Week 2:
A/B Testing I

AIM-5014-1A: Experimental Optimization

David Sweet // 20230608

Compare mean and expectation.

Coin Flipping

- Flip a coin.
  - Win $1 if heads
  - Lose $1 if tails
- Outcomes (observable): $1, -$1
- Expectation (unobservable): $0
- Mean: $\sum$ outcome_i / N

Measurement Estimates Expectation

- Measurement: *Estimate* of expectation of business metric (BM)

	Business metric	**Values logged**	**Post process**
Social media	Time spent per user per day	user id, date, time spent in a session	sum over sessions, avg. over users & dates
Credit card	P{fraud}	count of transactions, count of fraudulent	[num fraudulent] / [num transactions]
Trading strategy	PnL	trade prices and quantites	sum over returns on dollars held

Measurement Estimates Expectation

- observation / individual measurement
  - time spent by a *specific user* today
  - was *this* transaction fraudulent?
  - *today's* pnl
- Call one observation y_i
- Want to know expectation, $E[y_i]$

Measurement Estimates Expectation

- aggregate measurement == mean of observations
- Call it $\bar{y}$

$$\bar{y} = \frac{\sum_i^N y_i}{N}$$

- mean, $\bar{y}$, estimates expectation, $E[y_i]$
- Can’t observe expectation

Measurement Estimates Expectation

- Law of large numbers:
  - $\bar{y} \to E[y_i]$ as $N \to \infty$
  - Normal system operation: $mean(BM) \to E[BM]$
- Experiment estimates
  - What would normal operation look like if I ran this version of the system?

Measurement Estimates Expectation

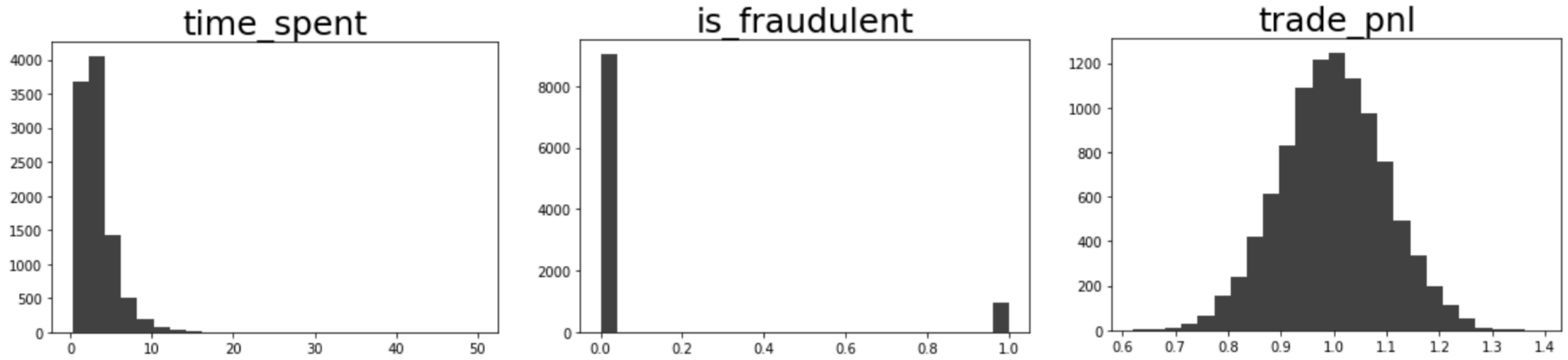
- Example: social media
  - logs record (like, no_like, like, like, no_like)
  - encode as array of y_i: [1, 0, 1, 1, 0]
  - $\bar{y} = (1 + 0 + 1 + 1 + 0)/5 = 3/5 = 0.60$

```
y = np.array([1,0,1,1,0])
y_bar = y.mean()
```

Compare the terms standard deviation and standard error.

Measurements are Uncertain

- Observations / ind. measurements vary from user to user, date to date, session to session, trade to trade, etc.



Measurements are Uncertain

- One measurement: N observations
  - produces a single number, $\bar{y}$
- Run M measurements
  - produce M numbers, $\{\bar{y}_m\}$

Measurements are Uncertain

- Std. dev quantifies uncertainty in $\bar{y}$

$$var[\bar{y}] = \frac{\sum_i^M (\bar{y}_m - \bar{\bar{y}})^2}{M}$$

$$sd = \sqrt{var[\bar{y}]}$$

- where $\bar{\bar{y}}$ is mean of $\{\bar{y}_m\}$
- Too hard. Only want to take *one* measurement, not M.

