Module 3:
A/B Testing II: Design, Measure, Analyze

DAV-6300-1: Experimental Optimization

David Sweet // 20240912

Review: Business Metric (BM)

- Observe business metric, y_i
- Ex: incident of fraud, clicking “like”, sharing a video, clicking on an ad, skipping a song, swiping left or right, lingering on a photo, watching a video until the end
- Ex, daily: revenue, active users, shares traded, pnl

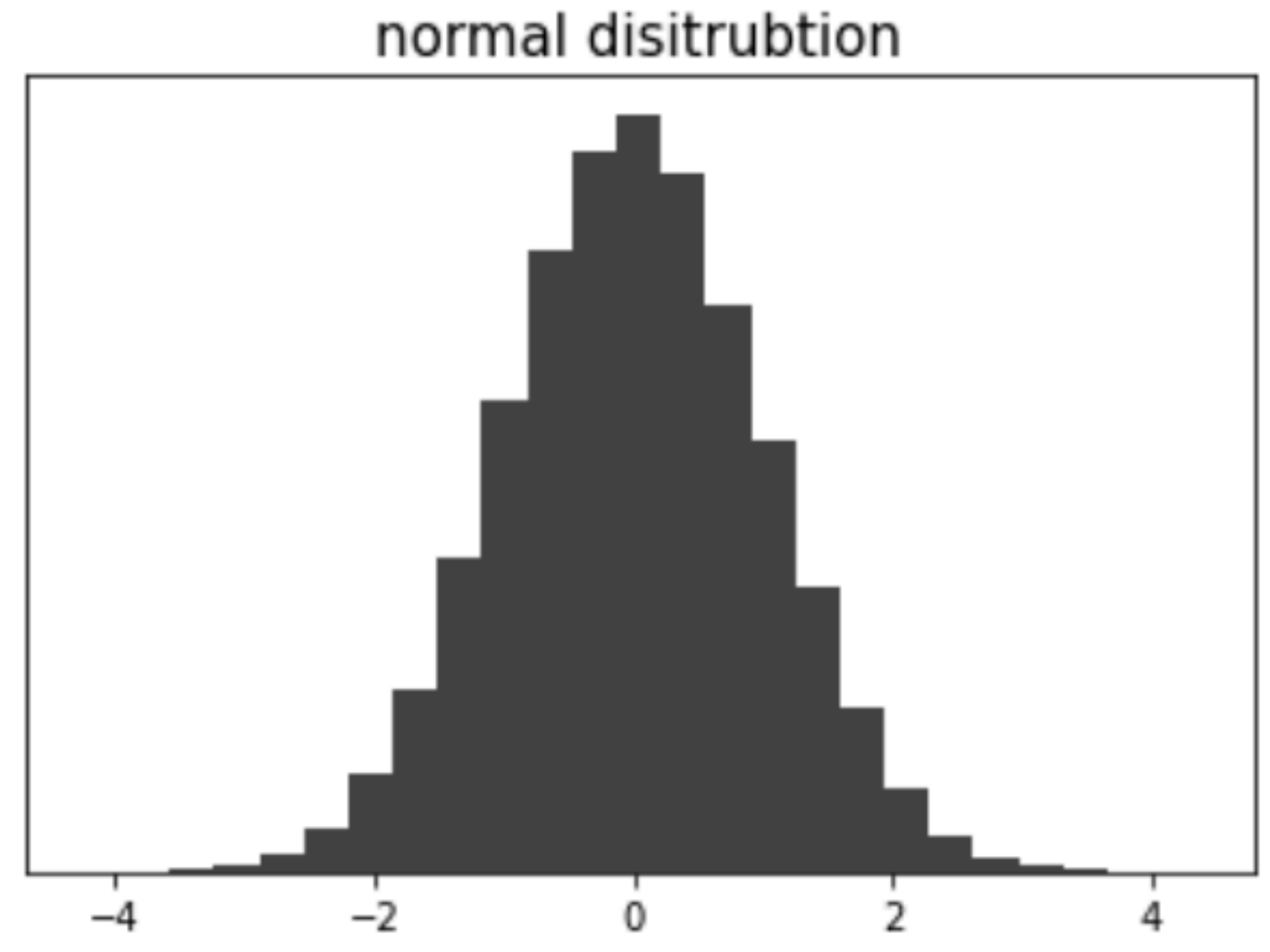
Examples?

Review: Law of Large Numbers

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

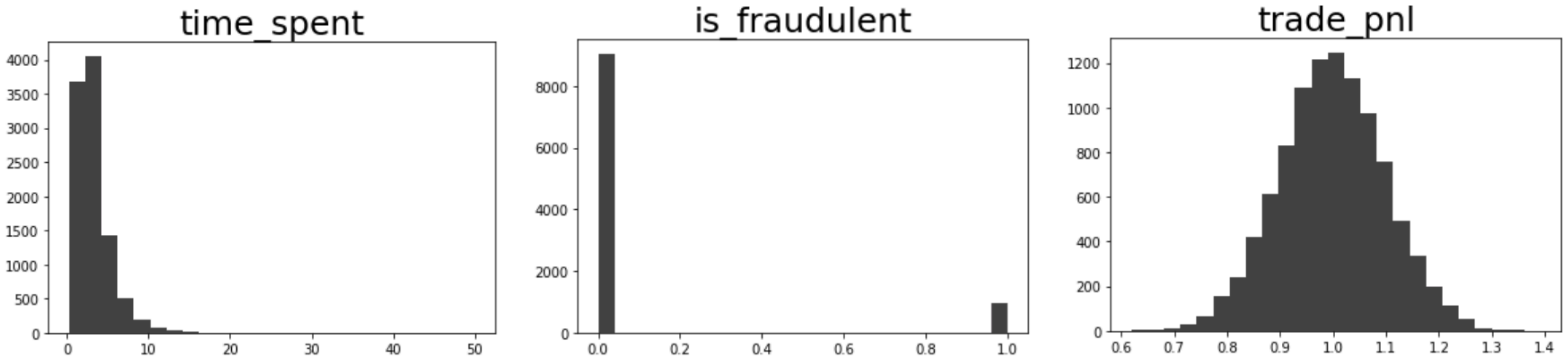
Review: Central Limit Theorem

- 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



Review: Central Limit Theorem

- Observations, y_i, may have any distribution:



- still, $\mu \sim \mathcal{N}(E[y], VAR[y]/N)$ for large N

$\mu = y$ when $N = 1$

Measurement

- Ex. vector of y_i: [1, 0, 1, 1, 0] — 1 == clicked on ad, 0 == ignored

$$\mu = (1 + 0 + 1 + 1 + 0)/5 = 3/5 = 0.60$$

$$\sigma = \sqrt{\frac{(1 - .60)^2 + (0 - .60)^2 + (1 - .60)^2 + (1 - .60)^2 + (0 - .60)^2}{5}} = 0.49$$

$$se = 0.49/\sqrt{5} \approx 0.22$$

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

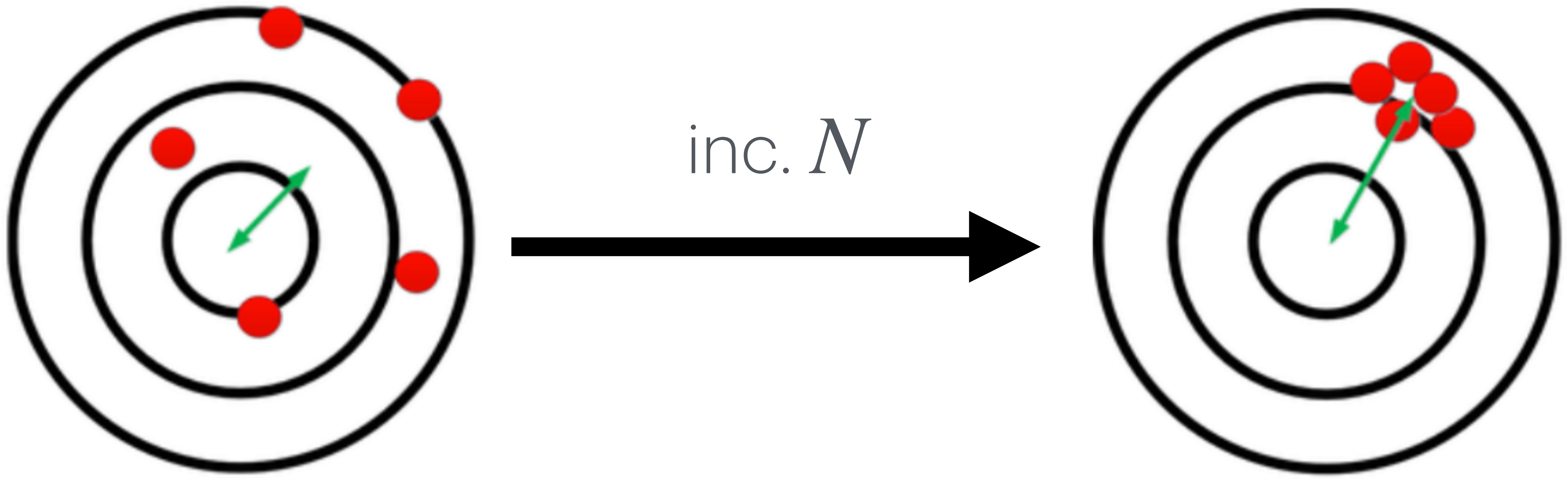
Key Terms

- Replication
- Randomization
- Design, Measure, Analyze
- False positive, False negative

Replication

- *Replication* decreases uncertainty / variance
  - Increasing N decreases se

$$se = \sigma/\sqrt{N}$$



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 $y_{B,i}$

$$\boxed{\delta = \mu_B - \mu_{A'} \quad se = \sigma_\delta / \sqrt{N}}$$

- $\sigma_\delta = \sqrt{\sigma_A^2 + \sigma_B^2}$

Measurement Bias

$$\frac{\mu - E[y]}{STD[\mu]} \sim \mathcal{N}(0,1)$$

- Bias is $\mu - E[y]$
- Unobservable … Yet controllable!
- Different types of biases

$$E[\mu] = E[y]$$

Measurement: Confounder bias

- Example, credit card fraud detection system, BM = 100% - [% lost to fraud]
  - Version A: Old ML model
  - Version B: New ML model
- For simplicity, run A in US and B in EU
  - Only have to change one config file

Measurement: Confounder bias

- But wait, EU has EMV chip card law
  - Measure BM(B) - BM(A) > 0
- Chip card law is a *confounder*
- Solution:

A,EU	A,US
B,EU	B,US

- What if there are 2 confounders? 3? 4?

Did you measure BM(B) - BM(A), or BM(EU) - BM(US)?

Impossible to know all confounders

Measurement: Randomization

- Randomly assign *each observation* to A or B
  - Ex: Transaction enters system, flip coin: Heads use A, Tails use B
  - Irrespective of US/EU
  - Irrespective of everything
- Random assignment breaks correlations between confounders and assignment to A or B

