Experimental optimization

Lecture 4: A/B testing III: Operational considerations

David Sweet

Mid-term project

Simulation server

```
url0 = "http://3.90.111.98"

def hello():
    return requests.get(url0).text

def start(id_sim, N):
    url = f"{url0}/start_measurement?id_sim={id_sim}&N={N}"
    return json.loads(requests.get(url).text)['id_experiment']

def check(id_experiment):
    url = f"{url0}/check_measurement?id_experiment={id_experiment}"
    d = json.loads(requests.get(url).text)
    ind_meas_a = np.array(d['a']['individual_measurements'])
    ind_meas_b = np.array(d['b']['individual_measurements'])

    return ind_meas_a, ind_meas_b
```

Review: A/B testing

- **Design**: Specify number of replicates / individual measurements
  - $N > (\frac{2.48\hat{\sigma}}{PS})^2$
- **Measure**: Randomize A & B

More on this today

- **Analyze**: If $z > 1.64$ and $\mu >$ PS, then switch to B.

Deploying an A/B test

Safety first

- Three steps
  1. Small-sized A/A test
  2. Small-sized A/B test
  3. Full-sized A/B test
- If any step fails, start over

Deploying an A/B test

Small-sized A/A test

- “A/A”, colloquialism
  - Create two branches of code, one for A and one for B
  - Run the A code in B’s branch

Use a config flag

- Set up production system to run experiment
  - Deploy experimentation tooling
  - Engage experimentation system
  - Send small amount of flow (users, trades, etc.) to second “A”

Deploying an A/B test

Small-sized A/A test

- Look for any deviations from normal behavior
  - Ex: $|z| > 1.64$, indicating BM has changed
  - Usually monitor several secondary metrics, too
- Should see that the new branch behaves no differently
- Should see that the experiment tooling is functioning properly
- “Small” is like 1% of N, (the # of individual measurements from your design)

Deploying an A/B test

Small-sized A/B test

- Activate B, i.e. flip the config flag to True
- Stay at 1% of N
- Look for bugs in B's code
- Too few individual measure,ents to measure precisely, but
  - Look for large, adverse changes in BM
  - Look for large, adverse changes in secondary metrics

Deploying an A/B test

Full-sized A/B test

- Increase the flow to full scale, collect N individual measurements
- DO: Monitor BM and secondary metrics for large adverse changes
- DON'T: Stop the experiment if you see $z > 1.64$
  - Called "early stopping"; generates tons of false positives