Measurements are Uncertain

- Estimate variance of $\bar{y}$ by

$$var[\bar{y}] = var\left[\frac{\sum_i^N y_i}{N}\right] = \frac{\sum_i^N var[y_i]}{N^2} = \frac{N\sigma^2}{N^2} = \frac{\sigma^2}{N}$$

- where σ^2 is $var[y_i]$, define

$$se = \sqrt{var[\bar{y}]} = \sqrt{\frac{\sigma^2}{N}}$$

Measurements are Uncertain

- $se = \sqrt{\dfrac{\sigma^2}{N}} = \dfrac{\sigma}{\sqrt{N}}$ is called the *standard error* of $\bar{y}$
- se decreases with N
- NB: Can't observe σ, either, but can estimate with sample std dev, $\hat{\sigma}$

Measurements are Uncertain

- Ex. again: vector of y_i: [1, 0, 1, 1, 0]
  - $\bar{y} = (1 + 0 + 1 + 1 + 0)/5 = 3/5 = 0.60$
  - $\hat{\sigma} = \sqrt{var[y_i]} \approx 0.49$
  - $se = \hat{\sigma}/\sqrt{5} \approx 0.22$

```
y = np.array([1,0,1,1,0])
y_bar = y.mean()
sigma_hat = y.std()
se = sigma_hat / np.sqrt(len(y))
```

Measurement in Brief

- Collect N observations of BM, y_i
- Calculate mean and standard error

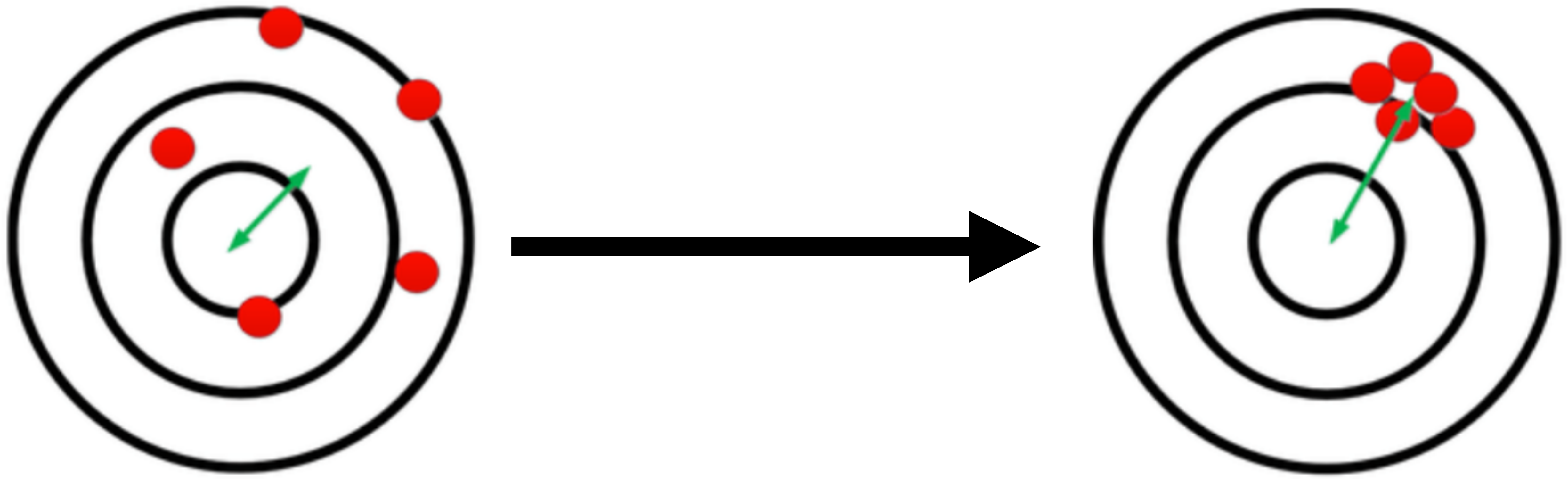
$$\bar{y} = \frac{\sum_i^N y_i}{N}$$

$$\hat{\sigma} = \frac{\sum_i^N (y_i - \bar{y})^2}{N}$$

$$se = \frac{\hat{\sigma}}{\sqrt{N}}$$

Measurements are Uncertain

- *Replication* decreases uncertainty / variance
  - from measurement to measurement



One measurement (one experiment) is one dart

Measure A & B

- A/B test compares BM(A) to BM(B)
- Collect N each of $y_{A,i}$ and $y_{B,i}$; $\delta_i = y_{B,i} - y_{A,i}$

$$\boxed{\bar{\delta} = \bar{y}_B - \bar{y}_A, \quad se = \hat{\sigma}_\delta / \sqrt{N}}$$

- $\hat{\sigma}_\delta = \sqrt{var[\delta_i]} = \sqrt{var[y_{A,i}] + var[y_{B,i}]}$

What is confounder bias?

Measurement: Confounder bias

- Example, credit card fraud detection system, BM = 100% - [% lost to fraud]
    - version A: old ML model
    - version B: new ML model
- A/B test: Collect N observations of BM, $y_{A,i}$ and $y_{B,i}$
- Run in EU: $\bar{\delta} = 0$ ==> B same as A

Measurement: Confounder bias

- But wait, EU has EMV chip card law.
- Chip card law is a *confounder*
- Run in US: $\bar{\delta} > 0$, i.e. $\bar{y}_B > \bar{y}_A$ ==> B wins
- Can't know all possible confounders

Measurement: Selection bias

- Usually don't run A/B test on all users, or all transactions, etc. [Risky]
- Select subset to run on
- Selection bias:
  - Subset not distributed like full population
  - Pop: 40% EU, 60% US
  - Subset: 10% EU, 90% US
- Biased estimate of BM value from subset

