Module 4
A/B Testing III: Practical Concerns

DAV-6300-1: Experimental Optimization

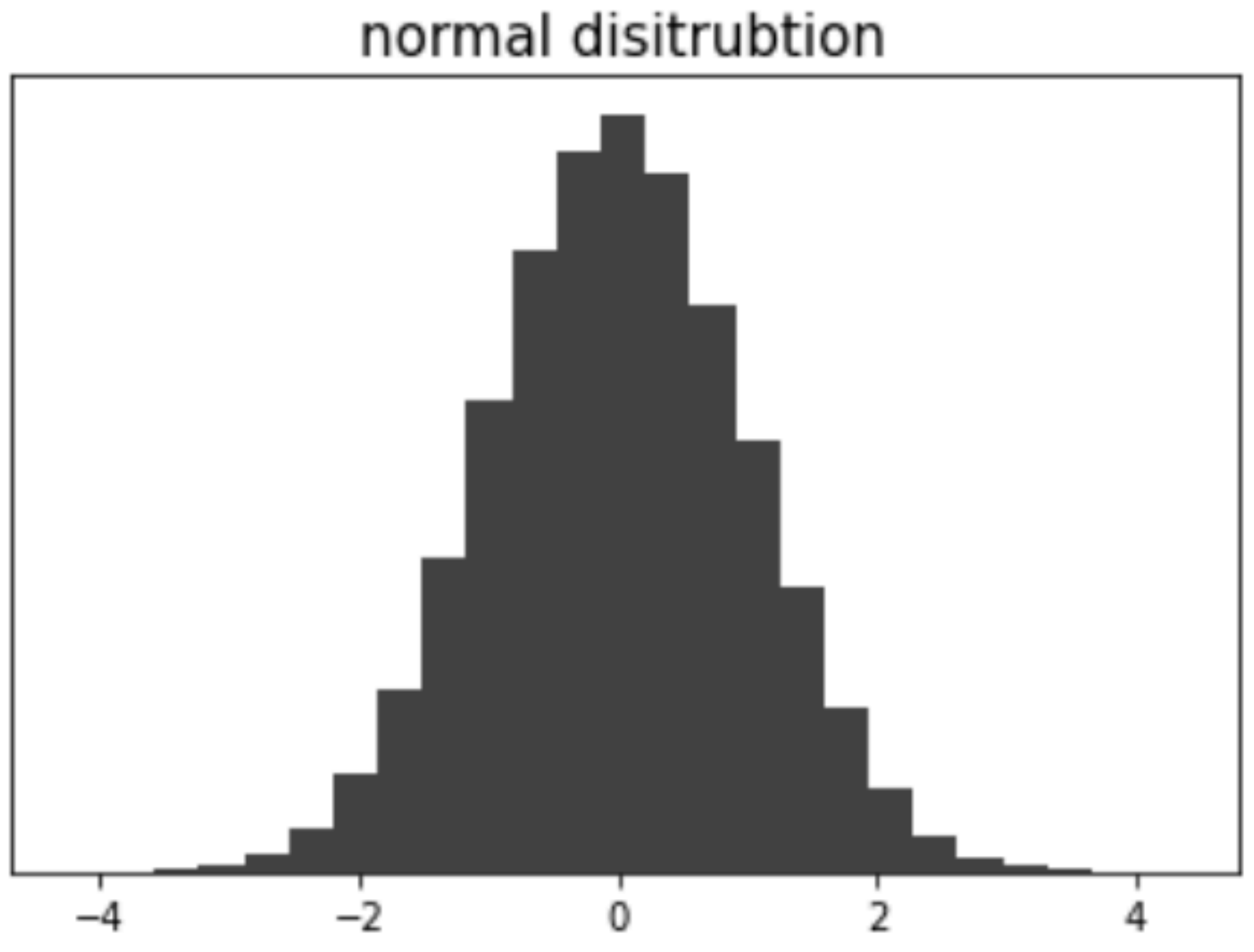
David Sweet // 20240919

Review: Law of Large Numbers

- N observations, y_i, the business metric
- w/mean $\mu = \dfrac{\Sigma_i^N y_i}{N}$
- As $N \to \infty$, $\mu \to E[y]$
- IOW: Our measurement (μ) estimates the "true" business metric

