Week 3:
A/B Testing II

AIM-5014-1A: Experimental Optimization

David Sweet // 20230615

Review: Law of Large Numbers

- N observations, y_i, of business metric
- w/mean $\bar{y} = \frac{\Sigma_i^N y_i}{N}$ as $N \to \infty$

$$\bar{y} \to E[y_i]$$

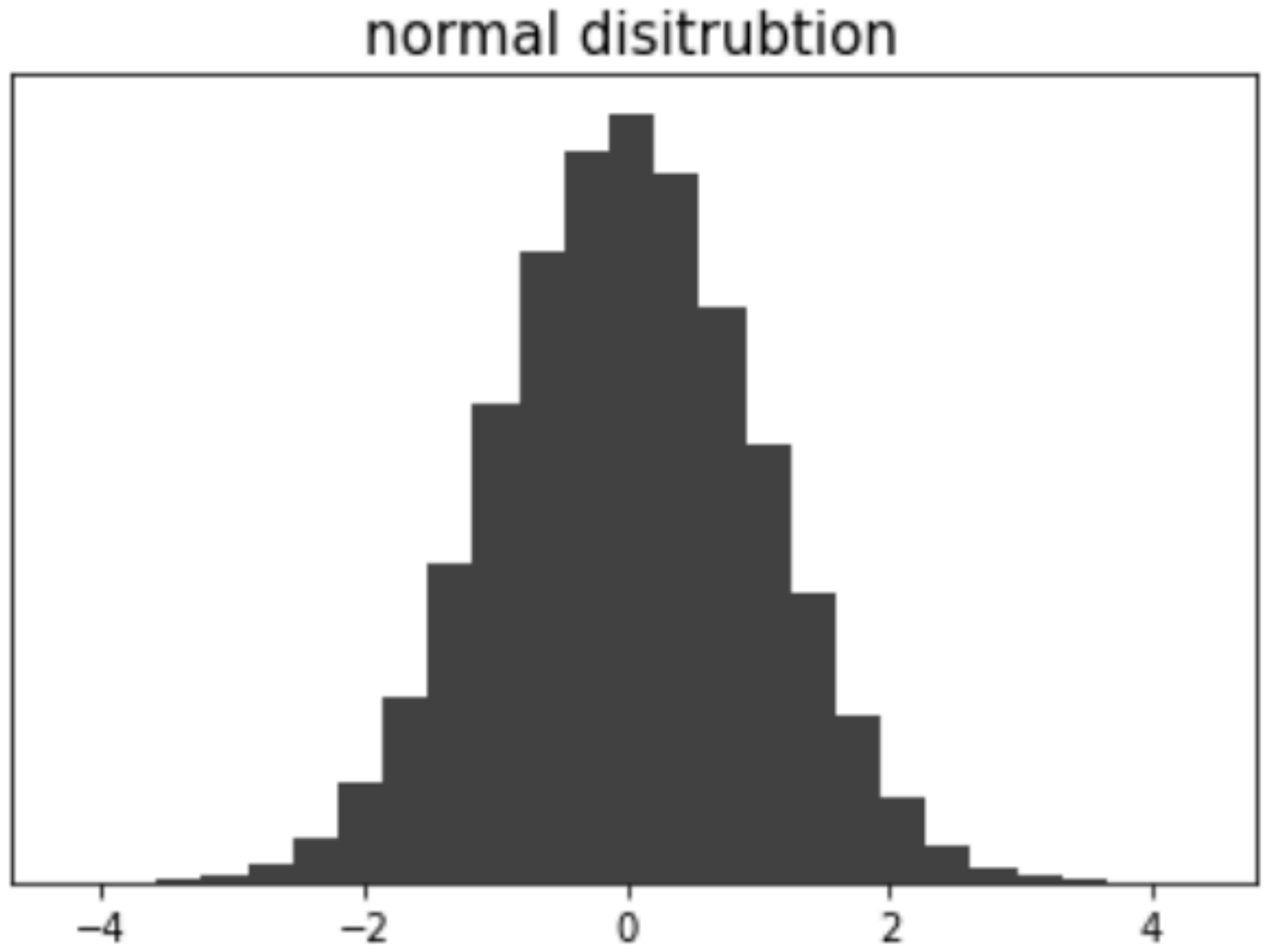
- IOW: Our measurement ($\bar{y}$) estimates the true business metric,
  - which we realize when we run our system for a long time