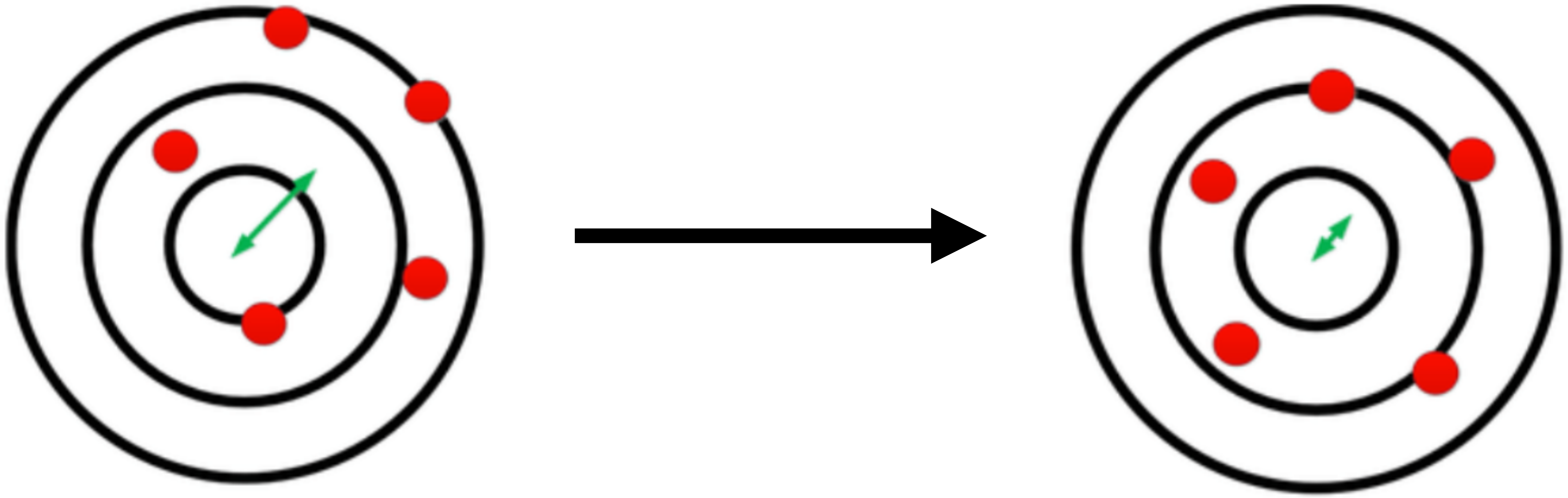
Measurement: Randomization

- Randomly assign each observation to A or B
  - Ex: Transaction enters system, flip coin: Heads use A, Tails use B
  - Ignore US/EU
  - Ignore everything, else too
- Random assignment breaks correlations between confounders and $\bar{\delta}$
- Random selection (from full population) makes subset look like population

Measurement bias

- Randomization decreases bias

One measurement (one experiment) is one dart



Analysis

- After measurement: $\bar{\delta} = \bar{y}_B - \bar{y}_A \qquad se = \hat{\sigma}_\delta / \sqrt{N}$
- Decision time: Accept or Reject B?
- Want to say: "If $E[y_B - y_A] > 0$, accept B."
    - But can't observe expectations