Review: Central Limit Theorem

- As $N \to \infty$, $\mu \sim \mathcal{N}(E[y], VAR[y]/N)$
- IOW: Measurement (μ) is normally distributed
  - …even if observations (y_i) are not
  - …when we have enough observations
- σ = estimates $STD(y)$
- $se = \sigma/\sqrt{N}$ = estimates $STD(\mu)$



$E[\mu]$

Review: A/B Test

- Goal: Accept or reject B
- Design: $N \geq (\frac{2.5\hat{\sigma}_\delta}{PS})^2$
- Measure: Replicate (reduce variance), Randomize (reduce bias)
- Analyze:

Criterion 1: $\delta > 1.6se \ \ (t > 1.6)$

Criterion 2: $\delta > PS$

Key Terms

- Optimism Bias
- Early Stopping
- Familywise error
- Bonferroni Correction

Toss 100 coins simultaneously.

Heads win \$1. Tails lose \$1.

How much do you expect to win?

Discard all coins that came up tails
ex., 58.

Play again with remaining 42 coins.

How much do you expect to win?

Optimism Bias

- Coins:

$$y_i = E[y] \pm \$1$$

$$E[y] = \$0$$

- Decision rule: If heads, coin is a “good coin”.
- False Positive: Thought you had a good coin but didn’t.

Optimism Bias

- Better decision rule:
- $\mu = \sum_{i}^{N} y_i$
  - Say "Good coin" if $\mu > \theta$; $\theta > 0$
  - False positive. (No "good coins". All fair.)
  - Optimism: Overestimate expectation

θ: "theta" for "threshold"

Optimism Bias

- Define "good coin": $E[y] > \$0$

weighted, unfair coin

- How do we tell?
  - A/B test, two coins: $E[y_B] - E[y_A]$
  - Measure: $\delta = \mu_B - \mu_A$
  - $N \geq (\frac{2.5\hat{\sigma}_\delta}{PS})^2$ where $\hat{\sigma}_\delta = \$\sqrt{2}$, $PS = \$0.10$

$N = 1250$ flips

- Decision rule: $\delta > 1.6se$

Optimism Bias

- Decision rule: $\delta > 1.6se$
- $P\{FP\} = 0.05$
- Acceptance is "optimistic"
  - 5% probability $\delta > 1.6se$ just by chance

Sequential Coin flipping

Optimism Bias

Repeated checking

Flip coin. Heads? ==> “It’s a good coin! Stop.”

Flip coin. Heads? ==> “It’s a good coin! Stop.”

Flip coin. Heads? ==> “It’s a good coin! Stop.”

