \dfrac{\sum_i^N (y_i - \bar{y})^2}{N}$ assumes

  - $cov(y_i, y_j) = 0$ <== independent

  - $E[y_i] = E[y_j], var[y_i] = var[y_j], \ldots$ <== identically distributed
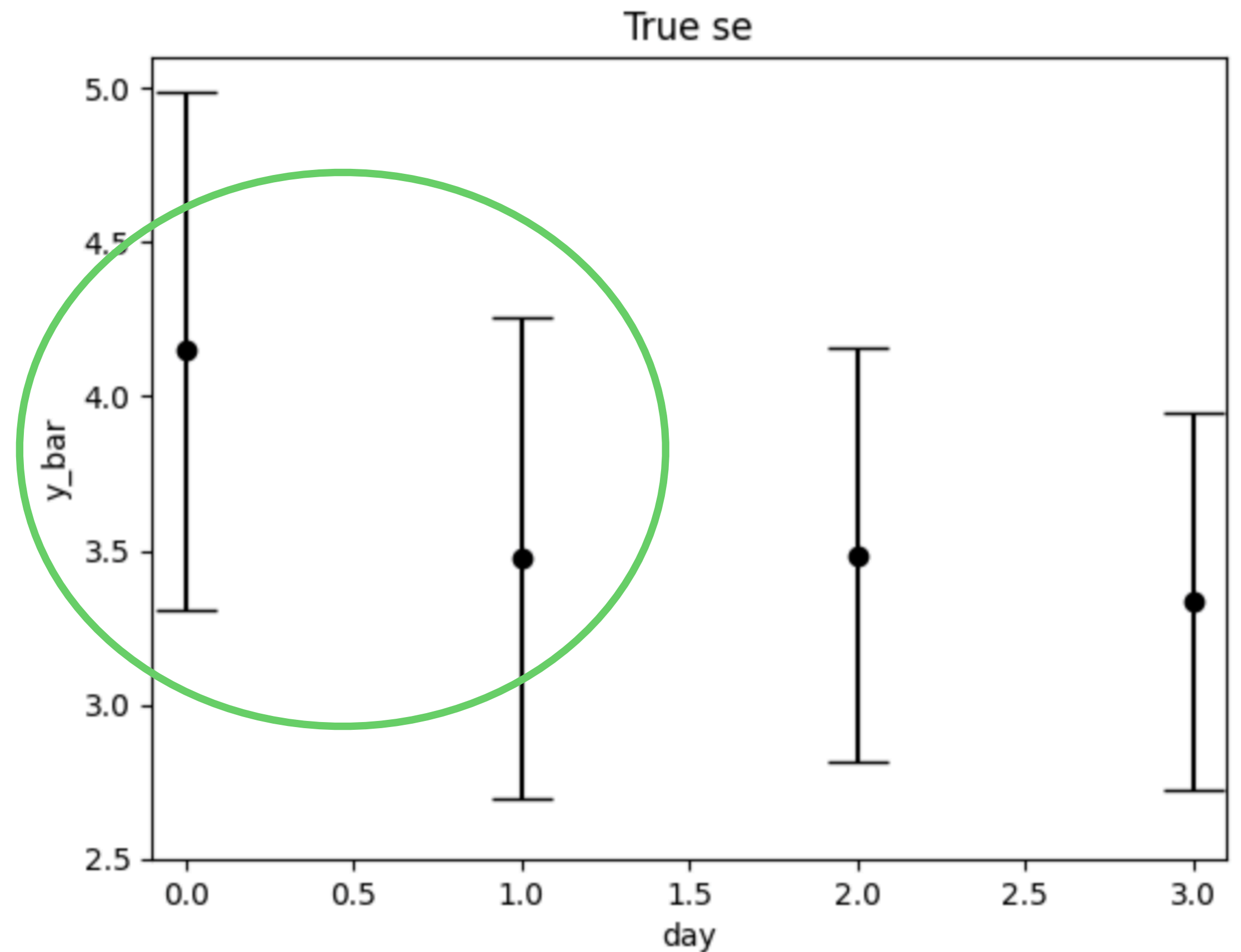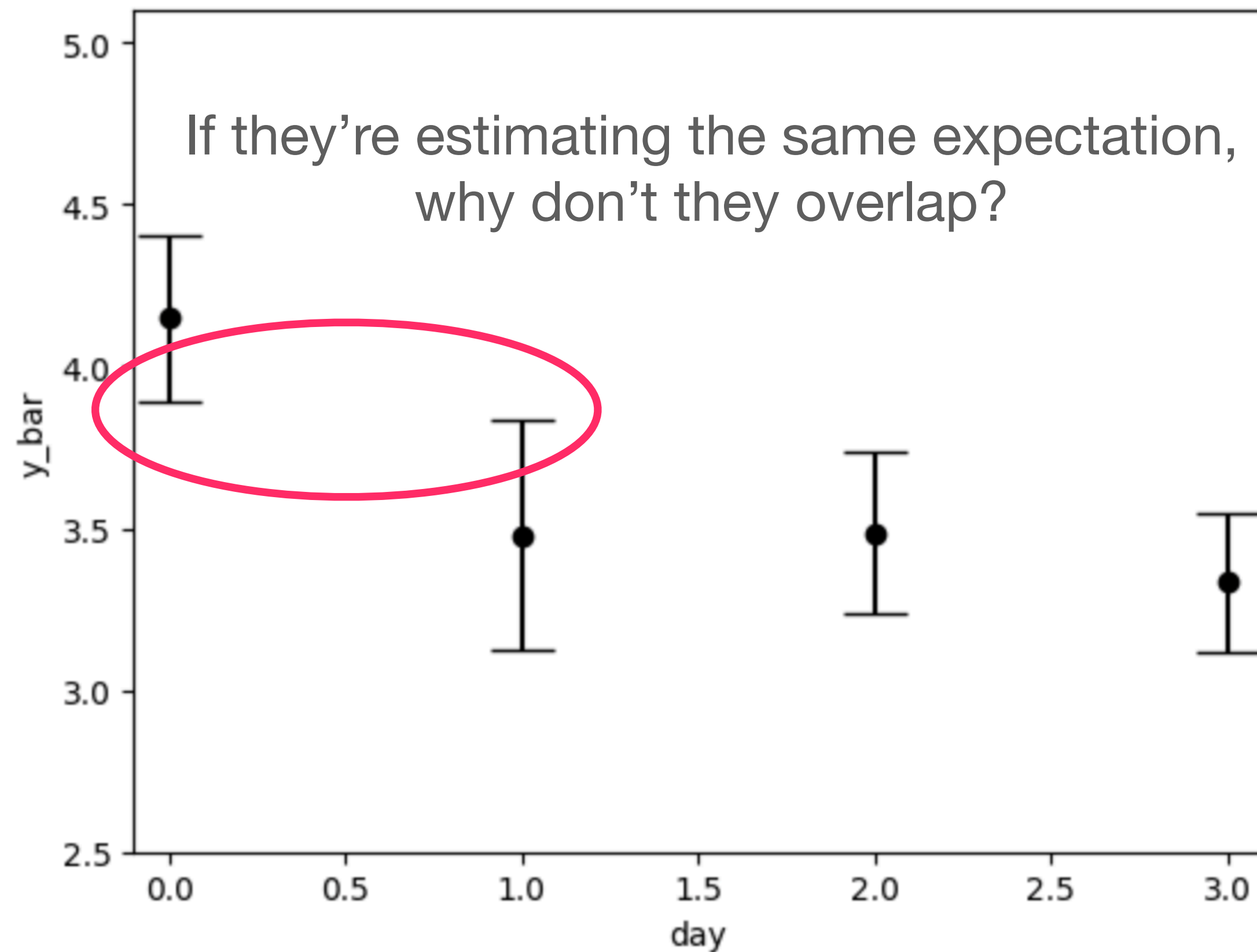
Standard Errors: iid violations

