Experimental optimization

Lecture 3: A/B testing II: Design

David Sweet

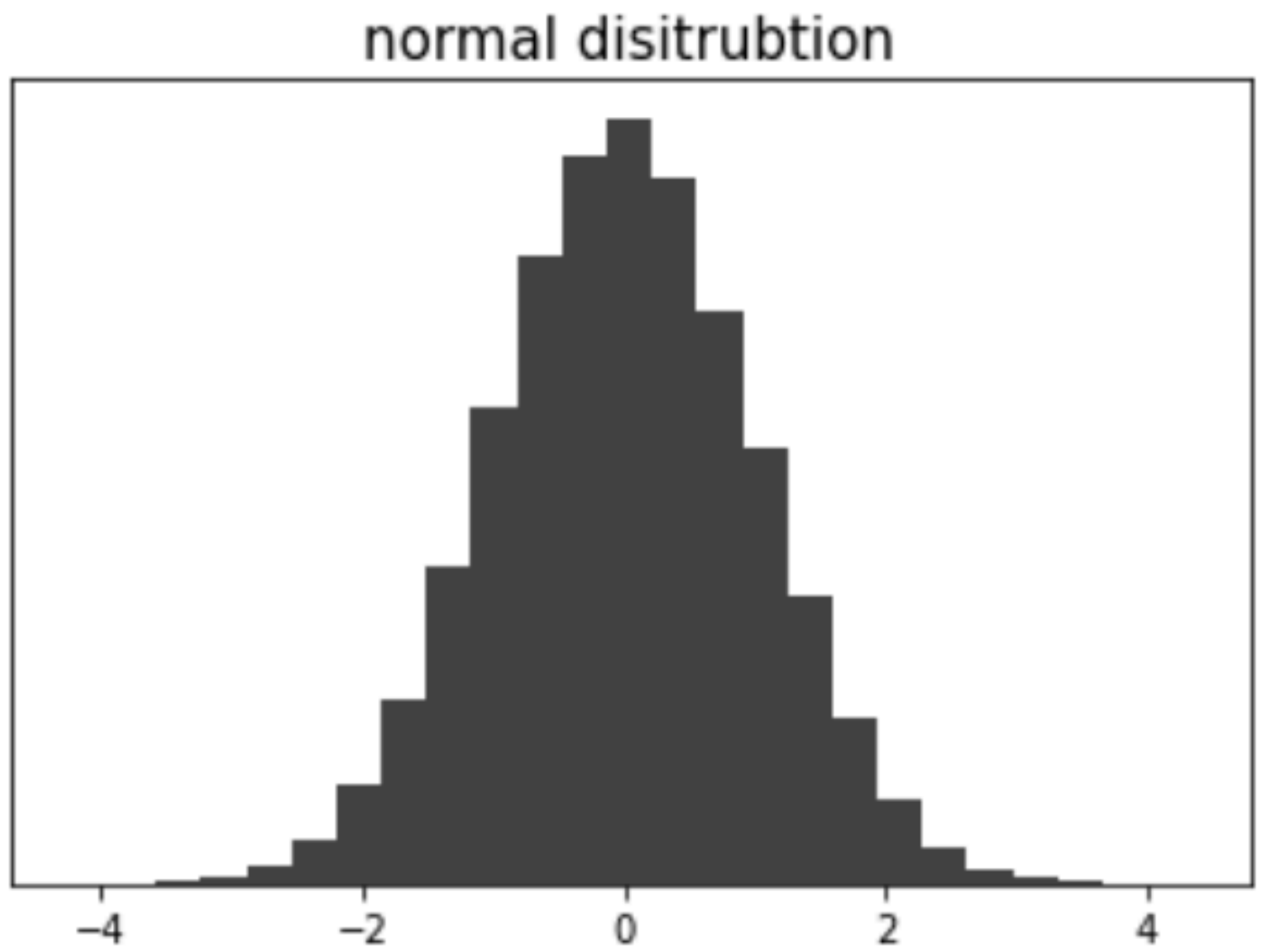
Mid-term project

Run an A/B test

- Three independent simulators, representing three engineered systems
- Simulators will return measurements via http
- Simulators will be slow
- You will run A/B tests on the simulators and
  - Write up your design, measurements, and analysis
  - Give a 5 minute presentation on the results during lecture #6

Review

Central limit theorem



- Given N samples $x_i \sim X$ of any distribution,
- with sample mean $\mu = \Sigma_i^N x_i / N$, as $N \to \infty$