⋮

Optimism Bias

Repeated checking

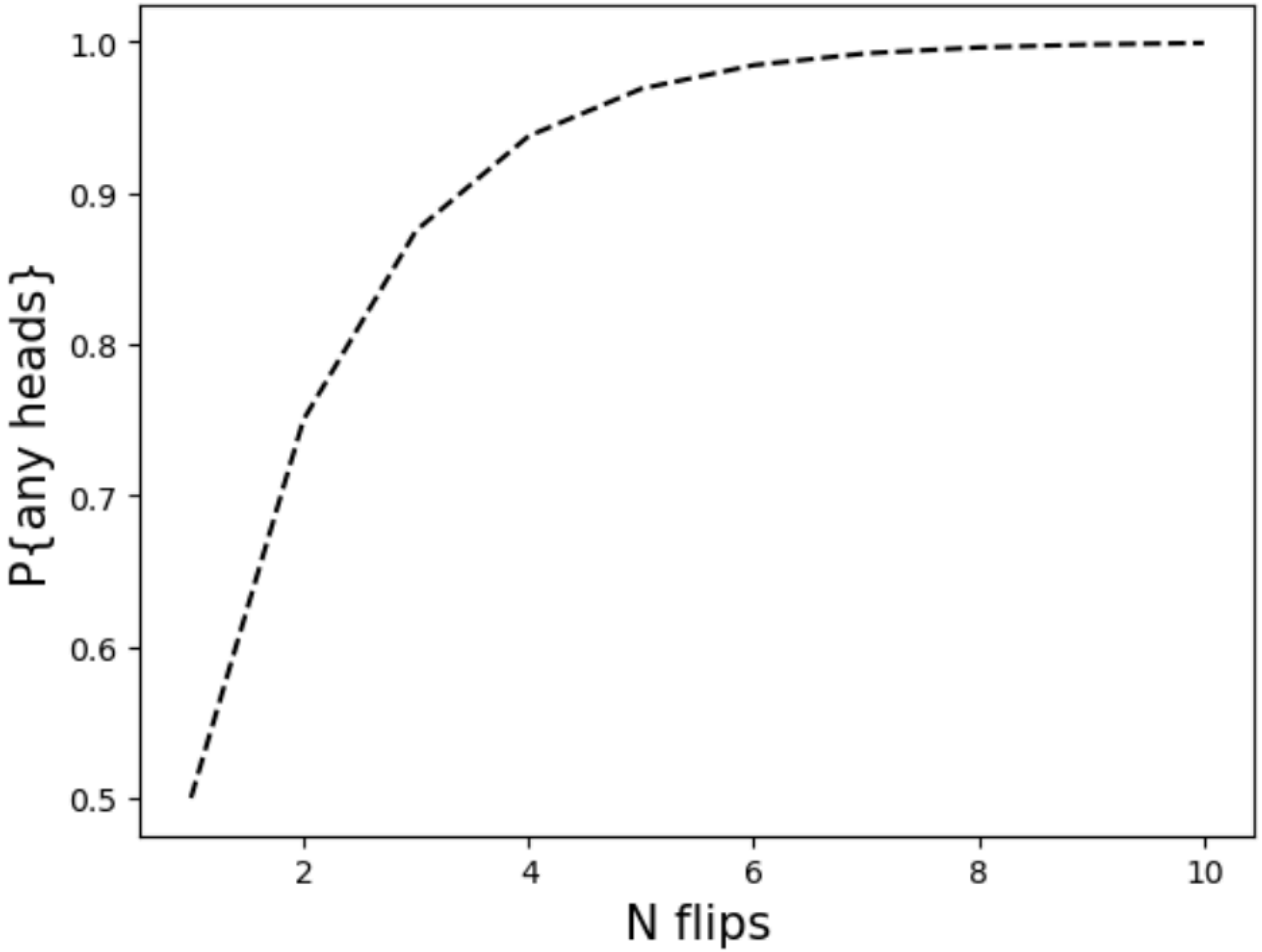
$$P\{\text{any heads} \mid 1 \text{ flips}\} = 1 - \frac{1}{2} = \frac{1}{2}$$

$$P\{\text{any heads} \mid 2 \text{ flips}\} = 1 - (\frac{1}{2})^2 = \frac{3}{4}$$

$$P\{\text{any heads} \mid 3 \text{ flips}\} = 1 - (\frac{1}{2})^3 = \frac{7}{8}$$

$$P\{\text{any heads} \mid 4 \text{ flips}\} = 1 - (\frac{1}{2})^4 = \frac{15}{16}$$

$$\vdots$$



Optimism Bias

Repeated checking

Measure for a day. $\delta > 1.6se$? ==> "B is better. Stop!"

Measure for a day. $\delta > 1.6se$? ==> "B is better. Stop!"

Measure for a day. $\delta > 1.6se$? ==> "B is better. Stop!"