Review: Central Limit Theorem

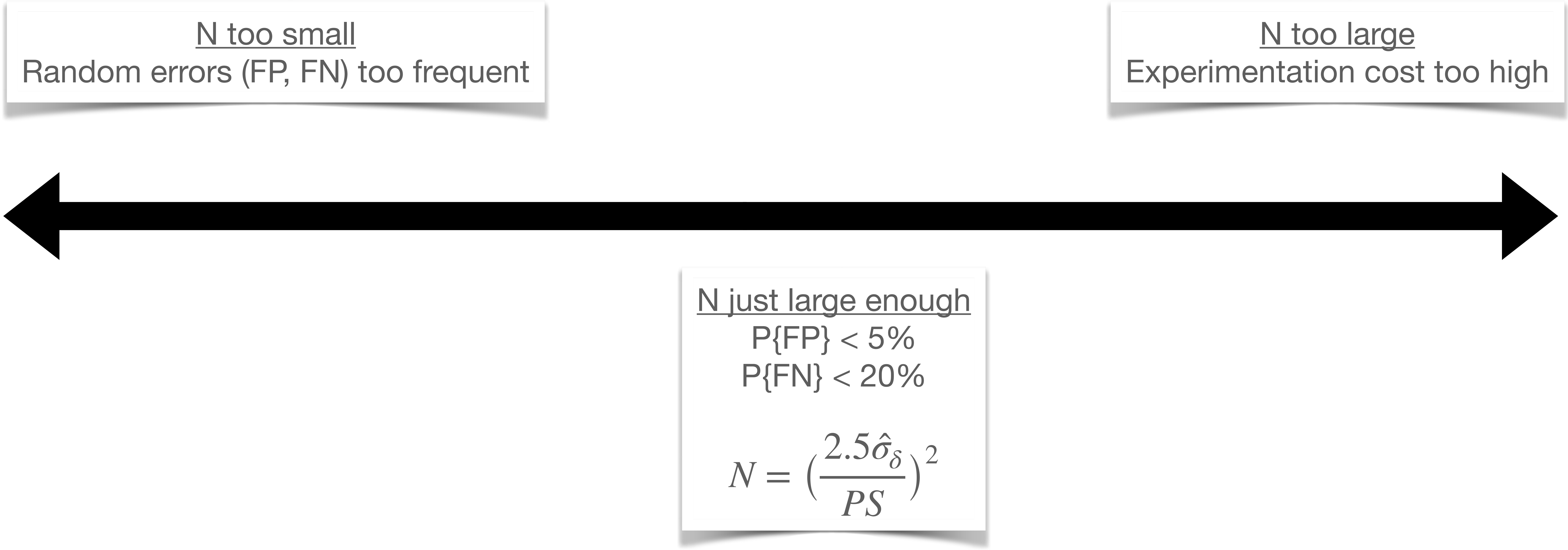
- As $N \rightarrow \infty$

$$\bar{y} \sim \mathcal{N}(E[y_i], \sigma^2)$$

- IOW: Measurement ($\bar{y}$) is appx. normally distributed
    - …even if observations (y_i) are not
    - …when we have enough observations
- Normality makes calculating P{FP} & P{FN} easy



