CS 260: Foundations of Data Science

Prof. Sara Mathieson

Spring 2025

HAVERFORD
COLLEGE

Admin

- **Lab 5 grades** on Moodle
- **Lab 7 / Project proposal** due last night
- **Lab 8 posted**, due ~~Wednesday~~ Thursday
- **Final project** instructions/rubric posted
- Next week:
  - Finish statistics, then midterm review
  - **Midterm in-class Thursday April 17**

Outline for today

- Randomized trials for the null distribution
- Are the means of two samples different?
  - t-tests
  - Permutation testing
- Bootstrapping

Outline for today

- Randomized trials for the null distribution
- Are the means of two samples different?
  - t-tests
  - Permutation testing
- Bootstrapping

Central Limit Theorem

- Assumptions
  - $X_1, X_2, \ldots X_n$ are iid samples
  - From a population with mean μ
  - Finite variance σ^2

- THEN

$$Z = \lim_{n \to \infty} \sqrt{n} \left(\frac{\bar{X}_n - \mu}{\sigma} \right)$$

is a standard normal distribution (i.e. mean 0 and variance 1)

Central Limit Theorem

- Last time we saw that the central limit theorem could be used to estimate a p-value

$$Z = \lim_{n \to \infty} \sqrt{n} \left(\frac{\bar{X}_n - \mu}{\sigma} \right)$$

- We first obtain a Z-score, then compute the probability of observing a result *as or more* extreme **under the null hypothesis**

Recap Handout 17

+

Continuous ⬄ Discrete features

Handout 17

① $E[X] = \sum_{x} x \cdot p(x) = 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = \boxed{\frac{1}{2} = \mu}$

② $Var(X) = E[(X-\mu)^2]$

$= \sum_{x} (x-\mu)^2 p(x)$

$= \left(0 - \frac{1}{2}\right)^2 \cdot \frac{1}{2} + \left(1 - \frac{1}{2}\right)^2 \cdot \frac{1}{2}$

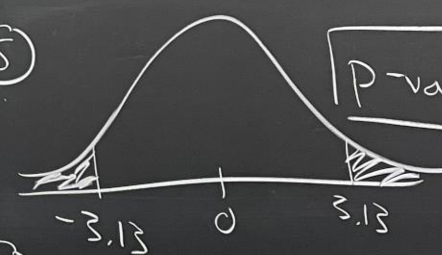
$= \boxed{\frac{1}{4} = \sigma^2}$

③ $\bar{X}_n = \frac{54}{80} = 0.675$

④ $Z = \frac{\bar{X}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}} = \frac{0.675 - 0.5}{\sqrt{\frac{0.25}{80}}}$

$\boxed{Z \approx 3.13}$

⑤



$\boxed{\text{p-value} = 0.001745}$

≤ 0.05

⑥ reject H_0!

continuous → discrete

(Lab 6)

- sort & find label break points
- ex: age ⇒ age ≥ 50
  age < 50

discrete → continuous

- binary → 0 & 1
- ordered → 0, 1, 2 ... V-1
- non-ordered
  ⇒ indicators →

is apple?	is pear?	is orange?
0, 1	0, 1	0, 1

Better way? Randomized trials

- Die example
  - n=10 rolls
  - [4, 2, 3, 1, 3, 1, 3, 3, 3, 1]
  - $\bar{X}_n = 2.4$

- H_0: null hypothesis (fair die)
  - What if we don't know mean & variance of null distribution?

- H_1: is the die weighted toward lower values? (one-sided)

Randomized trials: general idea

1. Run T trials that *mimic* our data under the null hypothesis

   roll a fair die

2. Record relevant information for each trial

   mean of the rolls

3. Count how many times you observe a result *as or more extreme* than your data (N_e)

   any trial with mean less than or equal to 2.4

4. p-value = N_e/T

Randomized trials: general idea

1. Run T trials that *mimic* our data under the null hypothesis

   Right now: each person does 1 trial!