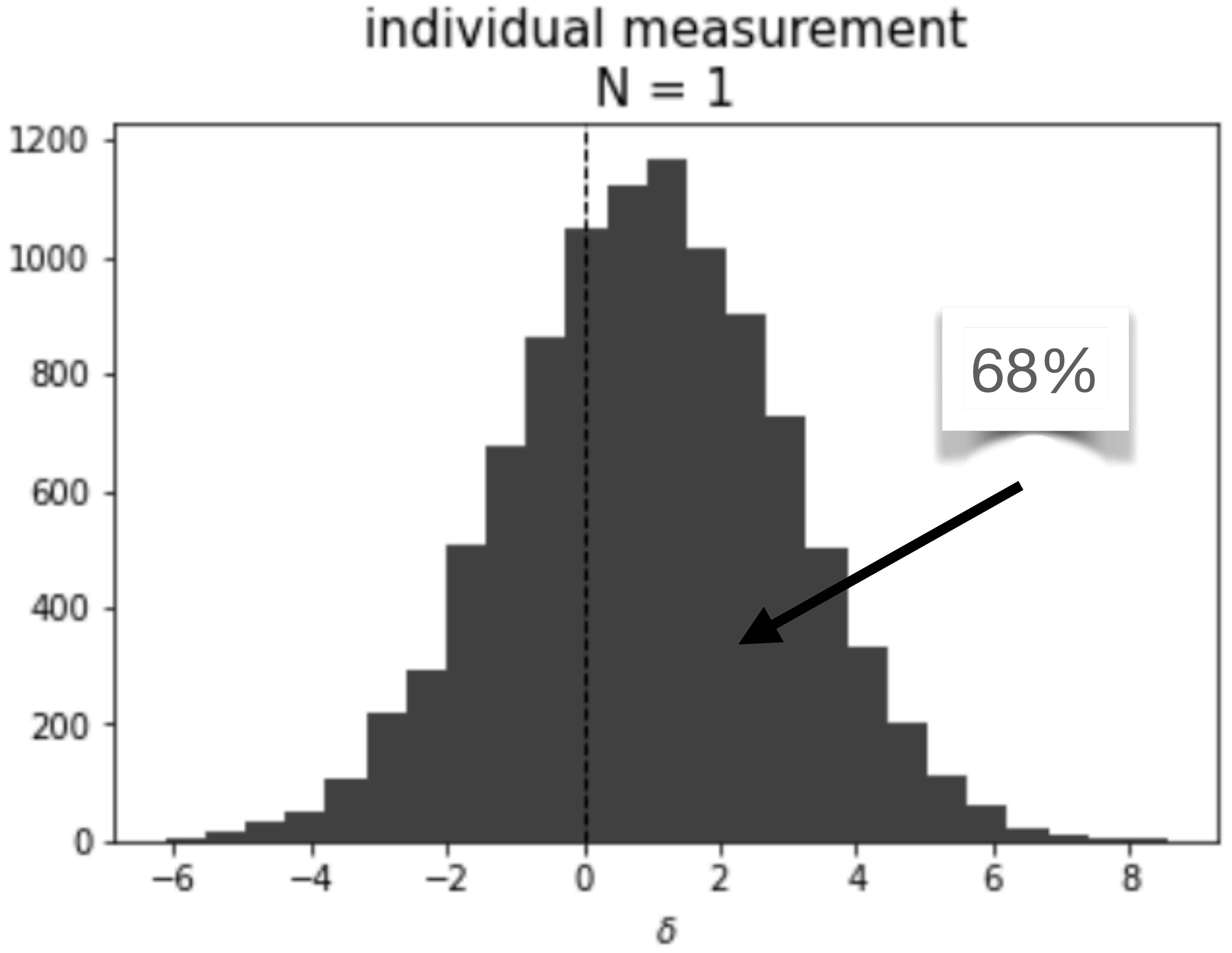
$$\mu \sim \mathcal{N}(\bar{x}, \sigma^2)$$

- IOW: Aggregate measurements are normally distributed
- …even if individual measurements are not
- (…when we have a lot of individual measurements)

A/B test

Is B better than A?

- Goal: correctly choose the better of versions A & B
- Define $\mu = \mu_B - \mu_A$
  (μ_A is agg. meas. of BM(A), resp. μ_B)
- Restate goal: Is $\mu > 0$?
- About 68% of ind. meas. have $\mu > 0$
- $P\{\text{wrong}\} = 1 - .68 = .32$