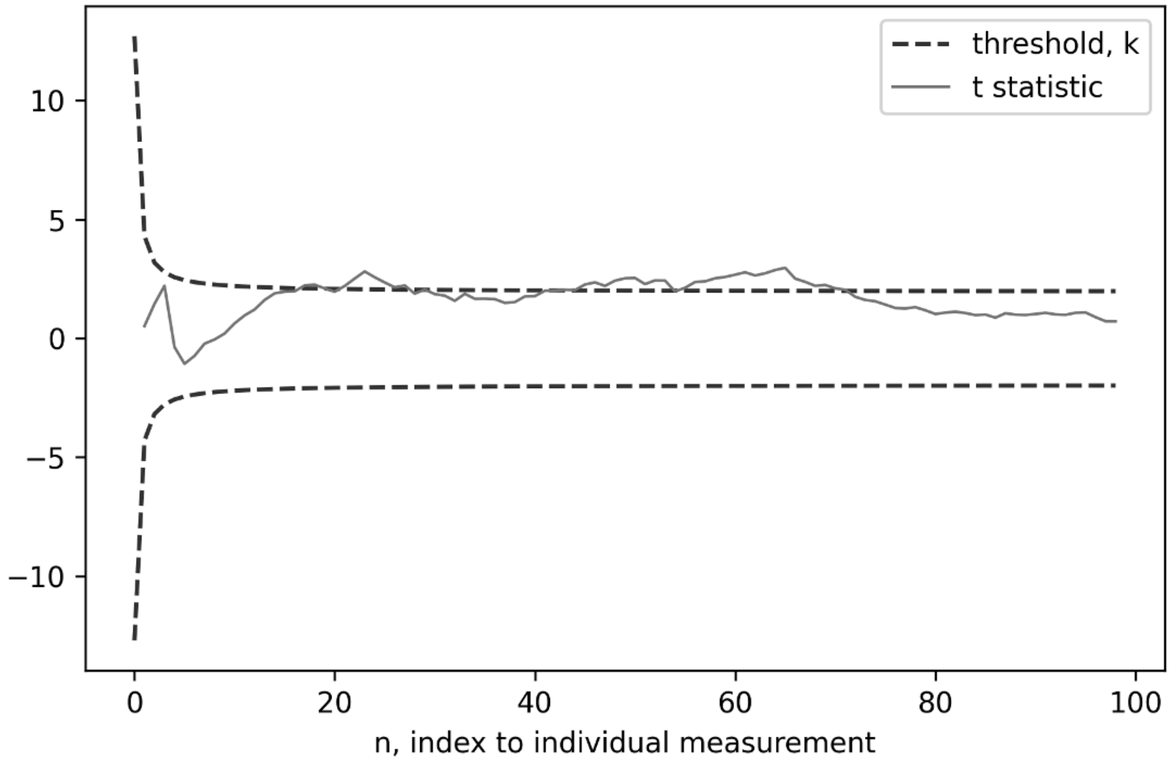
Unrelated to NN's regularization technique of the same name

Early stopping

Generates false positives

- Aka: “snooping”, “peeking”
- It’s ok to *look* at your z-score (or t-score), but
  - it’s NOT ok to stop the experiment and *accept* B.
- Why? z (or t) fluctuates

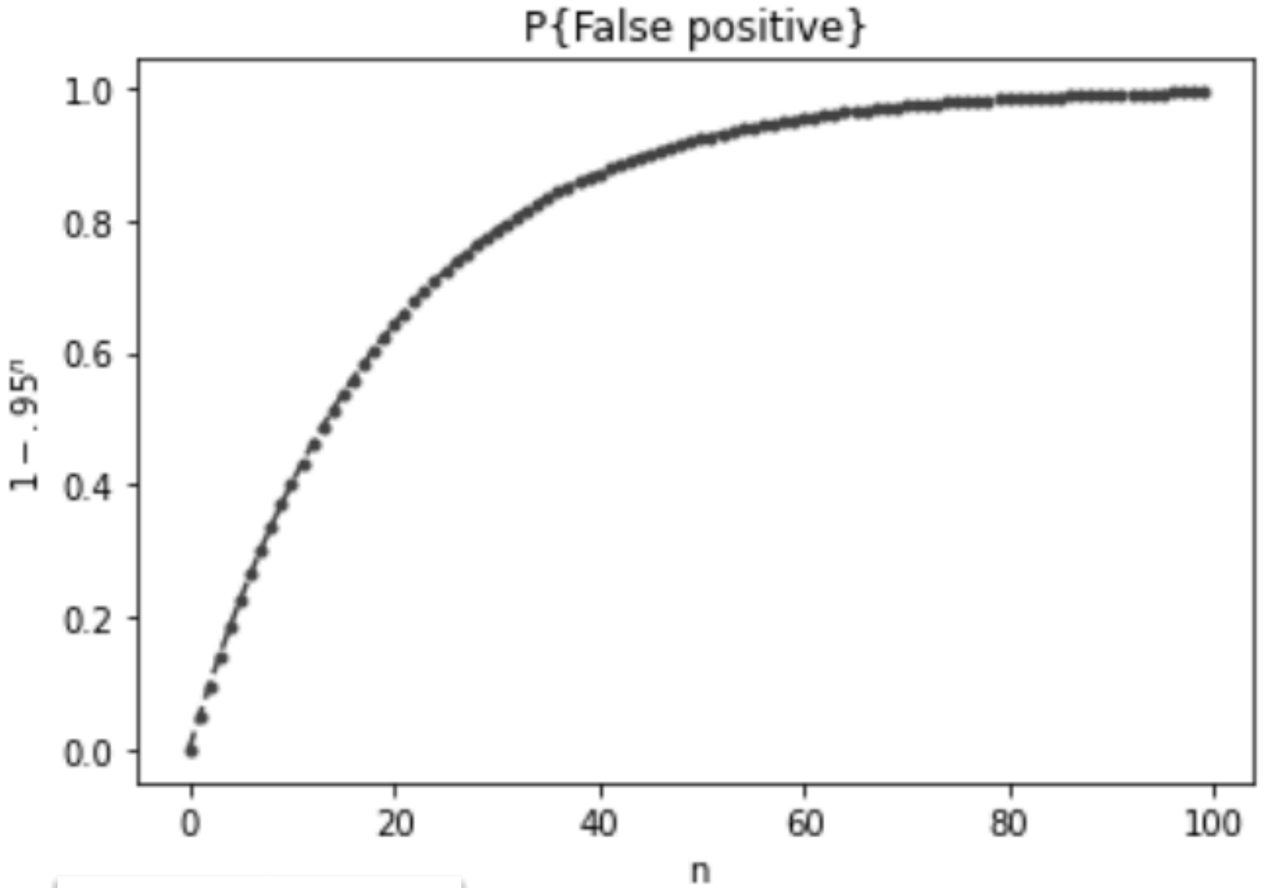
It’s safe to *reject* B at any time for any reason