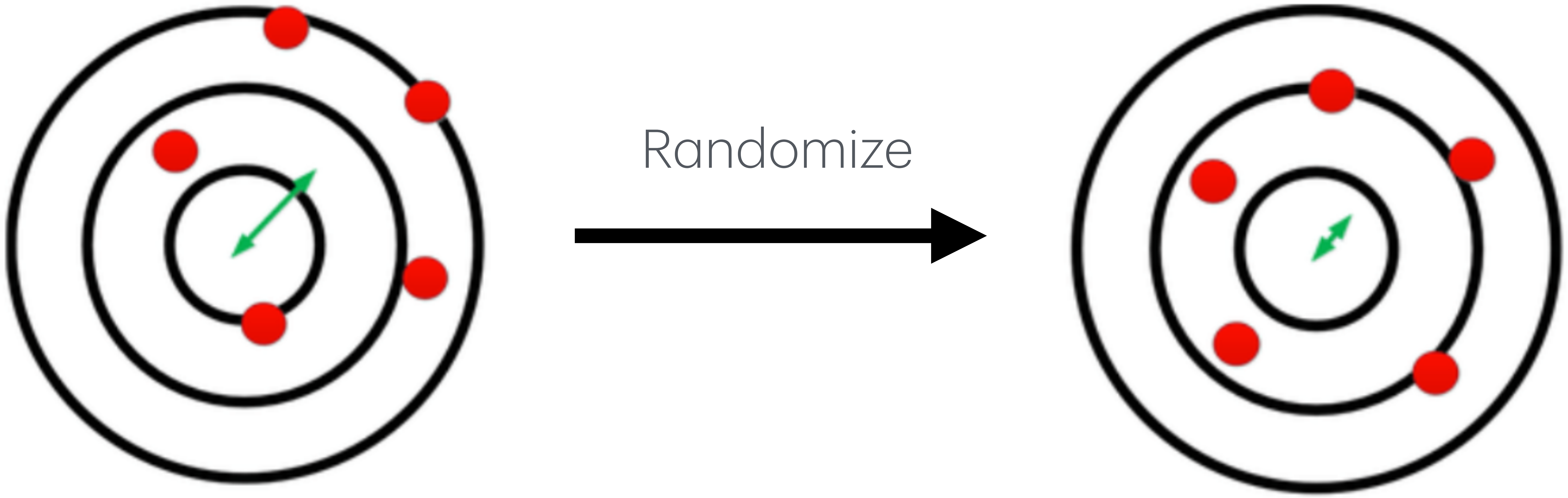
Measurement: Selection bias

- Usually don't run A/B test on all users, or all transactions, etc. [Risky]
- Select subset to run on
- Selection bias:
  - Pop: 40% EU, 60% US
  - Subset: 10% EU, 90% US
- Fix: Sample uniformly randomly from full population

Measurement bias

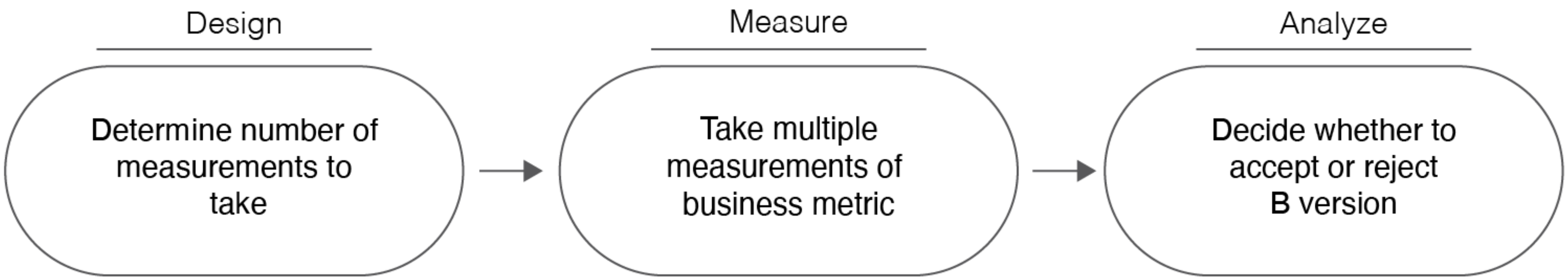
One measurement
(one experiment)
is one dart

- Randomization decreases bias



Randomization removes biases
that you don't even
know are there.

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

- Tension:
  - Goal: Keep the cost low ==> N small
  - Goal: Keep the uncertainty low ==> N large
- Resolution:
  - "Find the smallest N s.t. uncertainty is low enough."

Design

- A/B testing approach to uncertainty
  - Limit *false positives*
  - Limit *false negatives*
- False Positive (FP)
  - Measure BM(B) > BM(A), wrong

FP: Makes system worse

- False Negative (FN)
  - Measure BM(B) ≤ BM(A), wrong

FN: Misses opportunity

False Positives & Negatives

- Running ad system w/XGBoost model (Version A)
- Version B: Transformer model
- False Positive
  - A/B test says, “Switch to B”.
  - In long run, B gets fewer clicks than A.
- False Negative
  - A/B test say, “Stick w/A”
  - B *would have* gotten more clicks than A

But you’ll never know

Design

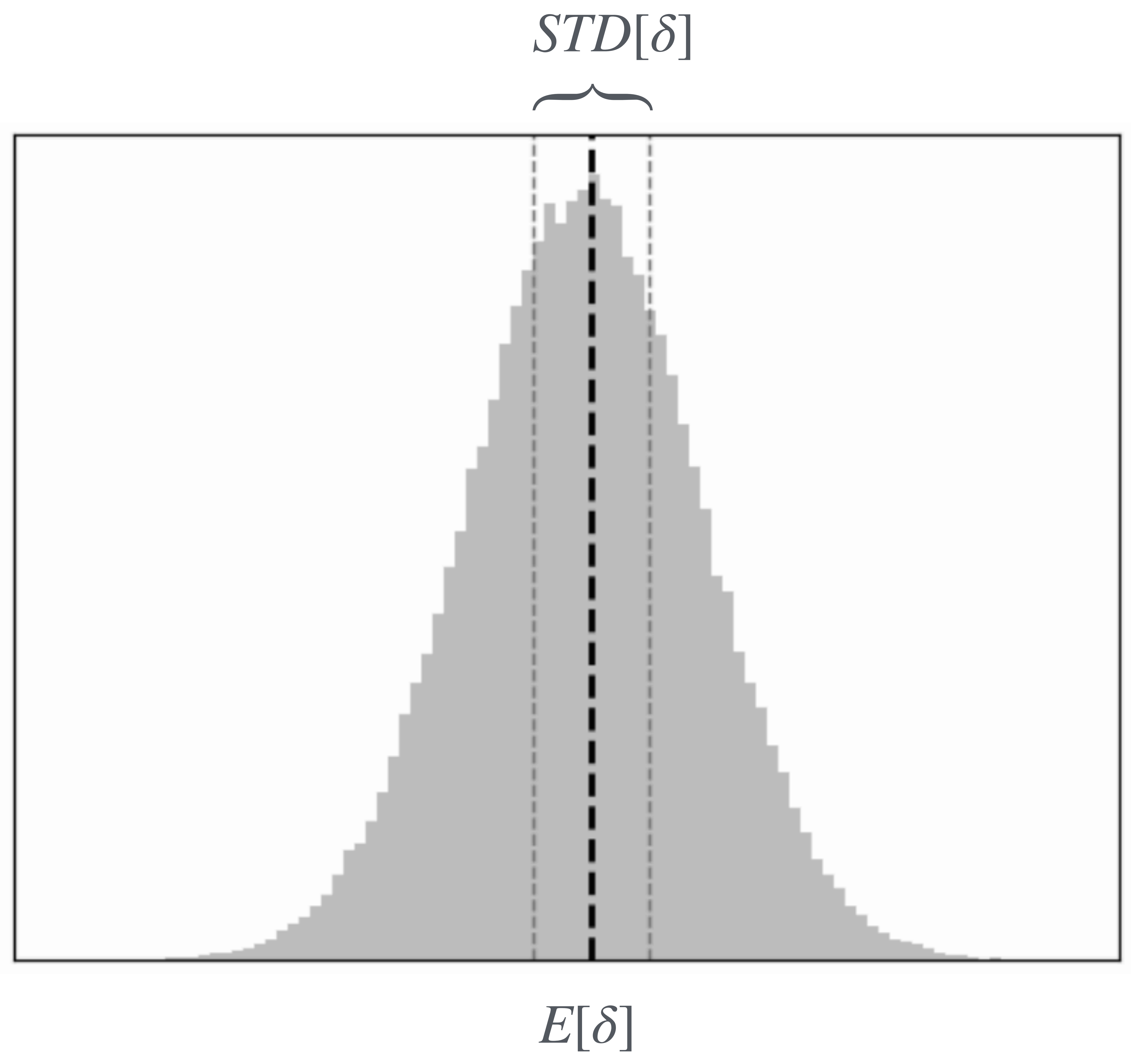
- Conventional limits:
  - $P\{FP\} < 0.05$
  - $P\{FN\} < 0.20$
- Notation:
  - $\alpha = 0.05 \qquad \beta = 0.20$