Review: A/B Test



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

N too large
Experimentation cost too high

N just large enough
P{FP} < 5%
P{FN} < 20%

$$N = (\frac{2.5\hat{\sigma}_\delta}{PS})^2$$

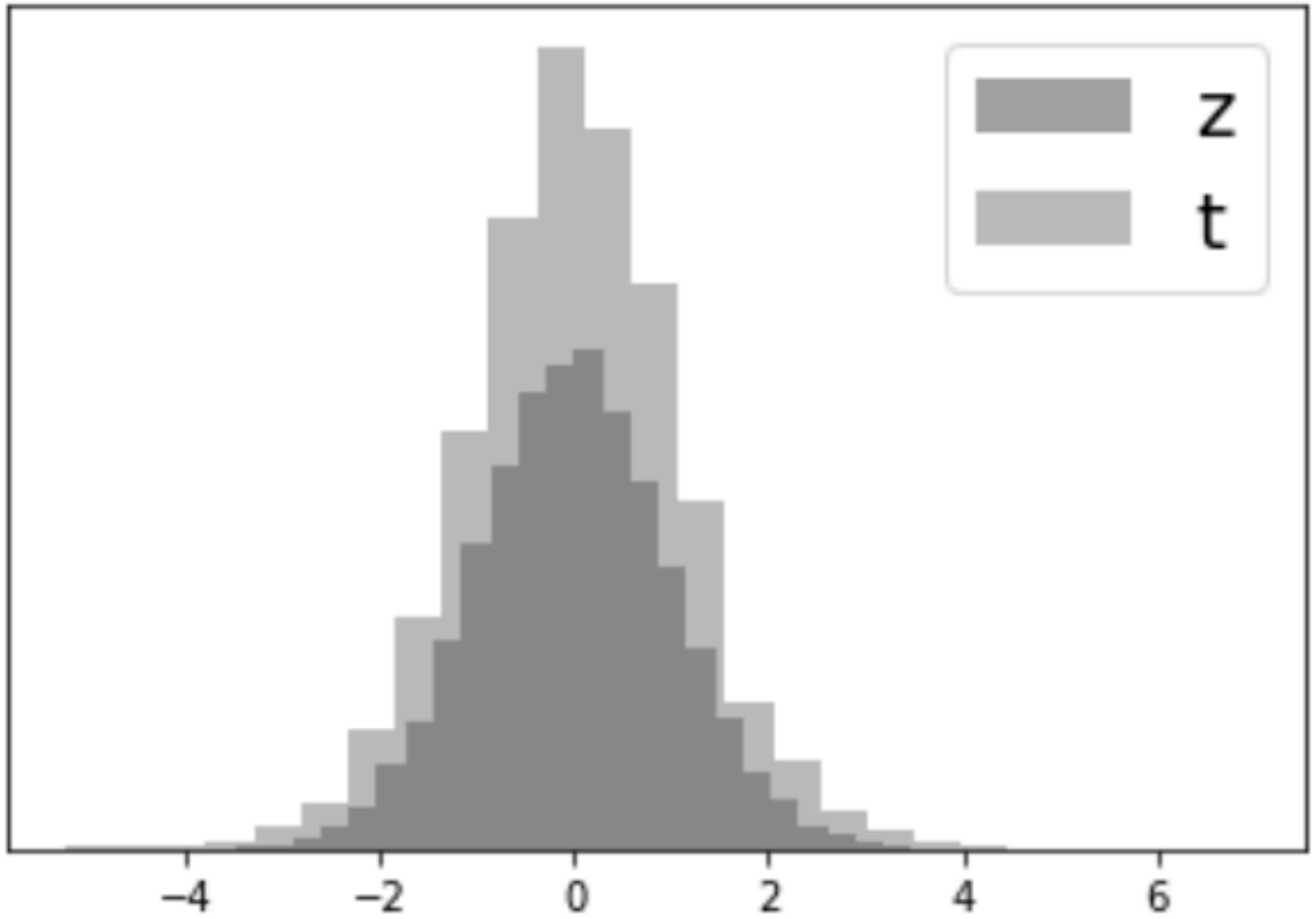
Review: A/B Test

- **Design**: $N \geq \left(\frac{2.5\hat{\sigma}_\delta}{PS}\right)^2$
- **Measure**: Randomize, $\bar{\delta} = \bar{y}_B - \bar{y}_A, \quad se = \sigma_\delta/\sqrt{N}$
- **Analyze**: If $\bar{\delta} > PS$ and $\frac{\bar{\delta}}{se} \geq 1.64$, then accept B.

Aside: t statistic

Student's t

- N large, z-stat (z-score):
  - $z = \frac{\bar{\delta}}{se} \sim \mathcal{N}(0,1)$
- N not large, t-stat: $t = \frac{\bar{\delta}}{se}$
- The uncertainty in se makes t follow "t distribution" (fatter tails)
- Usually called t



Let's say you play the coin-tossing game -- heads you win $1, tails you lose $1 -- with 100 coins simultaneously.