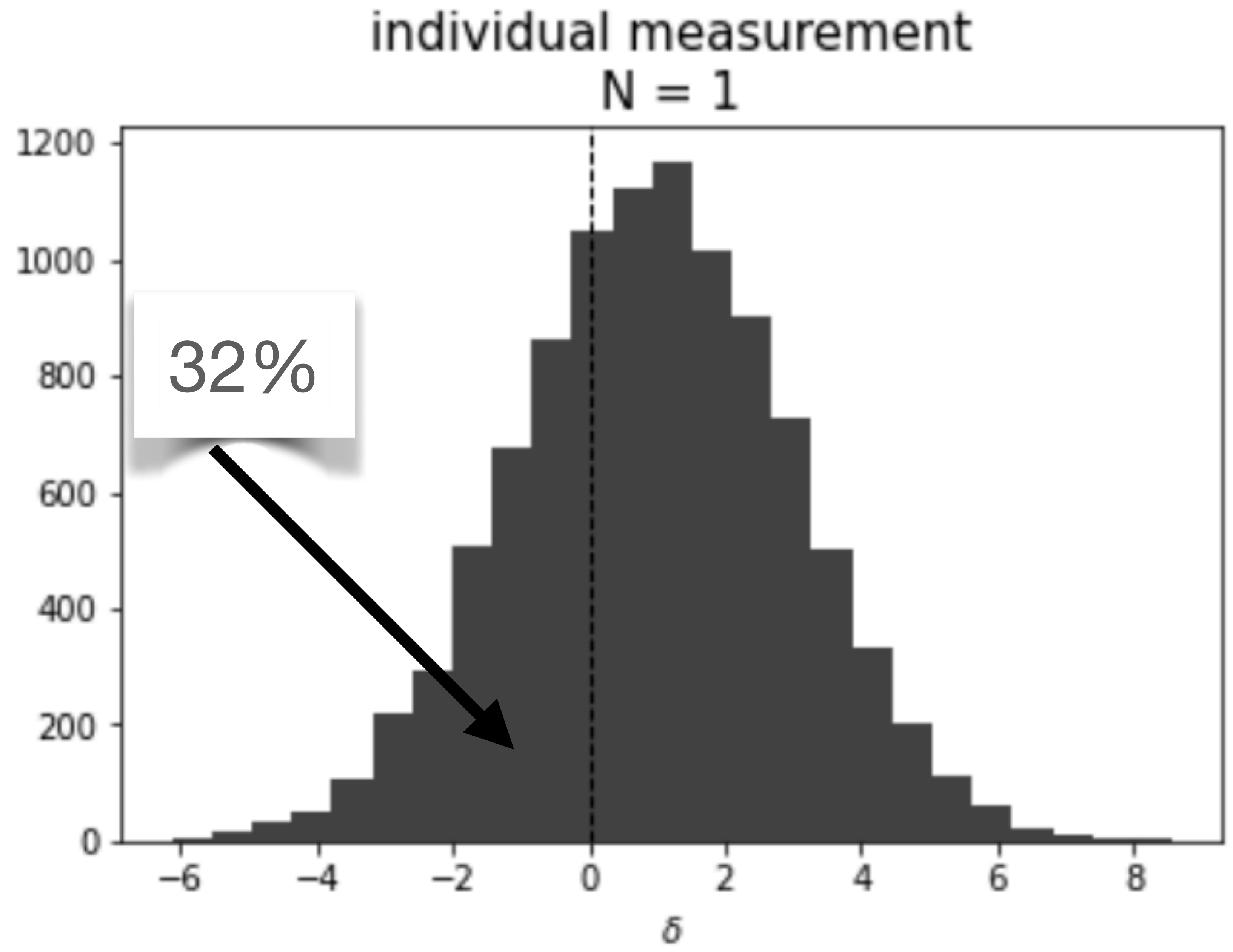


A/B test design

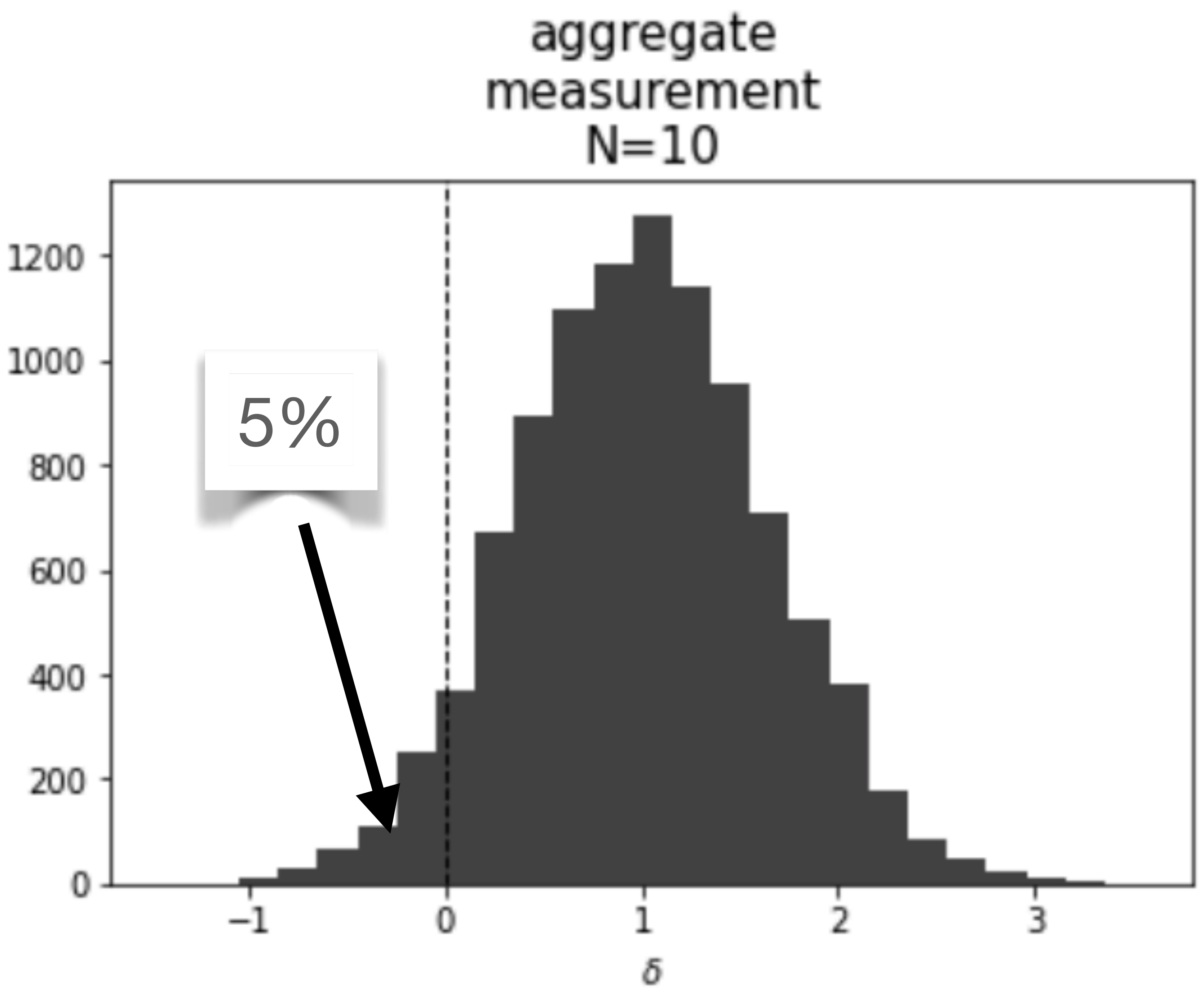
Probably not wrong

- Larger N ==> lower SE of agg. meas. ==> lower $P\{\text{wrong}\}$

$P\{\text{wrong}\} = .32$



$P\{\text{wrong}\} = .05$



A/B test design

Minimize N

- Larger N ==> higher experimentation costs, too
- A/B test design:

$$\textbf{Pick the smallest } N \textbf{ s.t. } P\{\textbf{wrong}\} < .05$$

- How? “Begin with the end in mind”