   roll a fair die

2. Record relevant information for each trial

   mean of the rolls

3. Count how many times you observe a result *as or more extreme* than your data (N_e)

   any trial with mean less than or equal to 2.4

4. p-value = N_e/T

4.3	4.1	4.1
3.8	4.1	3.2
4.7	3.3	
3.8	3.1	
3.2	3.9	
3.8	3.1	
4.0	4.0	
3.4	3.4	

$T = 18$

$N_e = 0$

$$\text{p-value} = \frac{N_e}{T} = 0$$

$\Rightarrow$ reject null hypothesis of a fair die

Handout 18

Handout 18

① $T = 20$

② $N_e = 8$

one-sided

13 or fewer

$N_e = 12$

two-sided

13 or fewer

or 17 or more

③ p-value = $\frac{12}{20}$ = 0.6 >> 0.05

fail to reject H_0

④

p-value = $\frac{\overbrace{▨}^{Ne}}{\underbrace{▨ + □}_{T}}$

⑤

~~$\frac{\bar{X} - \mu}{\sqrt{\sigma^2 / n}}$~~ don't know μ & σ^2

Outline for today

- Randomized trials for the null distribution
- Are the means of two samples different?
  - t-tests
  - Permutation testing
- Bootstrapping

Difference in means

(blood pressure)

example:

before drug: [117, 54, 96, 123, 157 ....] n examples

$\bar{X}_n = 112$

after drug: [72, 98, 105, 82 ....] m examples

$\bar{X}_m = 96$

H_0: all #'s are drawn from the same distribution

H_1: after the drug, blood pressure was lowered
(one-sided)

plot

permutation testing

- Simulate null distribution!
- permute the "labels" of the data (i.e. "before" & "after")
- for t in range(T):
  permute labels ← helper

plot: $\bar{X}_m^{(t)} - \bar{X}_n^{(t)}$

difference in means

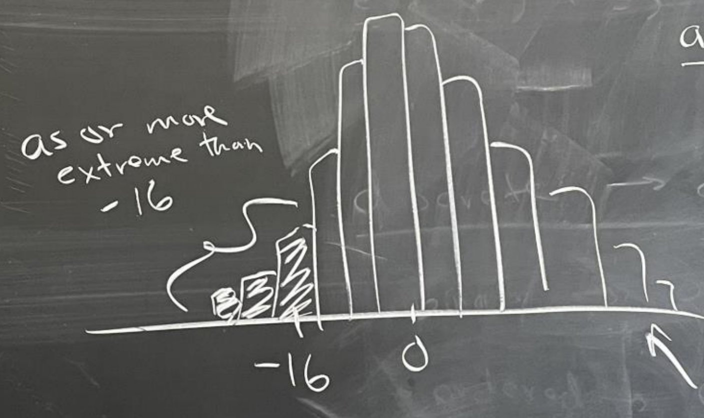
1 trial

"before": [98, 123, 105, 54 ....]

"after": [82, 72, 117, 157, 96 ...]

actual obs

$\bar{X}_m - \bar{X}_n = 96 - \ldots = \boxed{-16}$



Say: $T = 1000$, $N_e = 4 \Rightarrow \boxed{p = 0.004} \ll 0.05$

54 ....] still (n) $\bar{X}_n^{(1)} = 101$

[57, 96 ...] still (m) $\bar{X}_m^{(1)} = 105$

actual obs

$\bar{X}_m - \bar{X}_n = 96 - 112$

$= \boxed{-16}$

$\bar{X}_m - \bar{X}_n$

$\boxed{p = 0.004} << 0.05$

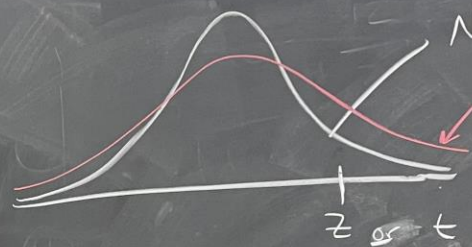
t-tests : CLT-inspired test

drawn from

$$\text{CLT: } Z = \frac{\bar{X}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$$