How much do you expect to win?

What if, after playing once, you discard all of the coins that came up tails -- let's say there were 58 of them -- then play the game again with the remaining 42 coins.

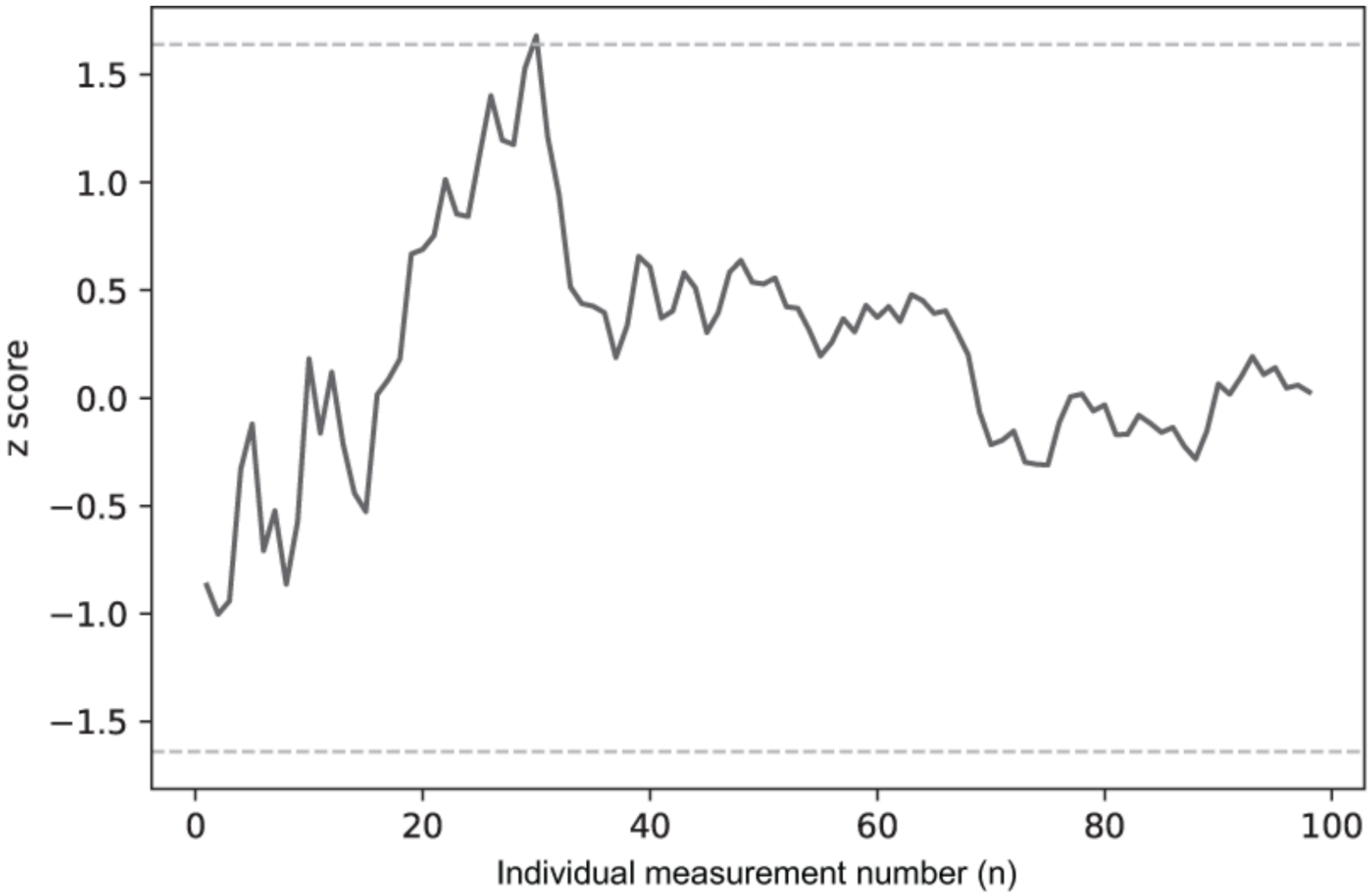
How much do you expect to win this time?