Design

- “Begin with the end in mind”
- Consider analysis
  - Accept: Switch from A to B
  - Reject: Stick with A, discard B
- Decision criteria?

Design

- CLT: $\delta \sim \mathcal{N}(E[\delta], STD[\delta])$
- Experiment is one draw



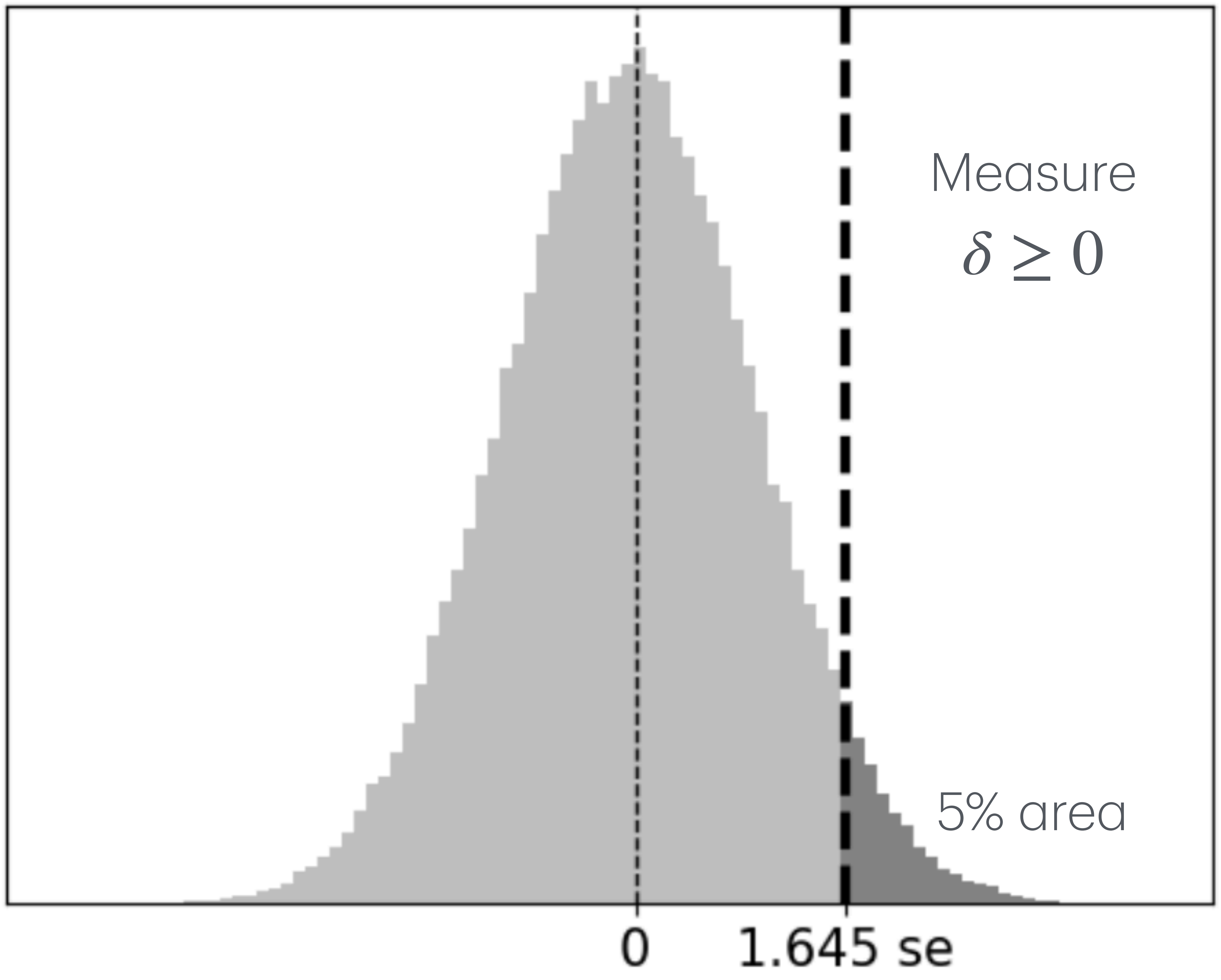
Analysis: False Positive

- Accept b/c $\delta > 0$
  - Wrong, ex $E[\delta] = 0$
- **Criterion 1**:

$$\delta > 1.645se$$

- Then

$$P\{FP\} < 0.05$$

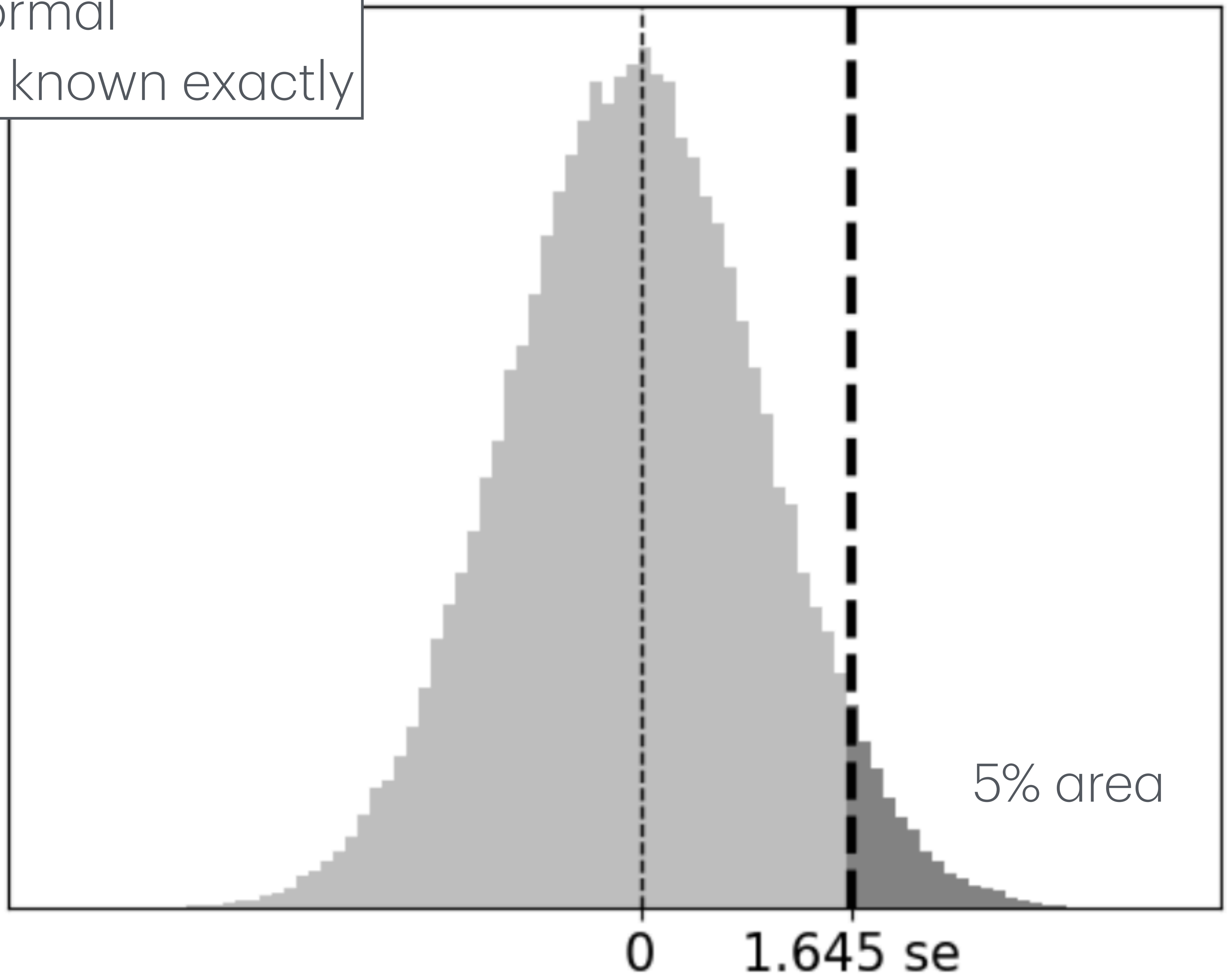


Statistical Significance

- $\delta \sim \mathcal{N}(E[\delta], STD[\delta])$
- Hypothesize: $E[\delta] = 0$
- se estimates $STD[\delta]$
- 5% of probability:

$$\delta > 1.645se$$

Normal
Probabilities known exactly



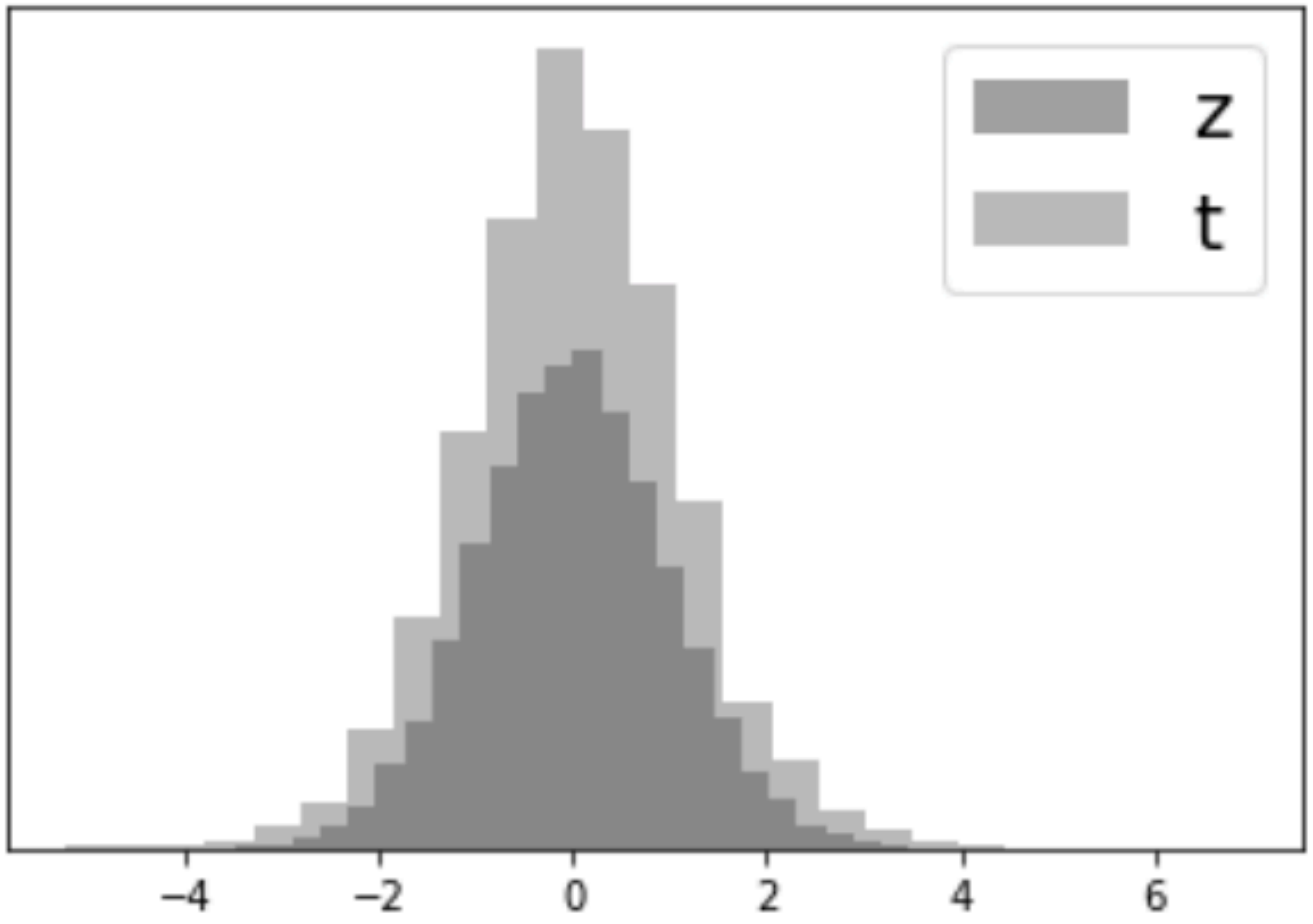
Aside: t statistic

Student's t

- t statistic: $t = \dfrac{\delta}{se}$
- The uncertainty in se makes t follow t distributior
  - Fatter tails than normal distribution
- Criterion 1: $\delta > 1.645se$

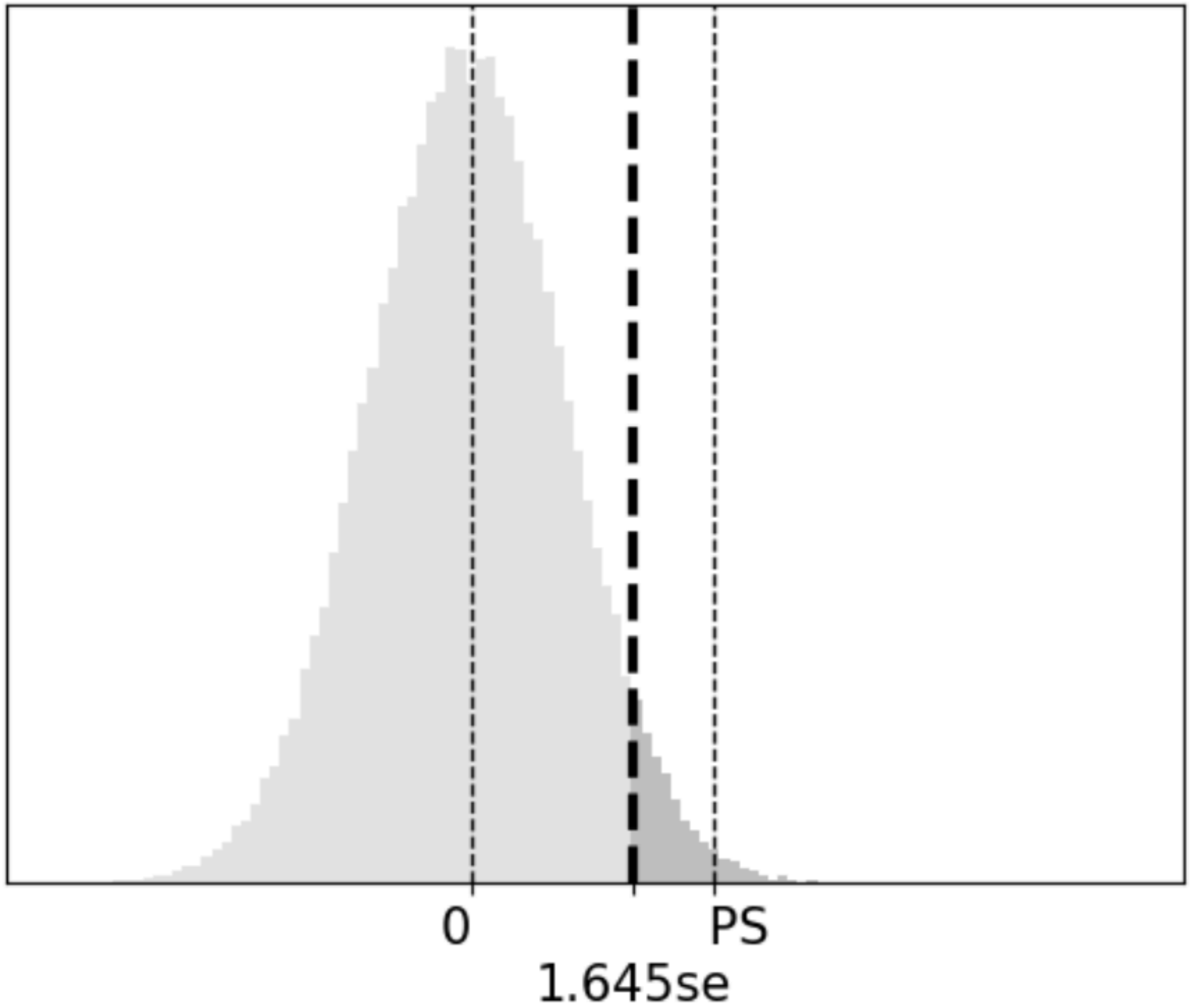
$$t > 1.645$$

- is called "t test"



Analysis: False Negative

- Measure $\delta < 1.645se$
- Decision: REJECT
- Really, though:
  - $E[\delta] \geq PS$