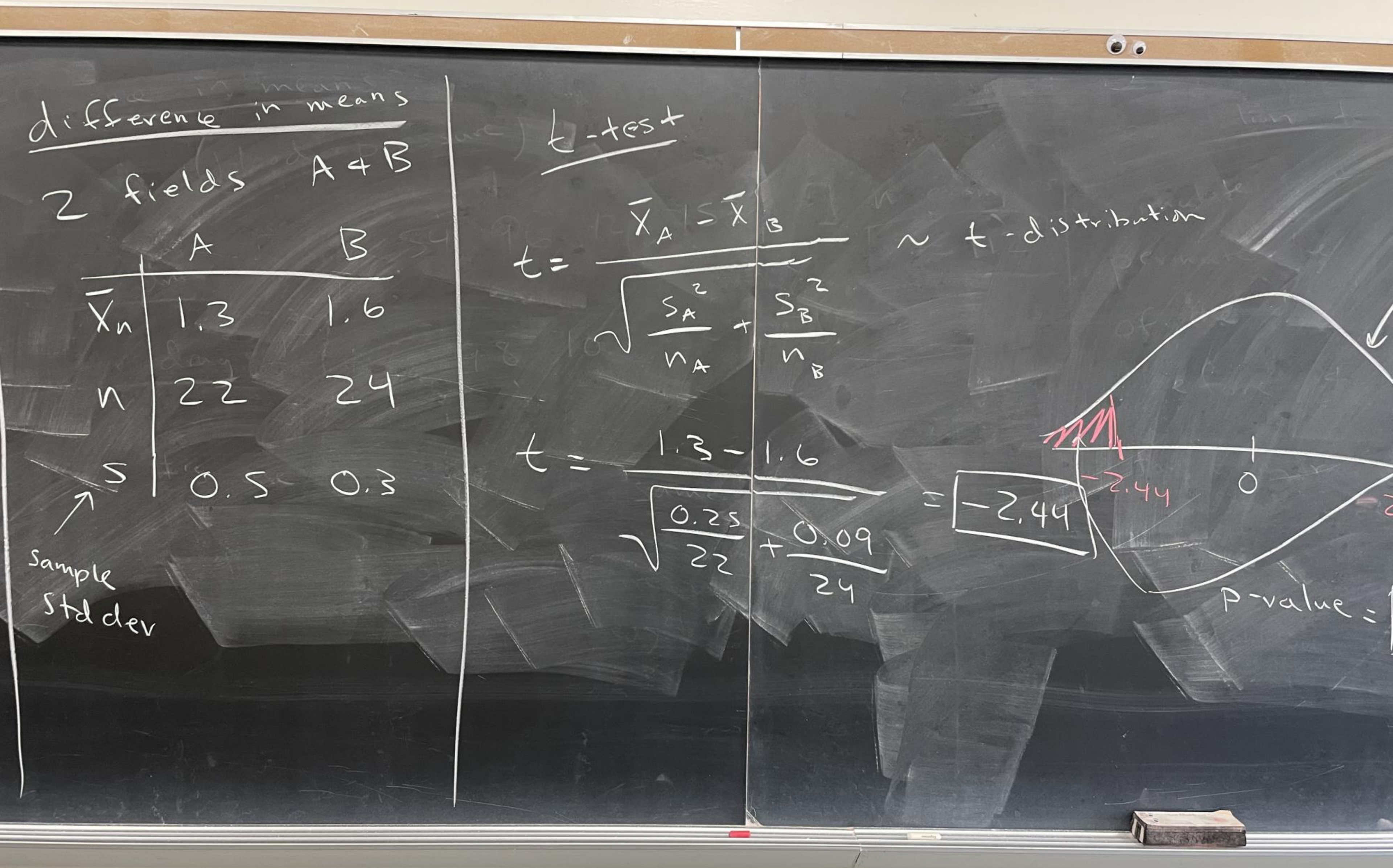
don't know σ? ⇒ use sample variance! (S^2)