Early stopping

Why false positives?

- Imagine:
  - D days of experiment
  - B is not better than A, $\bar{\delta} = 0$
  - Check for $z > 1.64$ periodically
- P{never seeing z > 1.64 at any n} >> 0.05
- False positive rate increases *dramatically*

Figure 8.2 Early stopping would generate a false positive at `n=30`.



Optimism Bias

- No “good coins”; all noise
- BUT: measurements = signal + noise
- $\bar{y} \sim \mathcal{N}(E[y_i], \sigma^2)$ ==> $\bar{y} = E[y_i] + \sigma\varepsilon$ where $\varepsilon \sim \mathcal{N}(0,1)$
  - $E[y_i]$ is signal, σ is noise level
  - ε is noise (like coins)
- Similarly: $\bar{\delta} = E[\delta_i] + se \times \varepsilon$

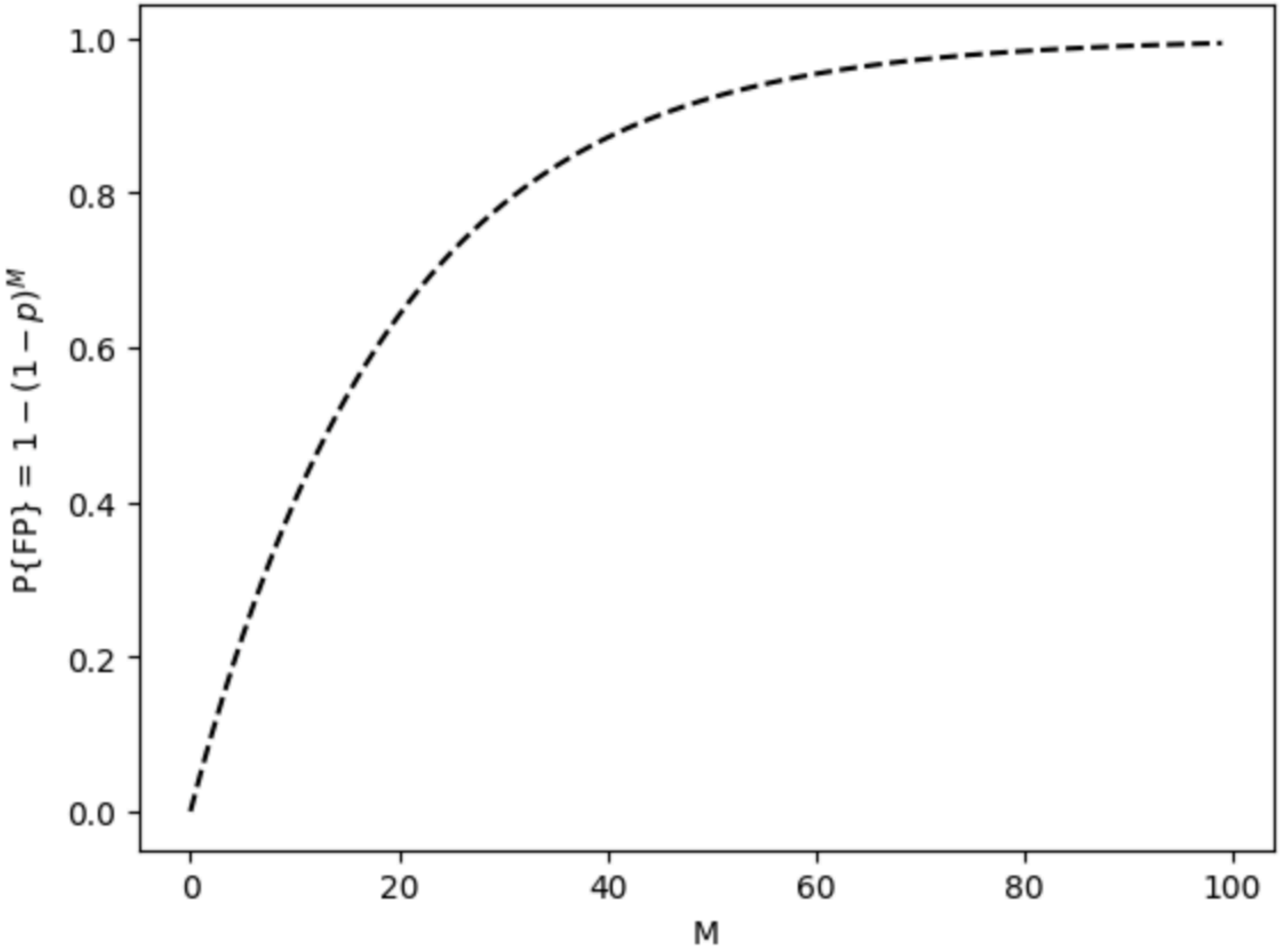
Optimism Bias

- $P\{\bar{\delta} > E[\delta_i]\} = P\{\bar{\delta} < E[\delta_i]\} = \dfrac{1}{2}$ <== like coin flips
- Get heads == \$1 now, expect \$0 later
- Get $t = \dfrac{\bar{\delta}}{se} > 1.64$ now
- Expect $t_{\text{future}} = \dfrac{E[\delta_i]}{se} <$ t later
- Noise + threshold decision rule ==> *optimism bias*

Regression to the Mean

Optimism Bias

- Run M experiments
  - Each w/$P\{FP\} = p = 0.05$
  - $P\{$Any FP$\}$= $1 - (1 - p)^M$
- Similar when repeatedly checking a single experiment



A/B Decisions

- Goal of A/B test is to make a decision
  - Decision: $t > 1.64$ and $\bar{\delta} > PS$
- Usually many metrics (10's) to consider
- Focus on one, main metric for measurement, t-stat
- Use other metrics as *guardrails*
  - $t > 1.64$ and $\bar{\delta} > PS$ and no guardrail metrics worsened

A/B Decisions

- Example:
  - BM: time spent viewing
  - Guardrails: ad revenue, posts viewed, likes
- Example:
  - BM: trading pnl
  - Guardrail: risk, volume, messaging

A/B Decisions

- Might even give up some guardrail metrics for gain in main metric
- Decision might involve multiple parties w/competing priorities

Recap