Analysis

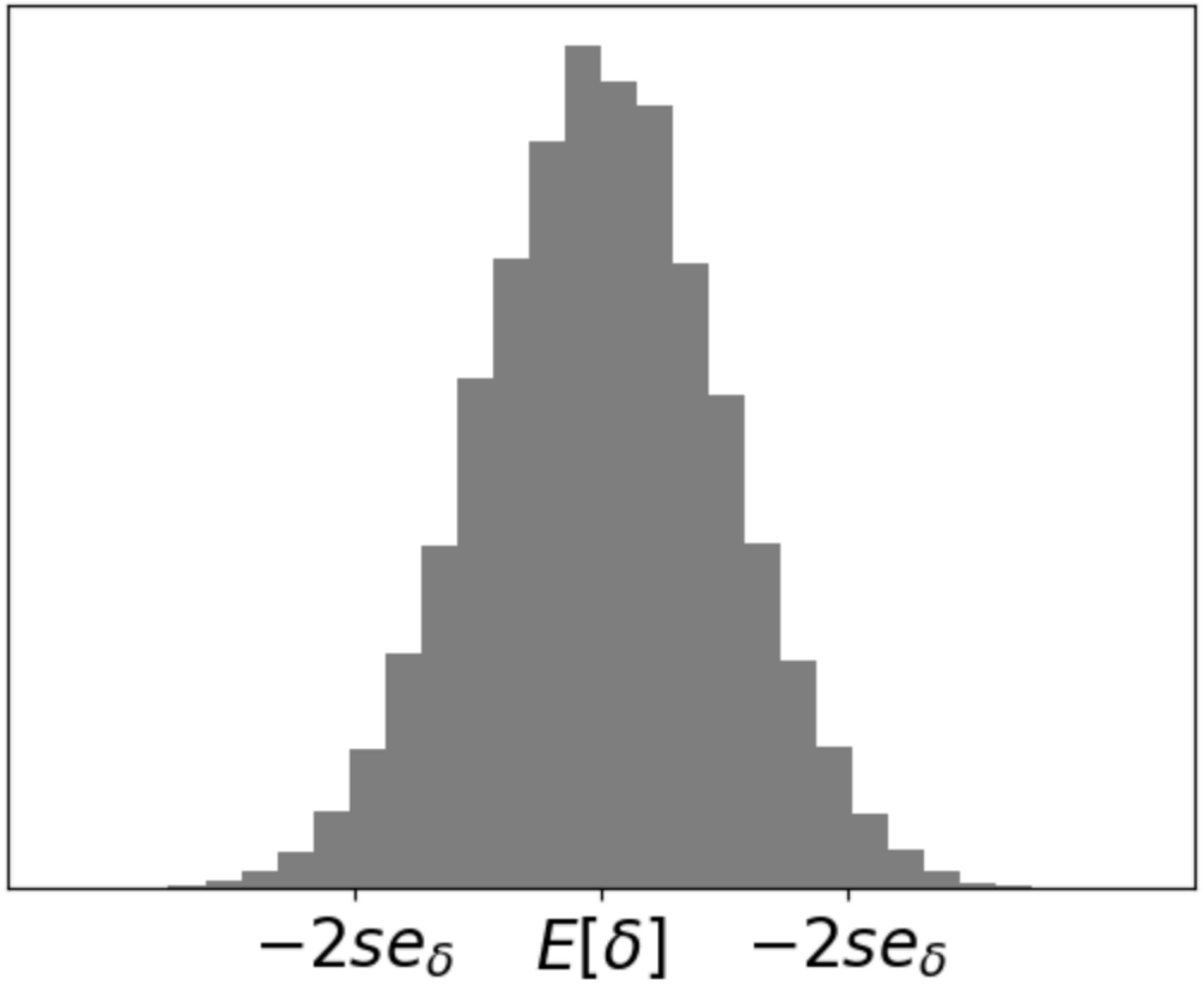
- Instead, ask: If $E[\delta]$ were 0, would it be that $P\{\bar{\delta} > 0\} \leq 0.05$?
  - If "B were the SAME as A"
  - could I have measured what I did ("B better than A")
  - with any meaningful probability (more than 0.05)?
- IOW:
  - Is the probability that this is a false positive (FP) less than 5%?

Analysis

- Asking:
    - “Is the probability — in a HYPOTHETICAL world — of what REALLY happened small enough?”
- Weird.
- FYI: Scientists and engineers often get this wrong.
    - Google “reproducibility crisis”.