$$t = \frac{\bar{X}_n - \mu}{\sqrt{\frac{S^2}{n}}} \sim t\text{-distribution}$$

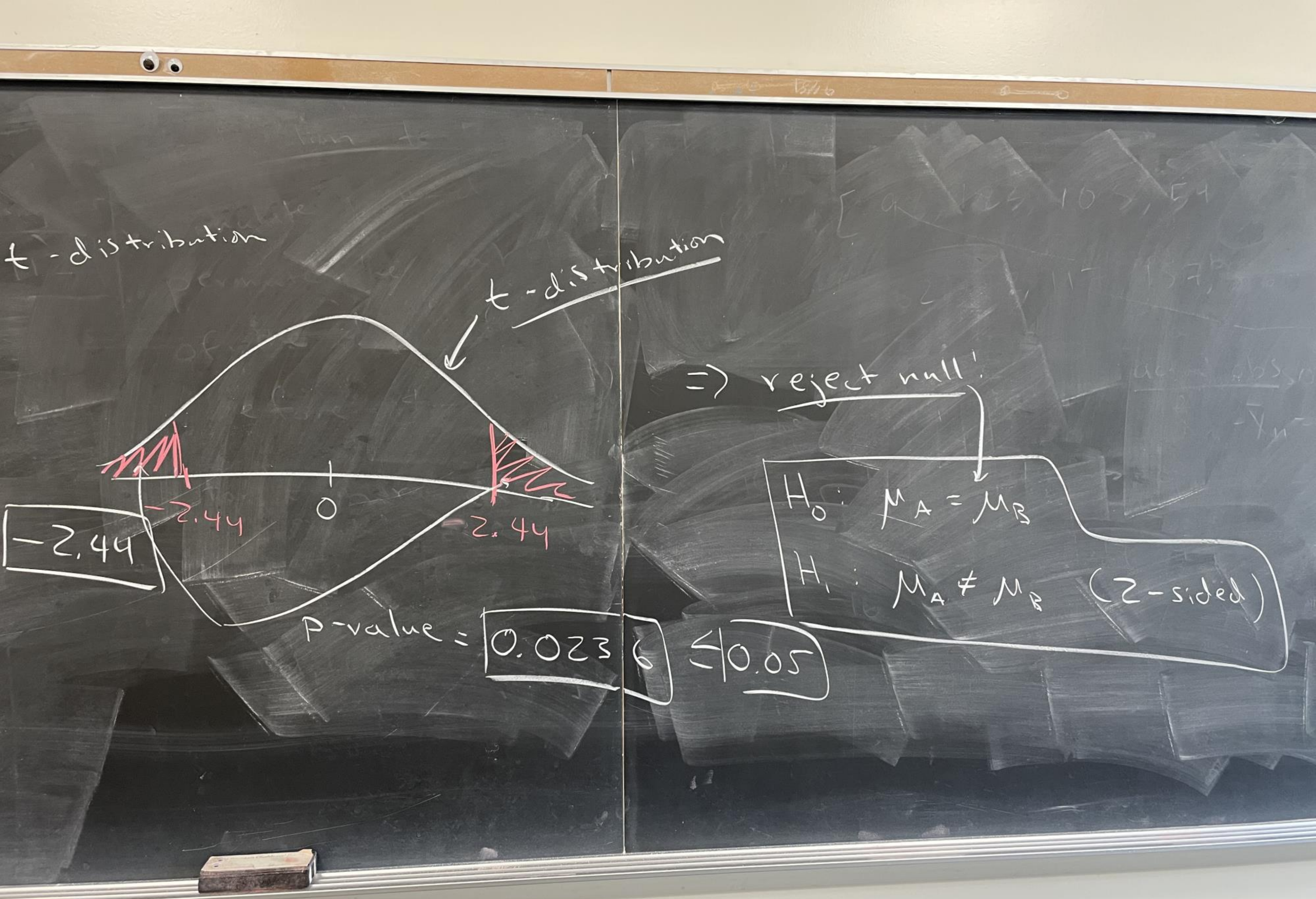


sample variance ⇒ $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}_n$

sample mean



https://www.khanacademy.org/math/ap-statistics/xfb5d8e68:inference-quantitative-means/two-sample-t-test-means/v/two-sample-t-test-for-difference-of-means



https://www.khanacademy.org/math/ap-statistics/xfb5d8e68:inference-quantitative-means/two-sample-t-test-means/v/two-sample-t-test-for-difference-of-means

Outline for today

- Randomized trials for the null distribution
- Are the means of two samples different?
  - t-tests
  - Permutation testing
- Bootstrapping

Next time!