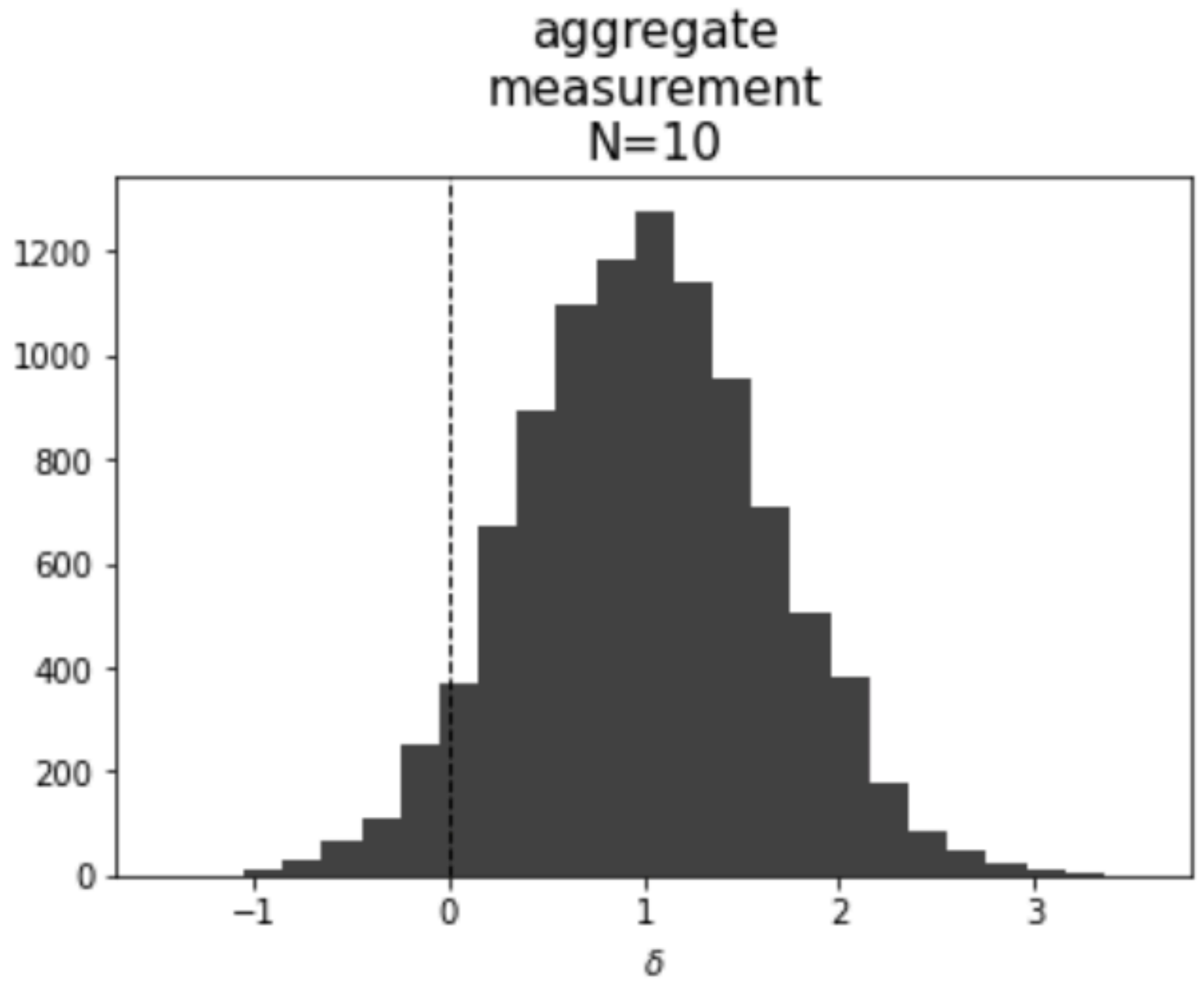
Analysis

Imagine measurement is done

- At end of A/B test you have measurements $x_{A,i}$ and $x_{B,i}$, from which you calculate
    - $\mu = \Sigma_i x_{B,i}/N - \Sigma_i x_{A,i}/N$

Aggregate measurement

- Ask: Is $\mu > 0$? (Is B better than A?)   <== easy
- But μ is one sample from a distribution like ==> so you could be wrong



Aside: notation

Notation for this lecture

- $x_{A,i}, x_{B,i}$ - individual measurements
- $\mu_A = \Sigma_i x_{A,i}/N$ - aggregate measurement of BM(A)
- $\mu_B = \Sigma_i x_{B,i}/N$ - aggregate measurement of BM(B)
- μ - aggregate measurement [of the difference BM(B)-BM(A)]
- $\bar{\mu}$ - expectation of aggregate measurement

Aside: notation

Notation for this lecture

- σ_A - standard deviation of $x_{A,i}$
- σ_B - standard deviation of $x_{B,i}$
- $\hat{SE} = \sigma/\sqrt{N}$ = standard error of μ, where $\sigma^2 = \sigma_A^2 + \sigma_B^2$

Analysis
False positive

- False positive (FP):
  If B is worse than A, i.e., expectation of BM(B) < BM(A), i.e. $\bar{\mu} < 0$, and
  you measure $\mu > 0$, i.e., you make the wrong decision

- rename: $P\{\text{wrong}\} = P\{FP\}$

Analysis

False positive

- Calculate: $\hat{SE} = \dfrac{\sigma}{\sqrt{N}}$
- Define $z = \dfrac{\mu - \bar{\mu}}{\hat{SE}}$ and hypothesize that $\bar{\mu} = 0$, i.e., BM(B)==BM(A)
  - then: $z = \dfrac{\mu}{\hat{SE}}$