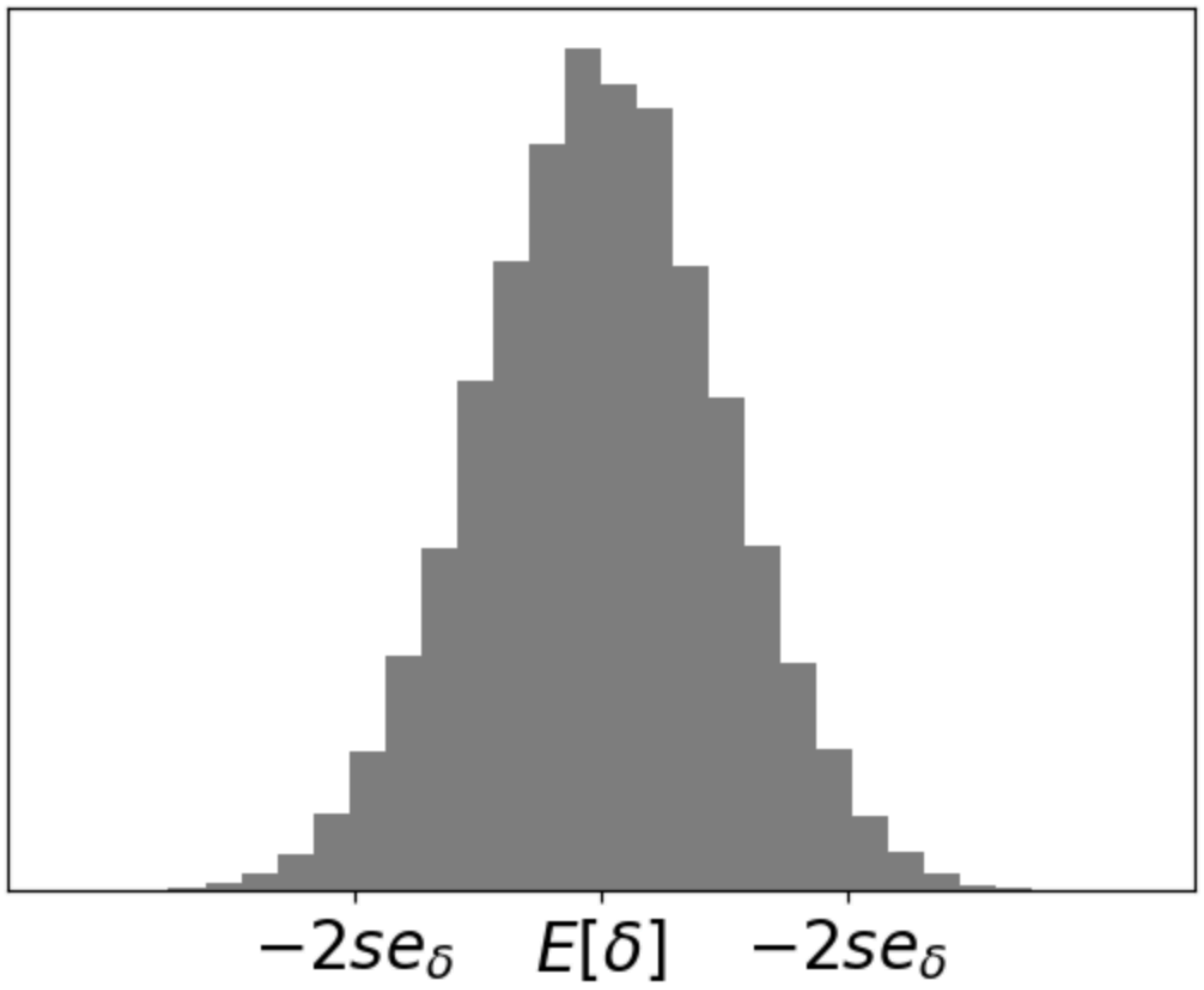
Analysis

- Central Limit Theorem says that $\bar{\delta} \sim \mathcal{N}(E[\delta], var[\delta])$ for large N
- Can't know $E[\delta]$; estimate $var[\delta]$ by se^2