- $cov(y_i, y_j) = 0$ violation

  - timeseries autocorrelation

  - correlated behavior across users

  - correlated behavior across stocks

- Common problem, big problem

Timeseries data are often autocorrelated. What could give rise to this? Gives examples.

Standard Errors: iid violations

- $cov(y_i, y_j) = 0$ violation

If they're estimating the same expectation, why don't they overlap?

True se

Standard Errors: iid violations

- $E[y_i] = E[y_j]$ violation

- Expectation can vary with

  - user, stock, time of day, day of week, genre of song, industry of stock, age of user, length of tweet, … (confounders)

  - transient effects (nonstationarity)

  - passage of time (nonstationarity)

- When running an experiment we try to isolate the effect of A/B on y_i

Standard Errors: iid violations

- $var[y_i] = var[y_j]$ violation

- called heteroscedasticity

- variance can vary with

  - user, stock, time of day, day of week, genre of song, industry of stock, age of user, length of tweet, …

  - any feature that might predict y_i might also predict $var[y_i]$

What is a holdout test and what is it used for?

Validation of Results

- Replication: Measure again

- No other tricks

- Replication crisis: Independent re-experiments don't reproduce original
  https://en.wikipedia.org/wiki/Replication_crisis

- Avoid crisis

Validation of Results

- Industry replication techniques

- Reverse A/B: Switch to B, but run small portion as A for a while

- Holdout

  - Start of quarter: Fix a set of users, NE, no experimentation

  - During quarter: Monitor difference between NE and rest of users

    - Is the difference growing as expected?

  - End of quart: Run A/B test comparing NE to "all changes from this quarter"

Recap

- Underestimating se increases false positives

  - Look for non-overlapping error bars, autocorrelation

- Replication is the only check on results

  - Reverse A/B

  - Holdout

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)

  - $N > (\dfrac{2.5 \times \sigma_\delta}{PS})^2 \sim 10{,}000$

- 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 10,452 sessions with A and 10,896 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
  - $\bar{\delta} = .004 \pm .0017$ clicks/page
  - $t = 2.35$

- Both criteria for switching to B are met
  - $\bar{\delta} > \text{PS} = 0.003$
  - $t > 1.64$

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 σ_δ 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 $\bar{\delta}$, t 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

Summary

- Create an experimentation process to reduce bias and conflict

- Include all stakeholders in decision-making

- Presenting results:

  - Describe BM, guardrails, design (N), measurement (randomization)

  - Report unusual events / problems

  - Report analysis: $\bar{\delta}$, t, guardrails

  - Interpret and recommend an action