- Early Stopping **DON'T**
  - Generates lots of false positives.
  - Suffers from optimism bias
- Guardrail metrics
  - Include in final decision
  - Importance varies between interested parties

Let's say you start an A/B test by switching (randomly, of course) 50% of your trades in a trading strategy to version B.

What risks are you taking?

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
  - Monitor guardrail, safety metrics, too
- Should see that the new branch behaves no differently
- Should see that the experimentation 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

Practical/Cavalier A/B Test

- You’ve run multiple A/B tests already
- You know they usually take about a week
- Procedure:
    - Set up arms, Ddeploy safely
    - Run for a week, then wait until $se < \frac{PS}{2}$
    - Helps correct for underestimated N

Practical/Cavalier A/B Test

- Note $se < \frac{PS}{2}$ equivalent to $\frac{PS}{se} < 2$

$2 \approx 1.64$

- NOT t-stat; Ignoring $\bar{\delta}$
- se varies, so we still have some optimism bias, but
  - We waited a week, so optimism/FP level is low enough
  - se much more stable than $\bar{\delta}$
- “Late stopping”, simple

Recap

- Deployment
  - Start small, scale up
  - Monitor main and guardrail metrics for safety
- Cavalier is ok if
  - N is similar from experiment to experiment
  - Err on side of caution: Later of N, $se < \frac{PS}{2}$

A/B/C/… Tests

- You have lots of ideas
    - and the capacity to run multiple arms simultaneously
- Measure versions A, B, C, … all at once.
- Versions called “arms”
    - A/B test has 2 arms
    - A/B/C test has 3 arms

A/B/C/… Tests

- Measure all arms, collect $\bar{y}$'s and se_y's
- Find the best of K arms:
    - Compare A to B w/$t_{A,B} > 1.64$, winner is m_1
    - Compare winner to C w/$t_{m_1,C} > 1.64$, winner is m_2
    - …
    - K-1 steps, best is m_k