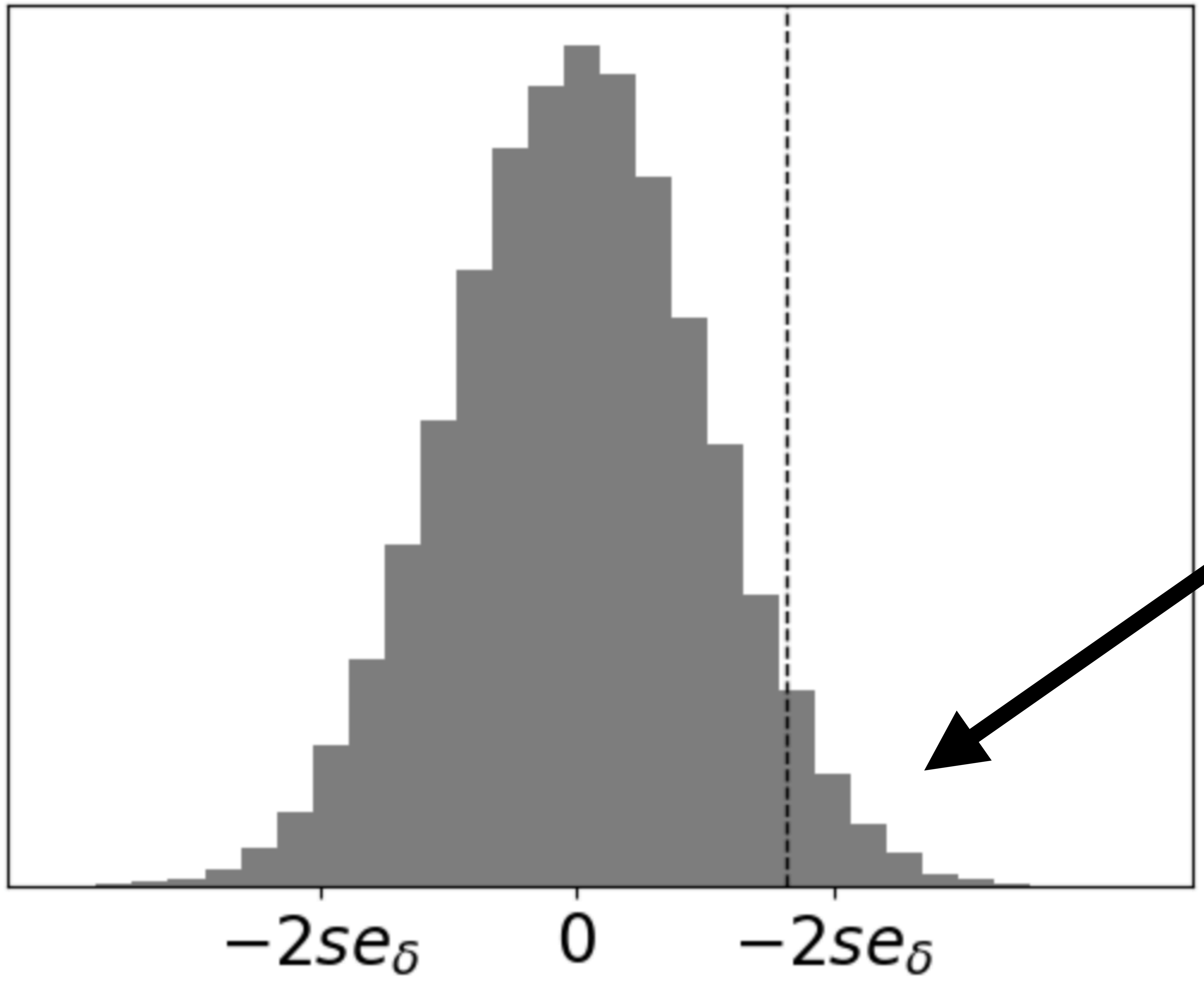
Analysis

- $\bar{\delta}$, one measurement, is a single draw from this distribution



Analysis

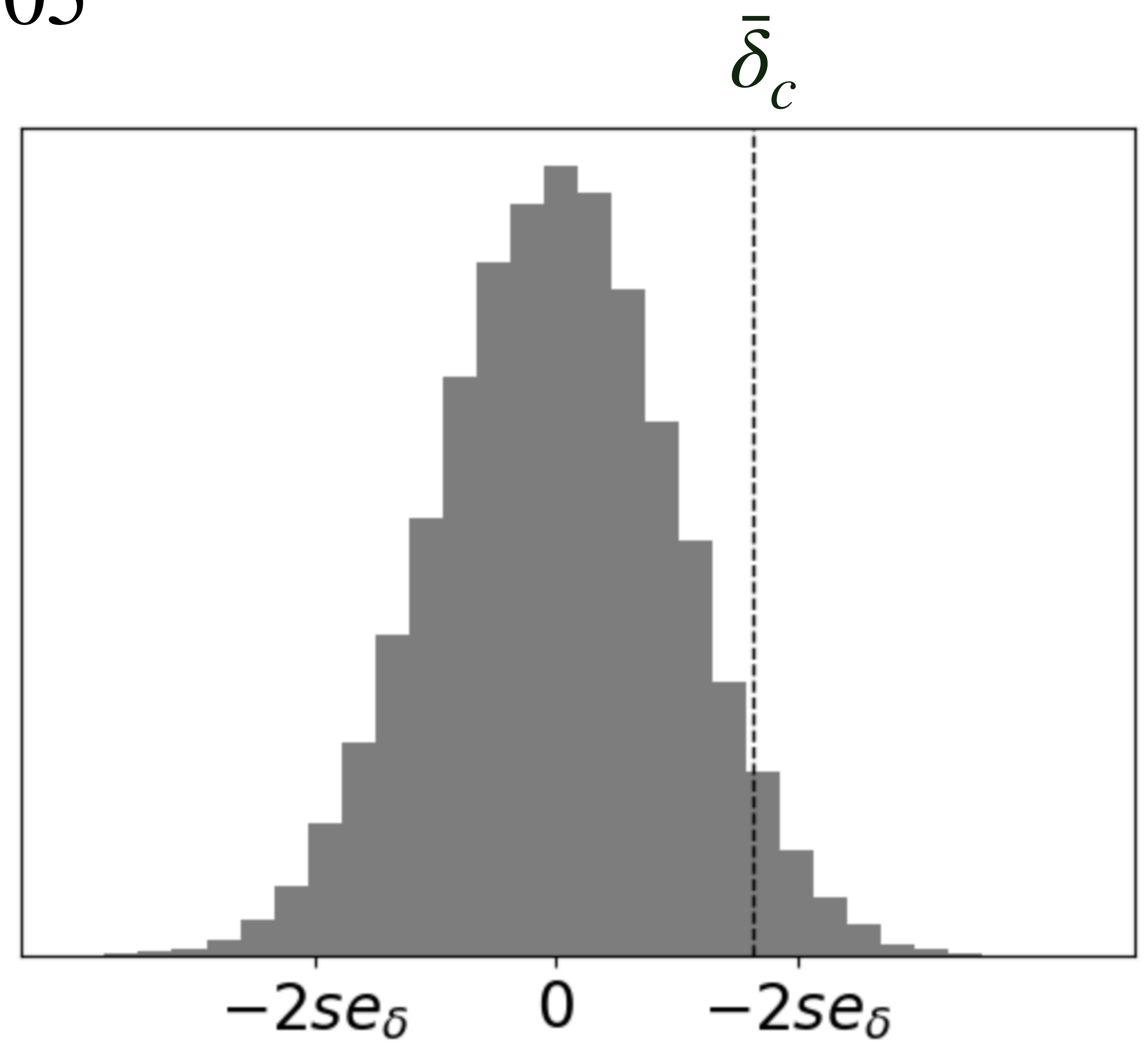
- Say we measured $\bar{\delta}$, w/$\bar{\delta} > 0$. Ask:
  - If $E[\delta] = 0$, what would be the probability of measuring $\bar{\delta} > 0$?



$P\{\bar{\delta} > 0\}$
= area to the right of line

Analysis

- Set a limit: $P\{\bar{\delta} \geq \bar{\delta}_c \,|\, E[\delta] = 0\} \leq .05$
- Find $\bar{\delta}_c$:
    - $0 + k \times se = \bar{\delta}_c$
    - What's k?