$\vdots$

Optimism Bias

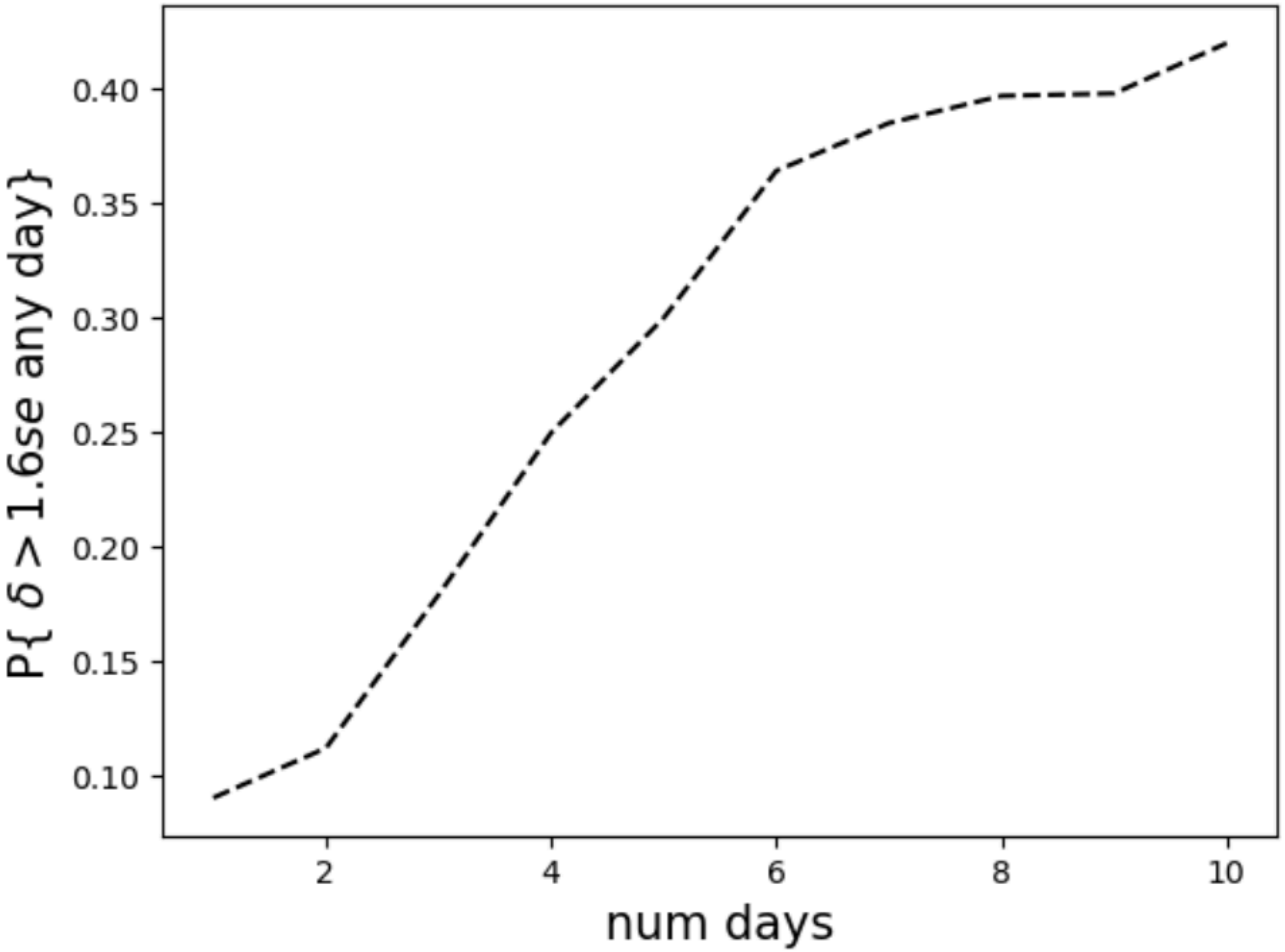
Repeated checking

$$P\{\delta > 1.6se \mid \text{1 day}\} = 1 - p_1$$

$$P\{\delta > 1.6se \mid \text{2 days}\} = 1 - p_1 p_2$$

$$P\{\delta > 1.6se \mid \text{3 days}\} = 1 - p_1 p_2 p_3$$

$$\vdots$$



Optimism Bias

Repeated checking

- Increases false positive rate dramatically
- aka Early stopping
- **BAD. Don't.**
- Take all N observations instead