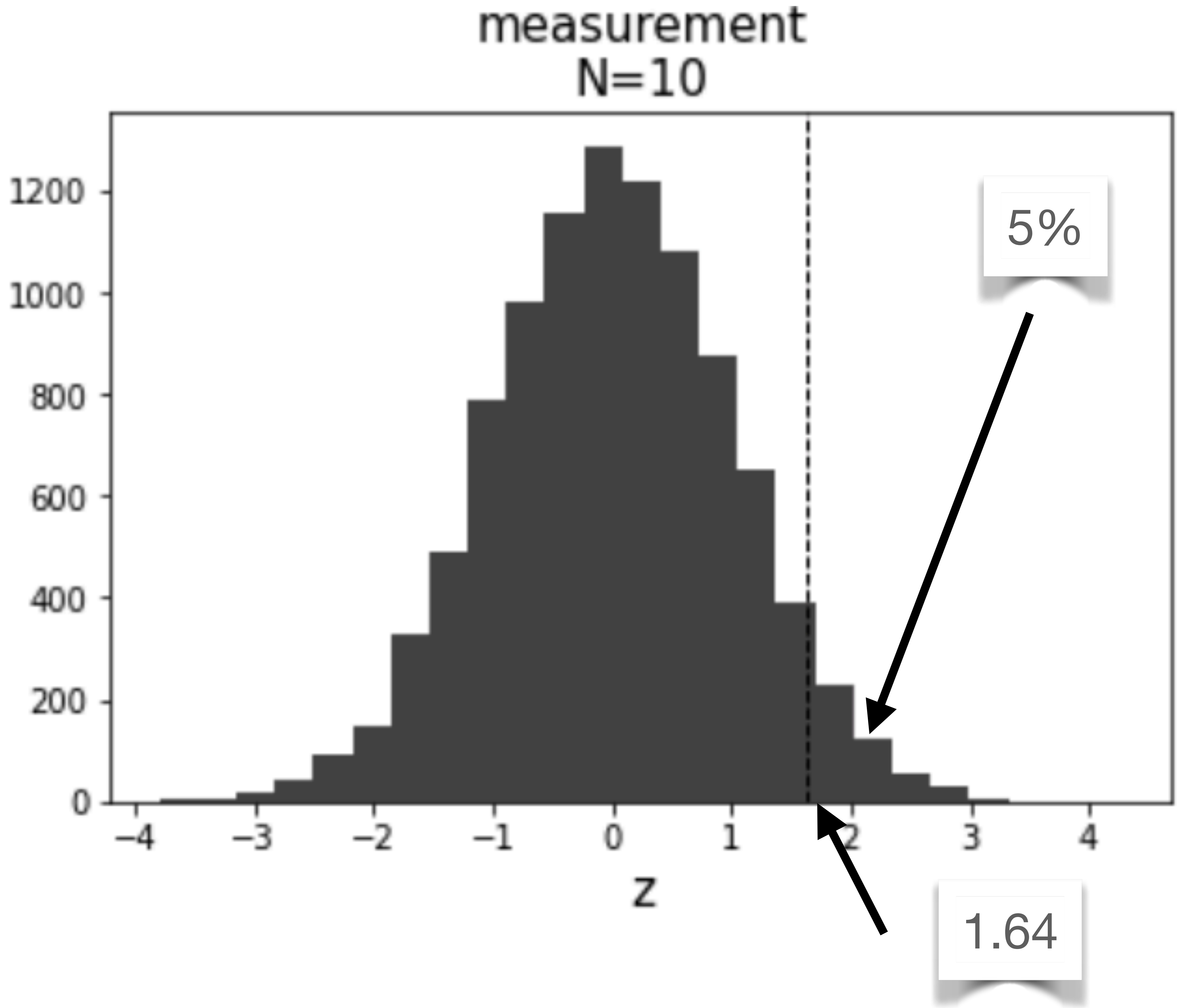
z-score of the aggregate measurement

- Then, by CLT, $z \sim \mathcal{N}(0,1)$

Analysis

Limit P{FP} to 5%

- Hypothesize $\bar{\mu} = 0$, then $\bar{z} = 0$
    - i.e., BM(B)=BM(A)
    - called *null hypothesis*
- $P\{\text{FP}\}$ = prob. of z falling to right of vertical line
- If $z > 1.64$, probably $\bar{z} \neq 0$ (5%)



Why 1.64?

Shape of normal distribution

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
										359
										753
										141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7854
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8804	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767

Why 1.64?