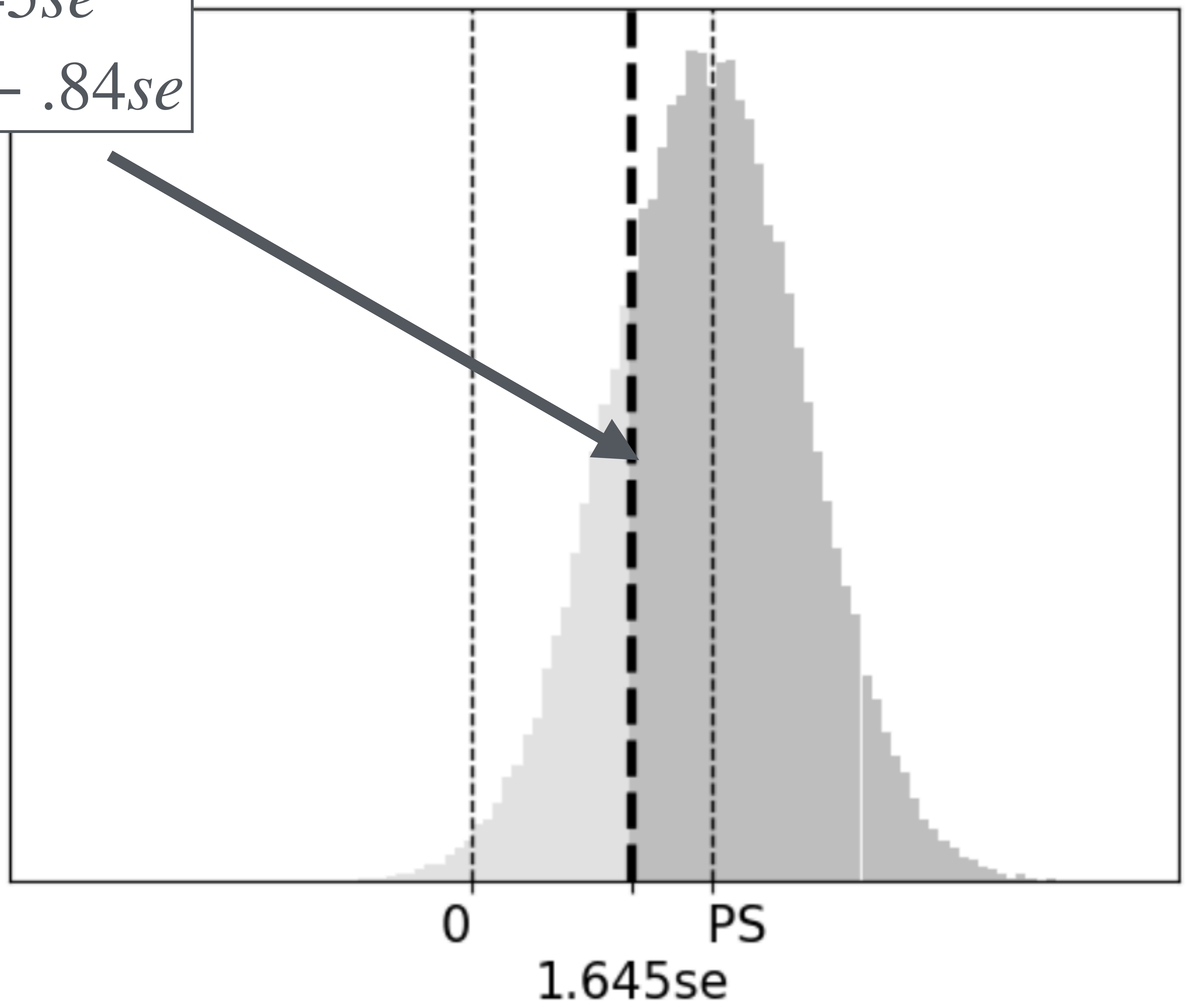
Analysis: False Negative

- $E[\delta] \geq PS$
- Simpler: $E[\delta] = PS$
- Limit to $P\{FN\} < 0.20$

$$PS - .84se \geq 1.645se$$

$$PS \geq 2.5se$$

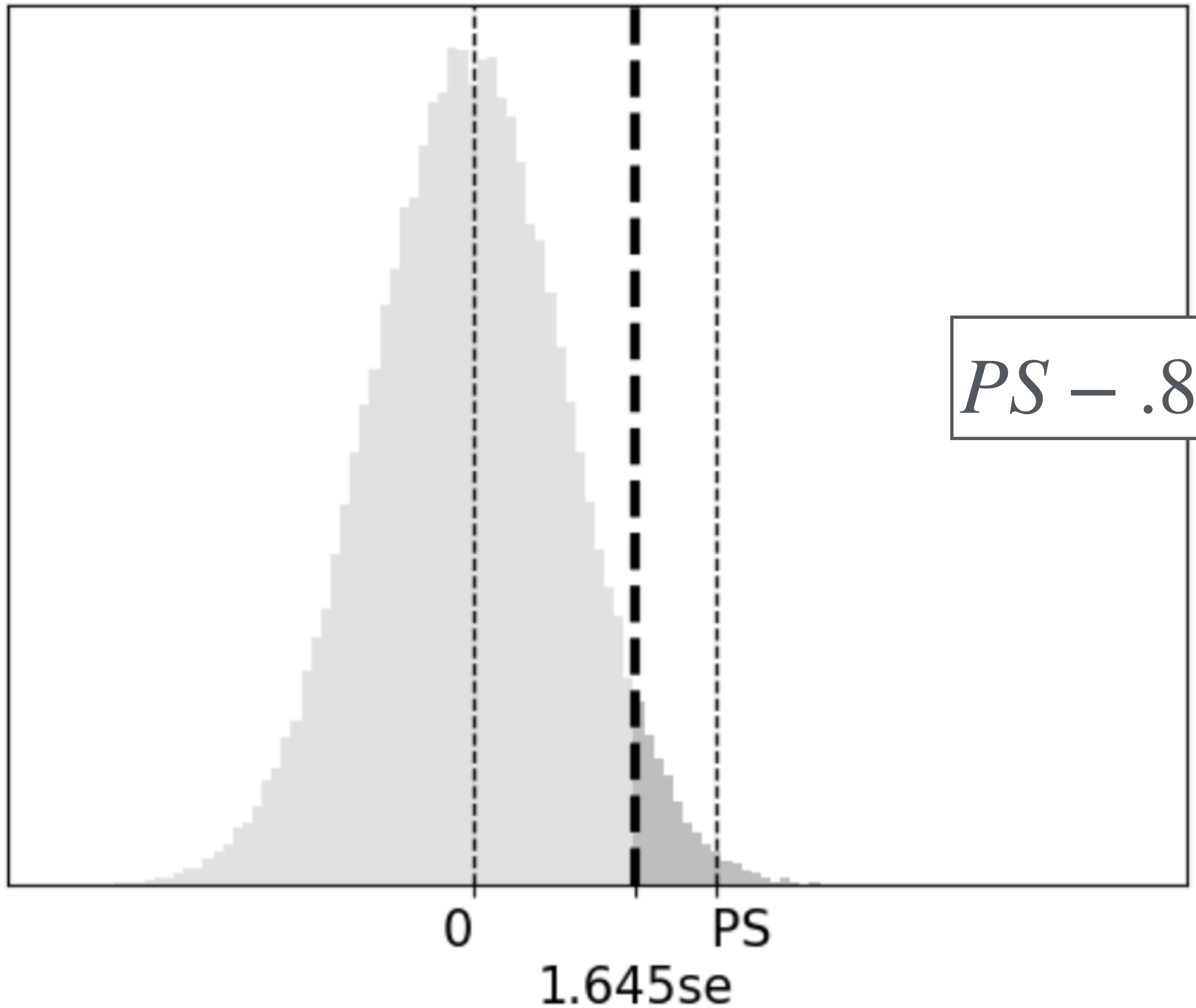
$1.645se$
$= PS - .84se$



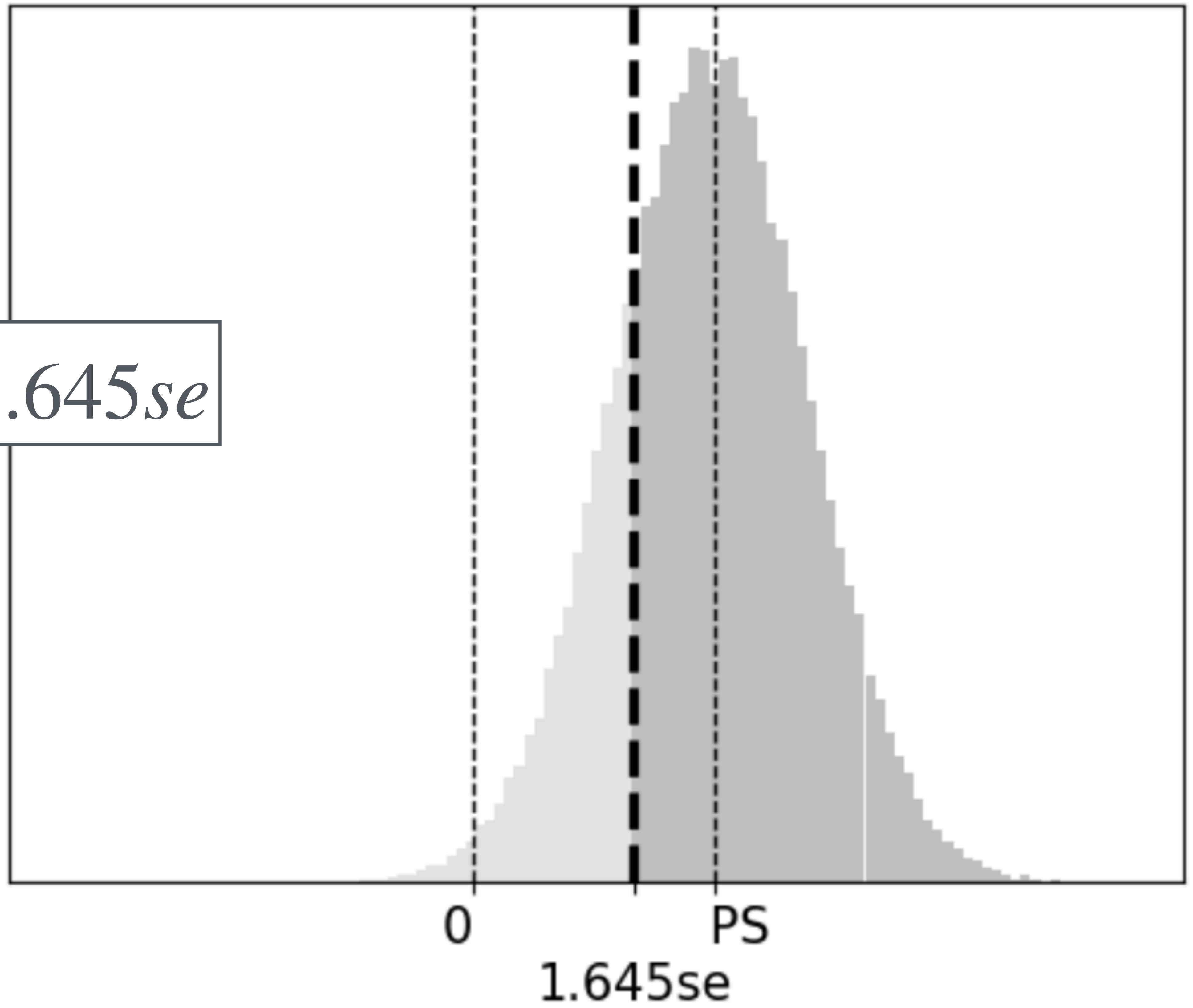
Design

$$P\{FP\} < 0.05$$
$$\delta > 1.645se$$

$$P\{FN\} < 0.20$$
$$\delta > PS - 0.84se$$



$$\boxed{PS - .84se \geq 1.645se}$$



Design

- Design criteria driven by acceptance criteria
- $PS - 2.5se \geq 0$
- $se \leq \frac{PS}{2.5}$
- Recall $se = \sigma_\delta / \sqrt{N}$

$$\sigma_\delta = \sqrt{\sigma_A^2 + \sigma_B^2}$$

$$\boxed{N \geq \left(\frac{2.5\sigma_\delta}{PS}\right)^2}$$

Design

- Don't know σ_δ before measurement
- Estimate:
  1. Existing data for A, $\sigma_B \approx \sigma_A$
  2. Pilot study: measure σ_A, σ_B
- Call it $\hat{\sigma}_\delta$

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

Design

- Decide PS
  - Business knowledge
  - Team/manager/customer/market determines goal
  - Somewhat subjective

Design

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%

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

Analysis: Practical Significance

- Practical Significance (PS): The smallest improvement you'd care about
- **Criterion 2**:

$$\delta > PS$$

Analysis: Decision Criteria

- Accept if

  Criterion 1: $\delta > 1.645se$

  Criterion 2: $\delta > PS$

- Reject otherwise
- Ex, reject if $\delta < 1.645se$

Design Summary

- Goal: Limits, $P\{FP\} < 0.05$ & $P\{FN\} < 0.20$
- Decide PS, Estimate σ_δ
- Find N, number of observations in measurement

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

Measurement Summary

- Replicate: N observations, prescribed by design
- Randomize: Each observation; 50% to A, 50% to B

Analysis Summary

- Calculate

$$\delta = \mu_B - \mu_A \quad \text{—-} \quad se = \sigma_\delta / \sqrt{N} \quad \text{—-} \quad \sigma_\delta = \sqrt{\sigma_A^2 + \sigma_B^2}$$

- Accept B if:

$$\delta > 1.645 se \quad \& \quad \delta > PS$$

- else reject B. Stick with A.