Analysis

- z-score table, or

```
scipy.stats.norm().ppf(1-.05)
```

```
1.6448536269514722
```

- $0 + 1.64se = \bar{\delta}_c$

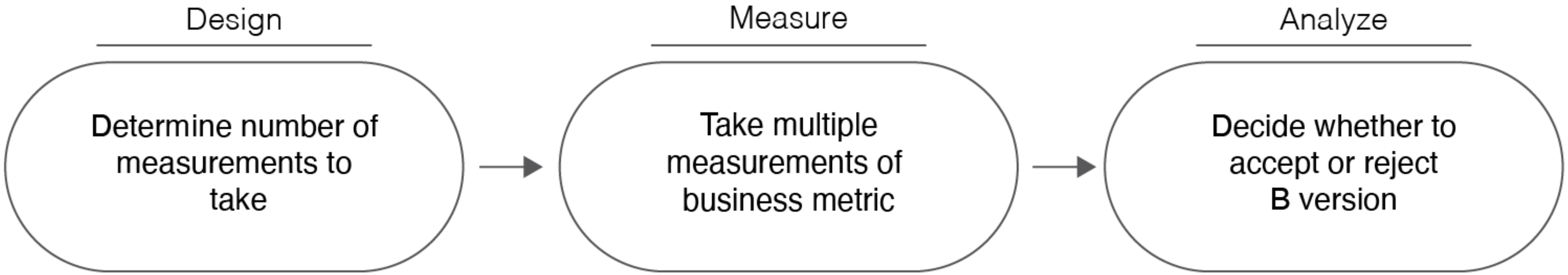
$$\boxed{\frac{\bar{\delta}}{se} \geq 1.64}$$

z	.00	.01	.02	.03	.04	.05	.06
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454
[illegible]	[illegible]	.7611	.7642	.7673	.7704	.7734	.7764
[illegible]	[illegible]	.7910	.7939	.7967	.7995	.8023	.8051
[illegible]	[illegible]	.8186	.8212	.8238	.8264	.8289	.8315
[illegible]	[illegible]	.8438	.8461	.8485	.8508	.8531	.8554
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846

Analysis

- Also ask, “Is $\bar{\delta}$ large enough to care?”
- If trading strategy makes \$1/day more, do you care?
- How about \$1000/day?
- How about \$10,000/day?
- Whatever the number, call it the
    - practical significance (PS) level

Experiment


Design
Determine number of measurements to take
Measure
Take multiple measurements of business metric
Analyze
Decide whether to accept or reject B version

Design

- Determine N
  - Minimize N b/c $, time, risk
  - Limit false positives to 5%.
- Recall check for P{FP} < 5%:
  - $0 + 1.64se = \bar{\delta}_c$ and $se = \sigma_\delta / \sqrt{N}$

Design

- Could solve for N
- But don’t know $\bar{\delta}_c$ or σ_δ
- Replace $\bar{\delta}_c$ with PS:
  - b/c you want to measure effects at least as large as PS
- Estimate σ_δ

Design

1. Estimate σ_δ from logs
    - $\sigma_\delta^2 \approx var[y_{A,i}] + var[y_{B,i}]$
    - Don't have B in logs, but $var[y_{B,i}] \approx var[y_{A,i}]$
2. Or, run *pilot study*
    - Run B in prod for a short time to estimate $var[y_{B,i}]$

- Either way: $\hat{\sigma}_\delta = \sqrt{v\hat{a}r[y_{A,i}] + v\hat{a}r[y_{B,i}]}$