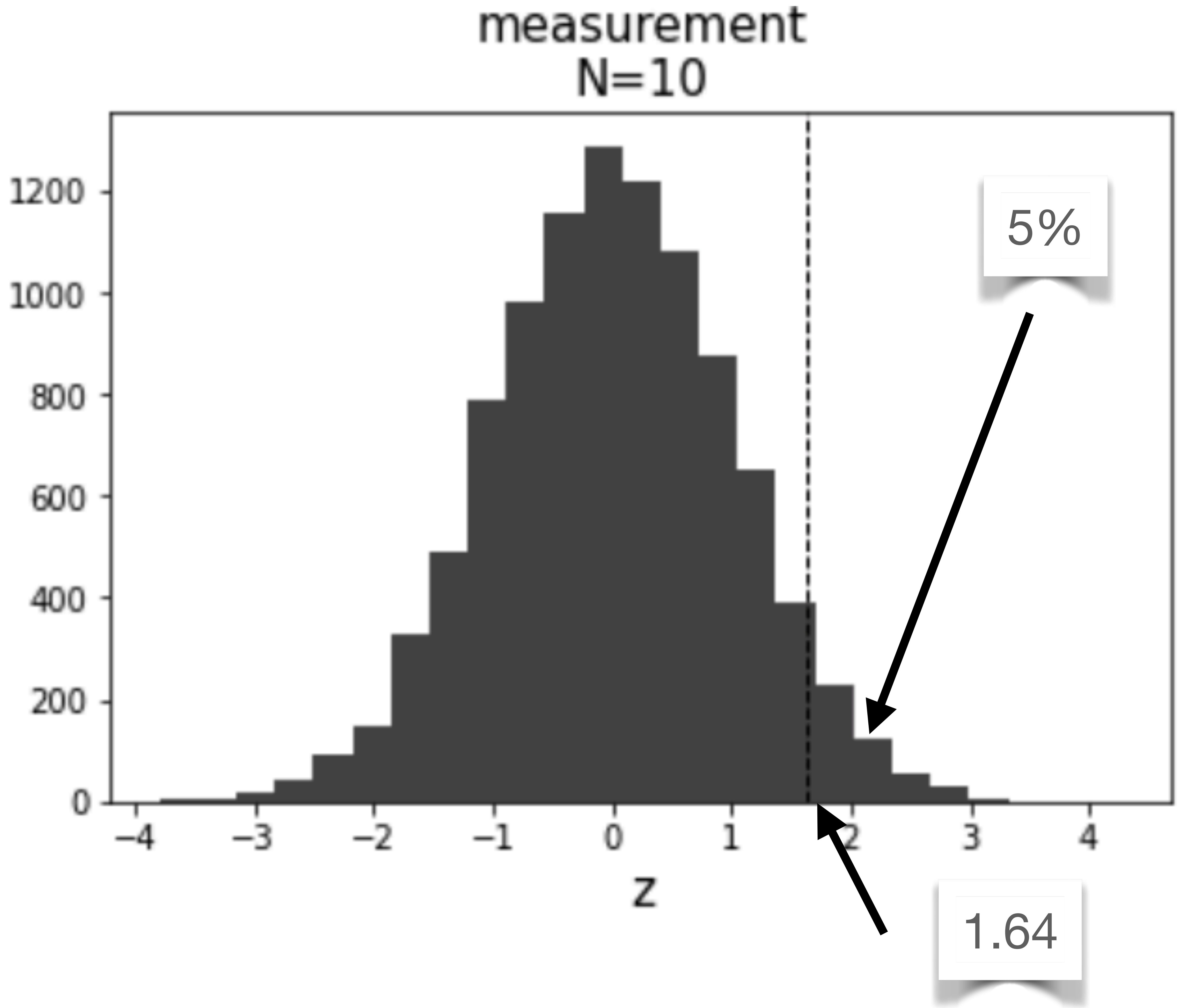
Shape of normal distribution

```
from scipy.stats import norm
norm.cdf(1.64)
```

```
0.9494974165258963
```

Analysis
Statistical significance

- Decision rule:
  - If $z > 1.64$, *act as if* $\bar{z} \neq 0$
  - If $z > 1.64$, *act as if* B beats A
- Agg. measurement is *statistically significant* when $z > 1.64$



Analysis

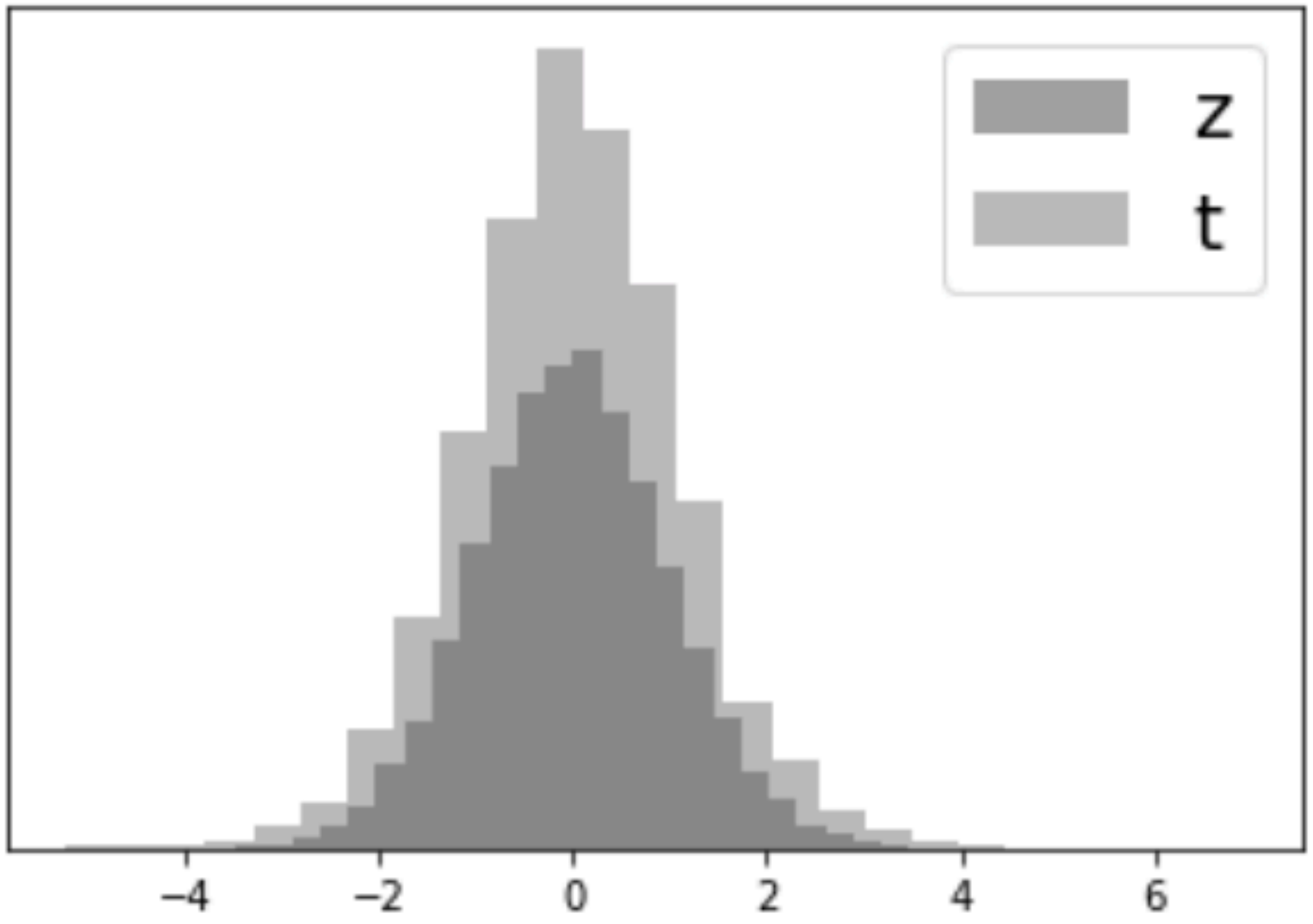
Practical significance

- Say $z > 1.64$, should you switch to version B?
- What if μ = $.01/day?
- Takes effort and risk to switch from A to B.
- Exogenous business decision (depends on you, team, firm, industry)
- Only switch if $z > 1.64$ **and** $\mu > PS$
- PS is *practical significance* level