Early stopping

Why false positives?

- Imagine:
  - D days of experiment
  - B is not better than A, $\bar{z} = 0$
  - Check for $z > 1.64$ periodically
- P{z>1.64 on check n} = p >= .05
- P{z>1.64 on **any** check up through n} >= $1 - (1 - p)^n = 1 - .95^n$



p=.05 when n=N

Early stopping

Consequences

- FP’s damages system
- Need to roll back changes once you realize they don’t work
- Lost time, lost money, etc.
- Other teams need to undo their changes or plans that depended on your false-positive results
- Have seen this enough times to want to warn you.

Running times

System	Business metric	Running time
Agency execution trading	Execution costs	1 week - 1 month
Infrastructure server	Latency	1 hour - 1 day
High-turnover hedge fund	just “safe to deploy”	1 month (small-scale A/B only)
Internet ads, recommender systems	revenue, engagement	1-2 weeks

Ethics

Example experiments

- A new trading strategy might over-message an exchange, disrupting service for all participants
- Say you want to remove posts about suicide and self-harm from a social media feed because they are unpleasant for the viewer. What about the poster?
- Does up-weighting misinformation (ex., elections, covid) encourage engagement? Are there negative side effects?
- If an ML fraud model holds payments for medicine or food, will customers (or fraudsters) suffer?

Ethics

- Controversy: 2021, Facebook ran “emotion contagion” study on users
  - manipulated the emotional content of users’ feeds: If a user sees more sad posts, does the user create more sad posts? (Yes.)
  - Experimented on ~600,000 users
  - Could users have been harmed?
  - Would users approve of having their posts used to make friends and family sadder? That’s not generally considered the intent of posting on Facebook.
  - see: https://www.pnas.org/content/111/24/8788

Ethics

What do you do?

- *Minimal risk*: "… the probability and magnitude of harm or discomfort anticipated in the research are not greater than those ordinarily encountered in daily life or during the performance of routine physical and psychological examinations or tests and that confidentiality is adequately protected. Be aware of ethical questions; include in your design process" [NIMH]
- No IRB in industry, so
  - Seek others' opinions
  - Larger companies might have internal reviewers / process
  - Seek outside counsel