A/B/C/… Tests

- Each comparison has $P\{FP\} = p = 0.05$
- Multiple comparisons ==> **high final FP**
  - $P\{\text{Wrong Max}\} = 1 - (1 - p)^{(K-1)}$
  - N.B.: $(1 - p)^n \approx 1 - np$
  - $P\{\text{Wrong Max}\} \approx 1 - (1 - (K - 1)p) = (K - 1)p$
  - (K-1)p > p = 0.05

Binomial approximation

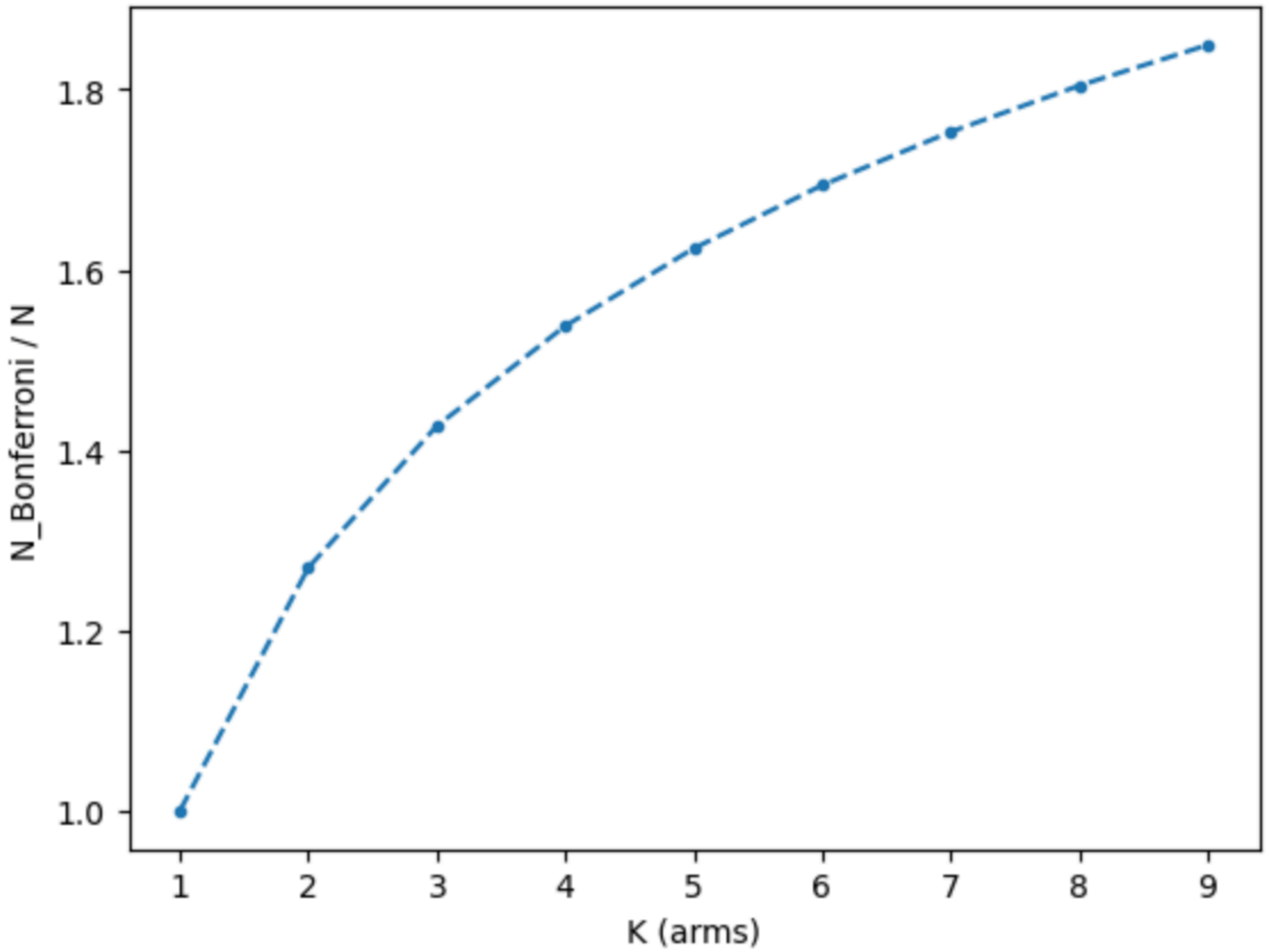
Familywise Error

A/B/C/… Tests

- $P\{\text{Wrong Max}\} \approx (K-1)p$
- Bonferroni correction
    - At design time, set FP limit to $p = \dfrac{0.05}{K-1}$
- $P\{\text{Wrong Max}\} \approx (K-1)\dfrac{0.05}{K-1} = 0.05$
- Usually see: $\alpha = \dfrac{0.05}{K}$ where K counts arms B, C, … (treatments)

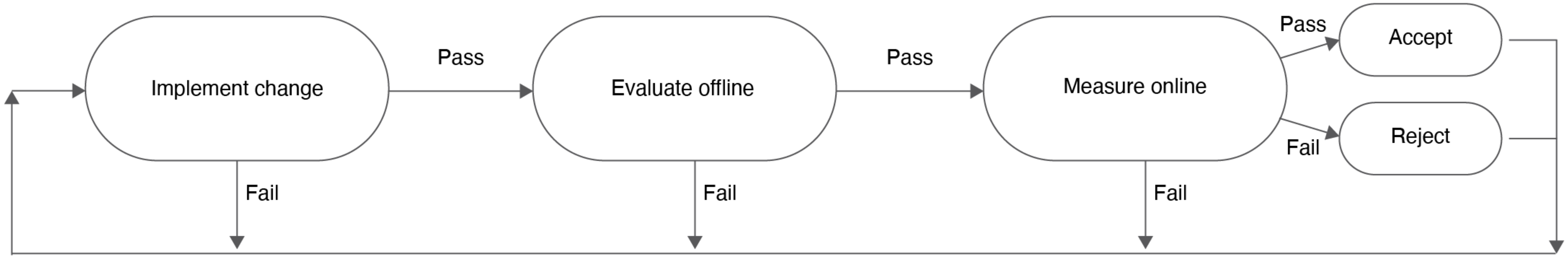
Alternative: Satisficing

- Bonferroni increases N
    - Not dramatically, though
- Alternative, naive approach
    - Select favorite arm, k
    - Run a second A/B test w/just arm k
    - Requires 2N observations in total
        - But simple & get P{FP} <= 0.05



Recap

- You can measure multiple arms simultaneously
  - Bonferroni: Long run can find best arm
  - Naive approach: Run second A/B test



When you run an A/B test on users
— on people —
what additional (non-technical)
risks are you taking?

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. How might this affect a suicidal 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
  https://www.pnas.org/content/111/24/8788