Cavalier Approach

- $\delta > 1.6se$
  - Bad for stopping
  - Good for picking winner (A or B)

- $PS > 1.6se$
  - Not so bad for stopping.
  - Cannot declare winner
  - Might stop early —
    underestimate of se

Cavalier Approach

- $PS > 1.6se$ same as $se < PS/1.6$
- I.e., just wait until se is small enough
- Approx. same as waiting until N is large enough
  - B/c $se \propto 1/\sqrt{N}$
- In practice: Repeating A/B tests, N & se similar every time.

Cavalier Approach

- In practice
  - Sequence of many A/B tests
  - N similar every time
  - Just start, wait until se small enough (not uncommon)

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

- Deviations from normal behavior?
  - Large change in BM?
  - Large change in *any* metrics?
- New branch behaves no differently
- Experimentation tooling functioning properly
- “Small” is ~1% of N

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 observations to measure precisely, but
  - Look for large, adverse changes in BM
  - Look for large, adverse changes in any metrics

Deploying an A/B test

Full-sized A/B test

- Increase the flow to full scale, collect N observations
- DO: Monitor BM and other 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 regularization technique of the same name

Recap

- Deployment
  - Start small, scale up
  - Monitor main and guardrail metrics for safety
- Cavalier approach ok if
  - N is similar from experiment to experiment

A/B/C/… Tests

- Lots of ideas (A, B, C, …)
- 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

μ_a for arm a

- Measure all arms, collect μ_a's and se_a's
- Find the best of K arms:
  - Compare A to B w/$t_{A,B} > 1.6$, call winner a_1
  - Compare winner to C w/$t_{a_1,C} > 1.6$, call winner is a_2
  - …
  - K-1 steps, best overall is a_k

No. BAD

A/B/C/… Tests

- Each comparison has $P\{FP\} = p = 0.05$
- Multiple comparisons
  - Optimism bias again
  - High final false positive rate
- *Familywise error*