Aside: t statistic

Student's t

- $z = \frac{\mu}{\hat{SE}}$, N large
- If N not large, write $t = \frac{\mu}{\hat{SE}}$
- The variation in $\hat{SE}$ makes t non-normal, "t distribution"



Design

Prepare for analysis stage

- Make N just large enough s.t. $z > 1.64$
  - — but no larger (to limit experimentation costs)
- Unpack:

$$z = \frac{\mu}{\hat{SE}} = \frac{\sqrt{N}\mu}{\sigma} > 1.64$$

$$\hat{SE} = \sigma/\sqrt{N}$$

- and solve for N:

$$N > (\frac{1.64\sigma}{\mu})^2$$

Design

Estimate inputs: σ

- Don't know σ at design time, so estimate
- Either
    - Note: $\mu = \mu_B - \mu_A, \quad \sigma^2 = \sigma_B^2 + \sigma_A^2$
    - Estimate $\hat{\sigma}_A$ by stddev of logged data from version A
- Assume $\hat{\sigma}_B = \hat{\sigma}_A$, then $\hat{\sigma}^2 = 2\hat{\sigma}_A^2$, or
    - Run a *pilot study* — go measure σ_B directly

Design

Estimate inputs: μ

- $N > (\frac{1.64\hat{\sigma}}{\mu})^2$
- Smaller δ ==> larger N
- What’s the smallest μ we’d actually care to measure?
    - $\mu = PS$
- Finally:

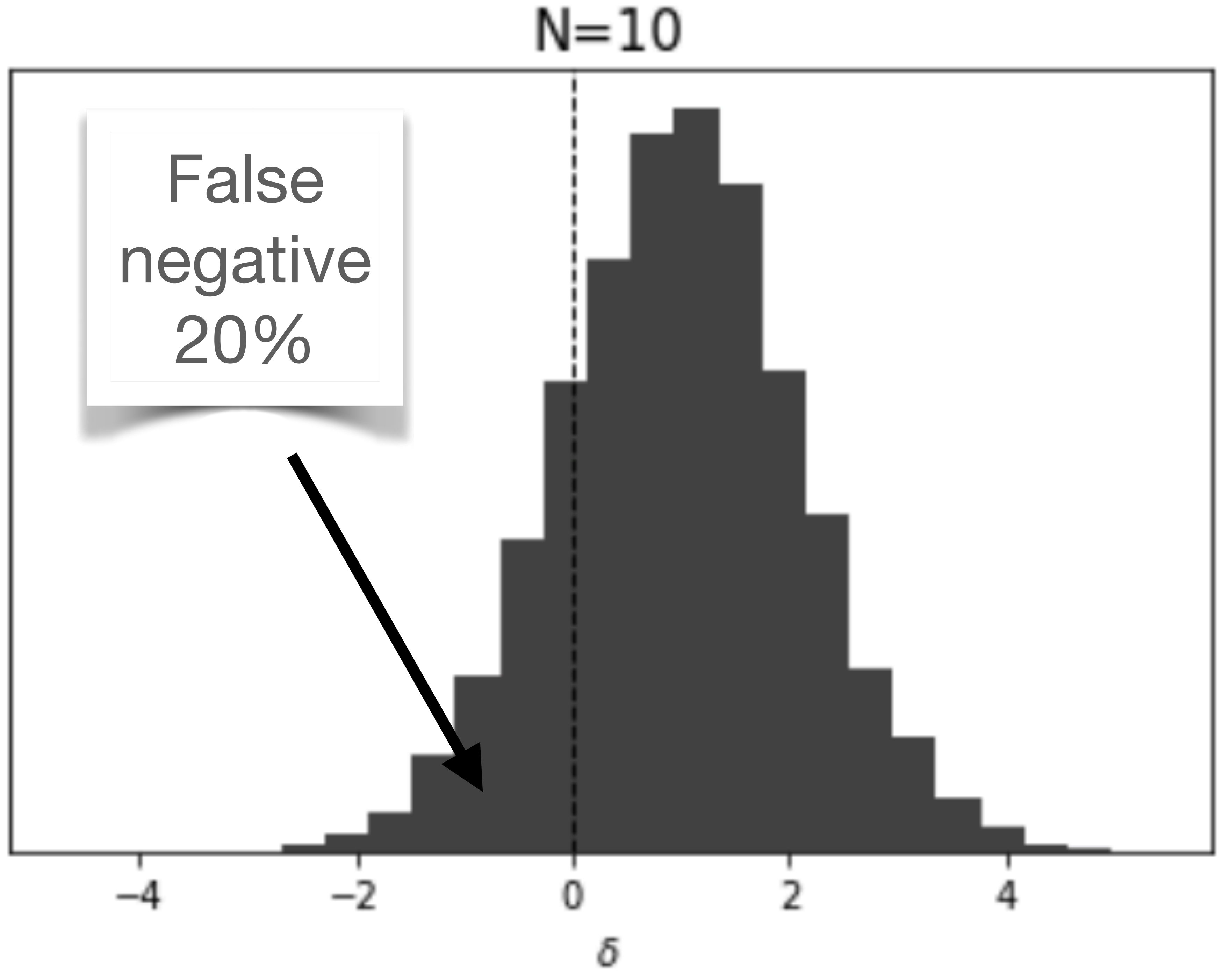
$$N > (\frac{1.64\hat{\sigma}}{PS})^2$$

The A/B test design

False negatives

The other way to be wrong

- False positive: You measure “B better than A”, but it isn’t
- *False negative*: You measure “B **not** better than A”, but it **is**
- Imagine that really $\bar{\mu} > 0$
- If you measure $\mu < 0$, that’s a false negative
- Limit: $P\{FN\} < .20$