Design

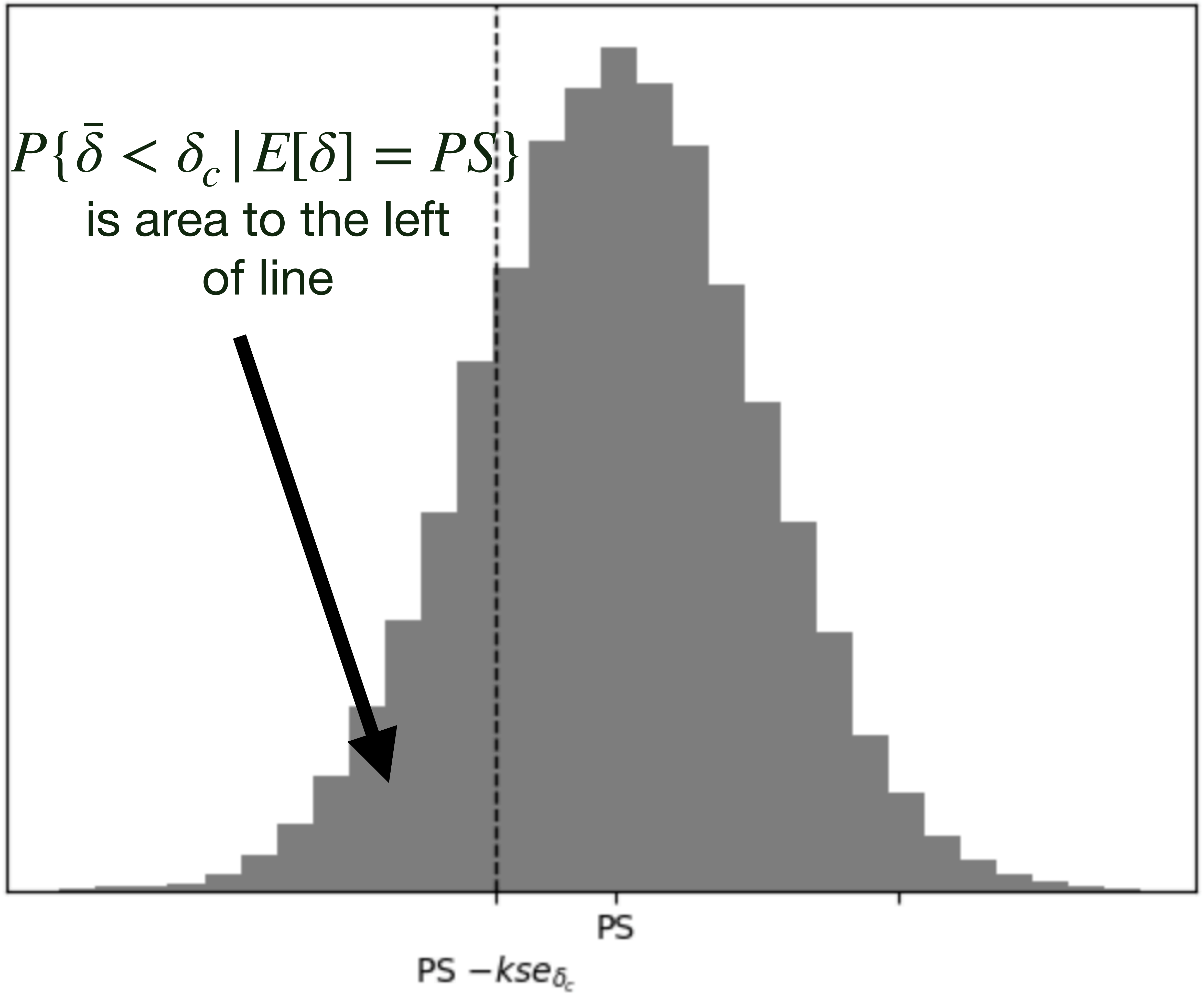
- One more thing… False negatives (FN)
- If $E[\delta] = PS$ *just big enough* to care
  - What would be probability we'd reject?
  - IOW, $P\{\bar{\delta} < \delta_c\}$?
- Limit this to $P\{\bar{\delta} < \delta_c \,|\, E[\delta] = PS\} \leq 0.20$

Design

```
scipy.stats.norm().ppf(.20)
```

```
-0.8416212335729142
```

- $PS - 0.84se = \bar{\delta}_c$
- Same $\bar{\delta}_c$ as earlier



Design

- $PS - 0.84 \times se = \bar{\delta}_c$ and $0 + 1.64se = \bar{\delta}_c$
- $\bar{\delta}_c = 0 + 1.64se = PS - 0.84se$, or

$$PS \approx 2.5se$$

- Now sub. in $se = \hat{\sigma}_\delta / \sqrt{N}$ and solve for N

$$\boxed{N_c = \left(\frac{2.5\hat{\sigma}_\delta}{PS}\right)^2}$$

Length of Experiments

- $N_c = \left(\frac{2.5\hat{\sigma}_\delta}{PS}\right)^2$
- $\hat{\sigma}_\delta$ is the "noise level"
- What happens to N if the noise level doubles?
- What happens to N is the PS level halves?
- As a product matures, what happens to N?

Length of Experiments

- Typical timescales
  - Tech. Products: 3 days - 2 weeks
  - HFT: 1-2 weeks; maybe 1 month
- Two views:
  - Given PS, how long do I have to wait?
  - Given a limited time, what's the smallest PS I can measure?

Terminology

- $\alpha = P\{FP\} = .05$
- $\beta = P\{FN\} = .20$
- False positive also called Type I error
- False negative also called Type II error
- Power = $1 - \beta = .80$ = P{True positive}
- Individual measurement: trial, sample, observation, replicate
- A/B test == Randomized Controlled Trial (RCT) == Controlled experiment
- A = control, B = treatment, PS = minimum effect size

Asymmetry in Limits

- $P\{FP\} \leq 0.05$
- $P\{FN\} \leq 0.20$
- Why use different limits?
- FP degrades BM, a real cost
- FN leaves BM same, an opportunity cost

Readings for Week 3

- Chapter 2 from Experimentation for Engineers (still)
- How do people actually operationalize ML in 2022?
  Josh Tobin
  https://gantry.io/blog/papers-to-know-20221207/
- Lecture 9: Ethics
  Charles Frye
  https://fullstackdeeplearning.com/course/2022/lecture-9-ethics/

Discussion Questions for Week 3

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

Summary: Experiment

- **Design**: $N \geq \left(\frac{2.5\hat{\sigma}_\delta}{PS}\right)^2$
    - Limits: $P\{FP\} \leq 0.05, P\{FN\} \leq 0.20$
- **Measure**: $\bar{\delta} = \bar{y}_B - \bar{y}_A, \quad se = \sigma_\delta/\sqrt{N}$
    - Randomize to reduce bias, replicate to reduce variance
- **Analyze**: If $\bar{\delta} > PS$ and $\frac{\bar{\delta}}{se} \geq 1.64$, then accept B.