A/B/C/… Tests

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

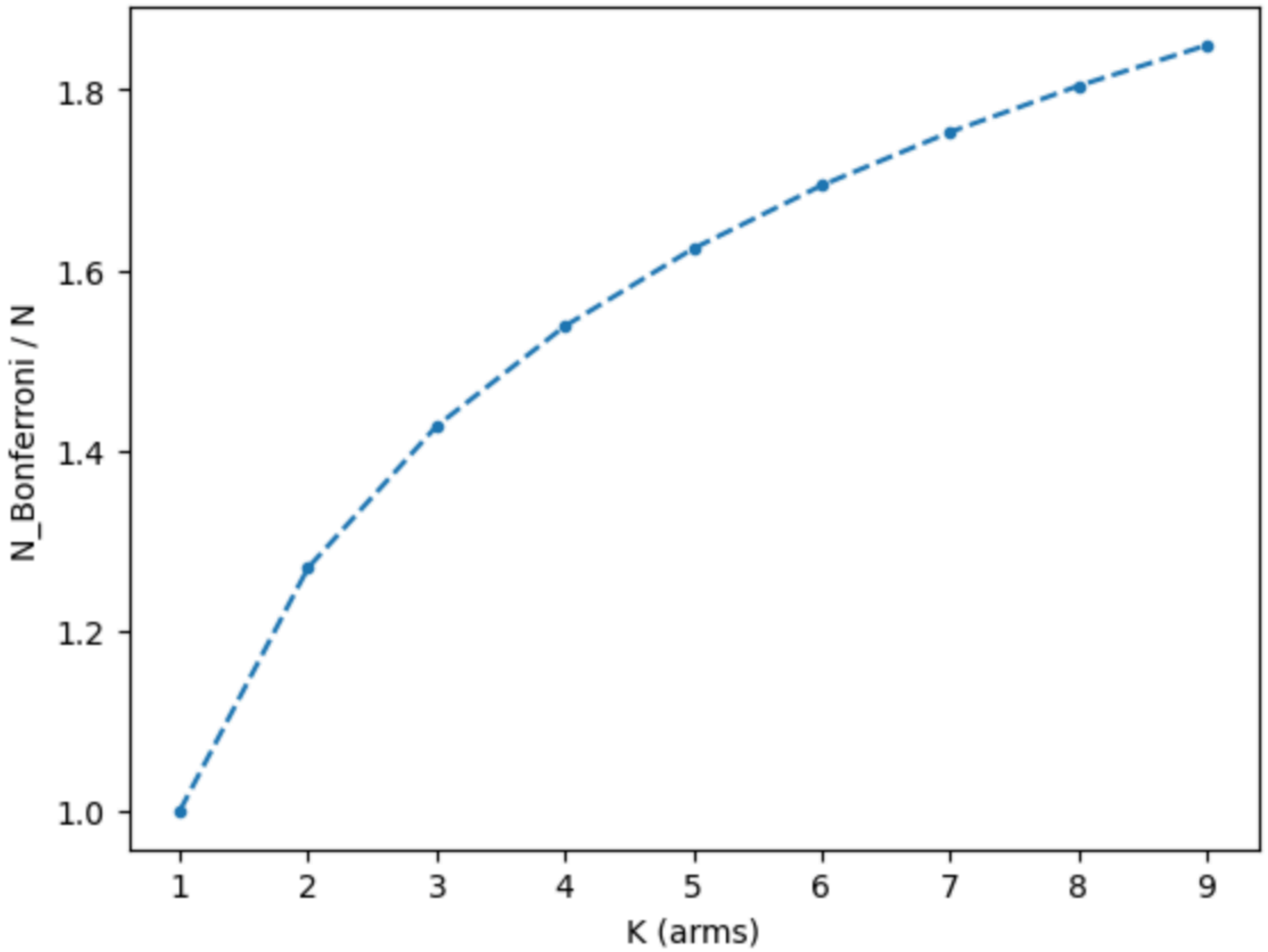
    Binomial approximation
  - $P\{\text{Wrong Max}\} \approx 1 - (1 - (K - 1)p) = (K - 1)p$
  - $(K - 1)p > p$

A/B/C/… Tests

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

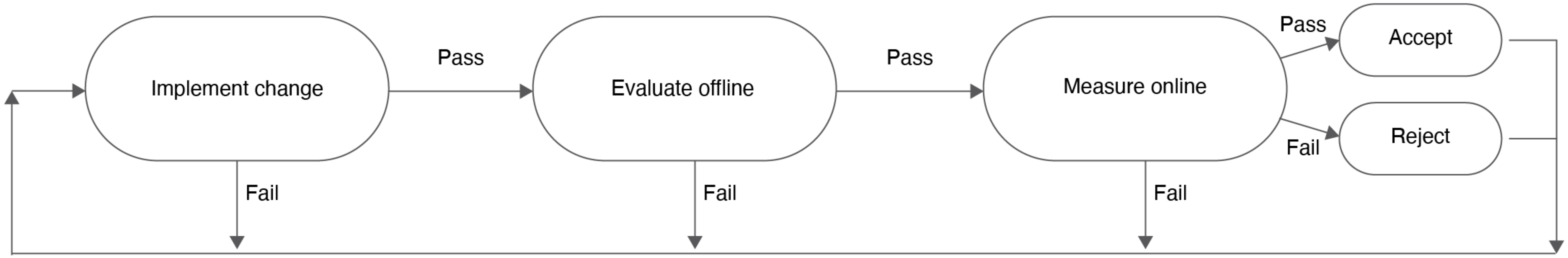
Alternative: Two Experiments

- Bonferroni increases N
  - Not dramatically, though
- Alternative, naïve approach
  - Select favorite arm, a
  - Run a second A/B test w/just A vs. a
  - 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



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 prevents 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