  - manipulated the emotional content of users’ feeds; Asked, 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

Experiment challenges: Ethical

- LinkedIn w/Harvard, Stanford, & MIT ran a study (2017-2022) on 20MM users to test whether weak ties provided better job leads than strong ties [Yes, BTW]
- Could some users have missed out on job opportunities because of this?
- Question was considered
  - Not actually experiments, but advanced observational analysis techniques
  - Ok’d by MIT’s **Institutional Review Board** beforehand
- https://arstechnica.com/tech-policy/2022/09/experts-debate-the-ethics-of-linkedins-algorithm-experiments-on-20m-users/

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

Readings for Week 4

- Chapter 7, *Experimentation for Engineers*
- Chapter 8, *Experimentation for Engineers*
- Present Your Data Like a Pro
  Joel Schwartzberg
  https://hbr.org/2020/02/present-your-data-like-a-pro

Discussion Questions for Week 4

- Where have we used the iid assumption so far in this class?
- What is a holdout test and what is it used for?
- Timeseries data are often autocorrelated. What could give rise to this? Gives examples.

Summary

- “Cavalier” design is ok if you measure conservatively
- Study multiple arms if you have the ideas & capacity
- Consider your experiment’s impact on people
- Deploy safely: start small, scale up
- Don’t stop early
- Use guardrails for monitoring and decision-making