Design

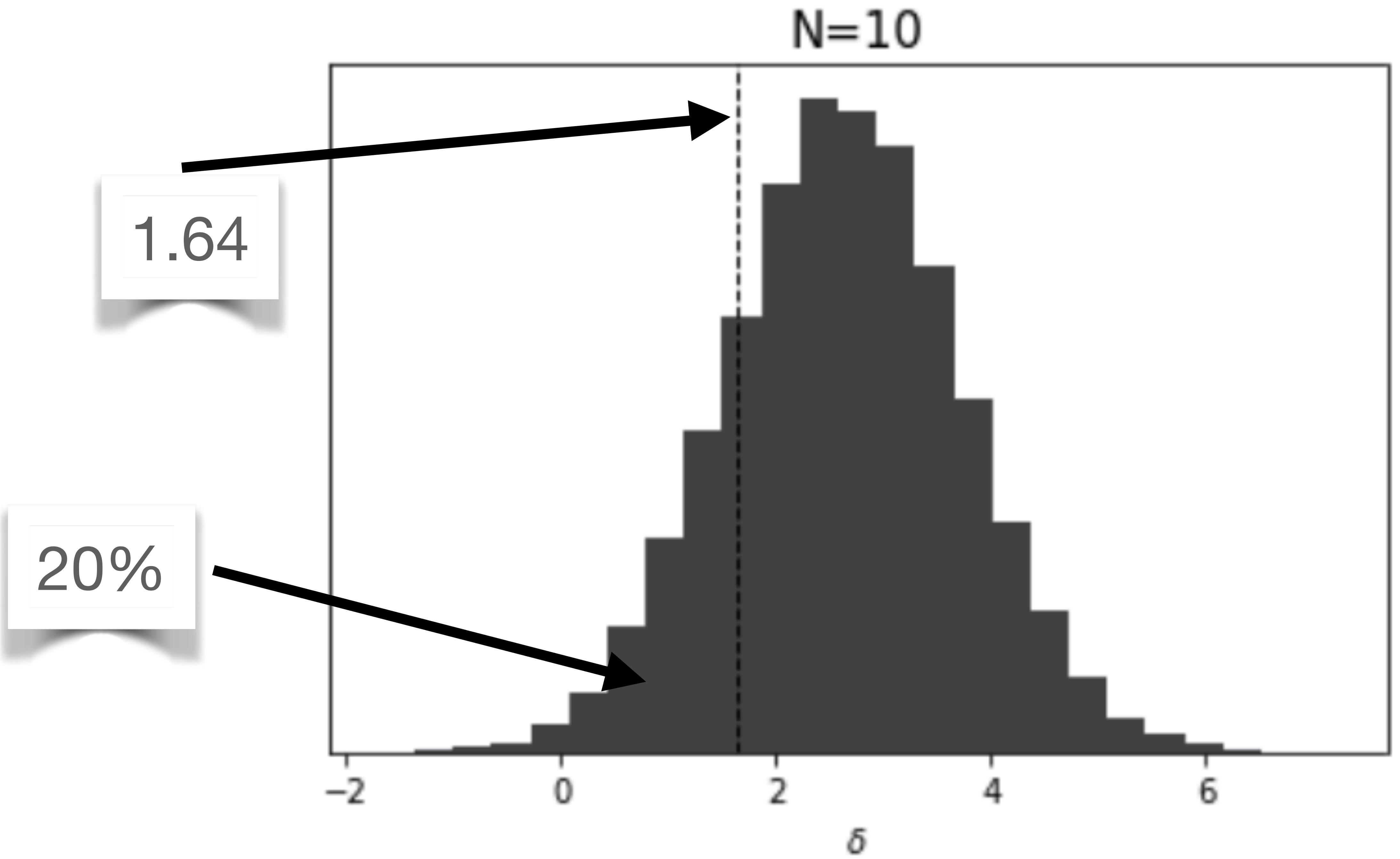
Power analysis

- Knowing that, during analysis, you'll accept a measurement that is
    - practically significant, — this is fixed
    - statistically significant, and — N controls this
- The worst-case false negative rate will be when:
    - true $\bar{\mu} = PS$ <== smallest it could be
- FP when you measure $z < 1.64$ (reject B incorrectly)

Design
Power analysis

- Solution: Keep threshold (z=1.64) far from $z = PS/\hat{SE}$
- Far enough so that you measure $z < 1.64$ less that 20% of the time
- Limits $P\{FN\} < .20$
- $z - .84 > 1.64$, or

$$N > (\frac{2.48\hat{\sigma}}{PS})^2$$



Design

Design summary

- "This A/B test measures a difference of precision PS with *power* of 80% at a *statistical significance* level of 5%"

$$N > \left(\frac{2.48\hat{\sigma}}{PS}\right)^2$$

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/B test design

Summary

- Limit false positives to 5%
- Limit false negatives to 20%
- Estimate σ with logs & $\sigma_B = \sigma_A$ OR run a pilot study

Design

- Switch to B if
  - statistical significance, $z > 1.64$, and
  - practical significance, $\delta > PS$

Analysis