CS 260: Foundations of Data Science

Prof. Thao Nguyen

Fall 2025

HAVERFORD COLLEGE

Materials by Sara Mathieson

Admin

- **Study guide & practice midterm** posted
- **Midterm 2** review next week

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 CLT could be used to estimate a p-value

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

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

- However, this only approximates a p-value

Randomized Trials

- Simulate the distribution under the null hypothesis
- Process:
  1. Run T trials that *mimic* our data under the null hypothesis
  2. Record relevant information for each trial
  3. Count $N_e =$ how many times we observe a result *as or more extreme* than the original data
  4. p-value $= N_e/T$

Data: $n = 20$ die rolls, $\bar{X}_n = 4.2$

- H_0: die is fair
- H_1: die is weighted towards higher values (one-sided)

die example

1 trial: 20 rolls of a fair die

(i.e. mean of the rolls)

any trial with mean ≥ 4.2

T trials

T=100

$$\text{p-value} = \frac{7}{100} = \underbrace{0.07}_{} > 0.05$$

☆ die not unfair

Handout 17

① $T = 20$

② $N_e = 8$

one-sided

13 or fewer

$\boxed{N_e = 12}$

two-sided

13 or fewer

or 17 or more

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

$\boxed{\text{fail to reject } H_0}$

④

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

Outline for today

- Randomized trials for the null distribution

- Are the means of two samples different?
  - t-tests
  - Permutation testing

- Bootstrapping

Difference in means

- Example blood pressure data:
  - Before meds: [117, 54, 96, 123, …] n examples, $\bar{X}_n = 112$
  - After meds: [72, 98, 105, 82, …] m examples, $\bar{X}_m = 96$

- H_0: all #'s are drawn from the same distribution
- H_1: after the medication, blood pressure was lowered (one-sided)

Permutation Testing

- Simulate the null distribution
- Process:
  1. Run T trials that *permute* the "labels" of the data
  2. For each trial, record the *difference in means* between the labels
  3. Count $N_e =$ how many times we observe a result *as or more extreme* than the original data
  4. p-value $= N_e / T$

Permutation testing: simulate null distribution!

permute the "labels" of the data

~~1 trial~~

"before": [98, 123, 105, 54 ...] still n examples $\boxed{\bar{X}_n^{(1)} = 101}$

"after" [82, 72, 117, 157, 96 ...] still m examples $\boxed{\bar{X}_m^{(1)} = 105}$

for t in T trials ($T \approx 1000 - 100{,}000$)

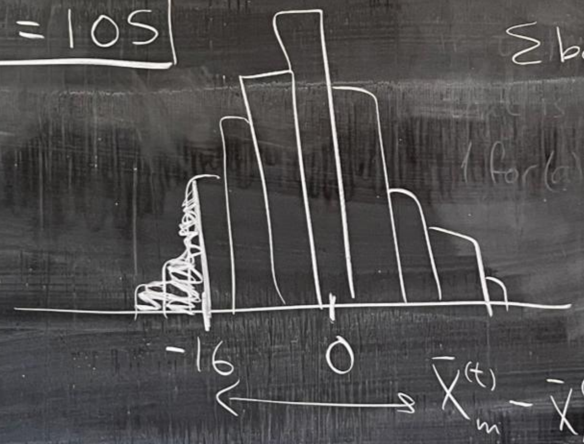
compute $\boxed{\bar{X}_m^{(t)} - \bar{X}_n^{(t)}}$

p-value?

$N_e = \#(\underbrace{\bar{X}_m^{(t)} - \bar{X}_n^{(t)} \leq -16}_{\text{one-sided}})$

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

Σ bars $= T$



t-tests

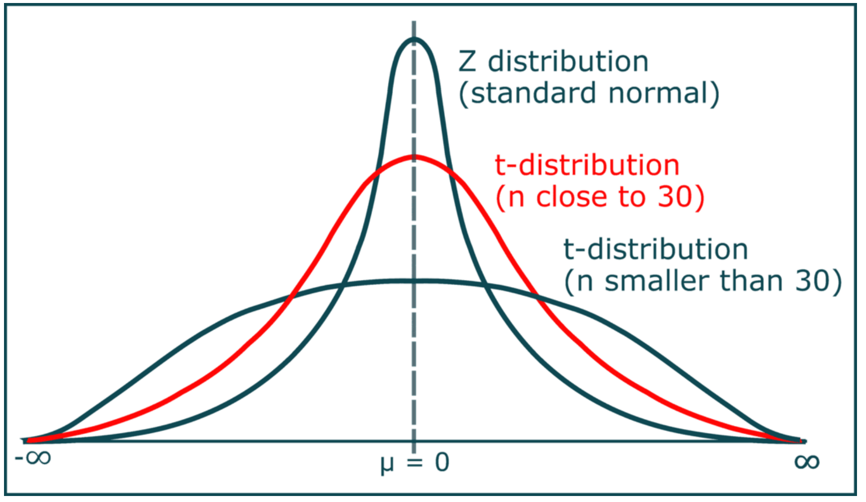
- CLT-inspired test:

drawn from

$$Z = \frac{\overline{X_n} - \mu}{\sqrt{\sigma^2/n}} \sim N(0,1)$$

- Don't know σ^2? Use sample variance

$$t = \frac{\overline{X_n} - \mu}{\sqrt{s^2/n}} \sim \text{t-distribution}$$



https://www.geeksforgeeks.org

difference in means

example pops (Khan)

$H_0: \mu_A = \mu_B$

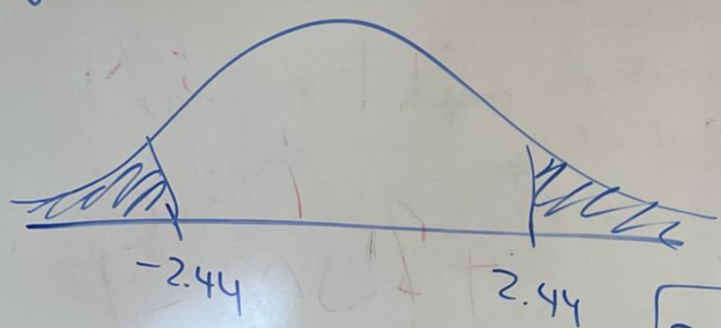
$H_1: \mu_A \neq \mu_B$ (2-sided)

	A	B
$\bar{X}_n$	1.3m	1.6m
S	0.5m	0.3m
n	22	24

sample stddev

$$t = \frac{\bar{X}_A - \bar{X}_B}{\sqrt{\frac{S_A^2}{n_A} + \frac{S_B^2}{n_B}}} \sim t\text{-distribution}$$

$$= \frac{1.3 - 1.6}{\sqrt{\frac{0.25}{22} + \frac{0.09}{24}}} = \boxed{-2.44}$$



$\boxed{\text{p-value} = 0.0236} < 0.05$

reject null!

https://www.khanacademy.org

Outline for today

- Randomized trials for the null distribution

- Are the means of two samples different?
  - t-tests
  - Permutation testing

- Bootstrapping

Example: estimating the mean

Data: ***X*** = [2, 3, 4, 8, 0, 6, 1, 10, 2, 4]

From some distribution with mean μ - we want to learn about μ

Estimate of the mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ = 4

How good is this estimate?

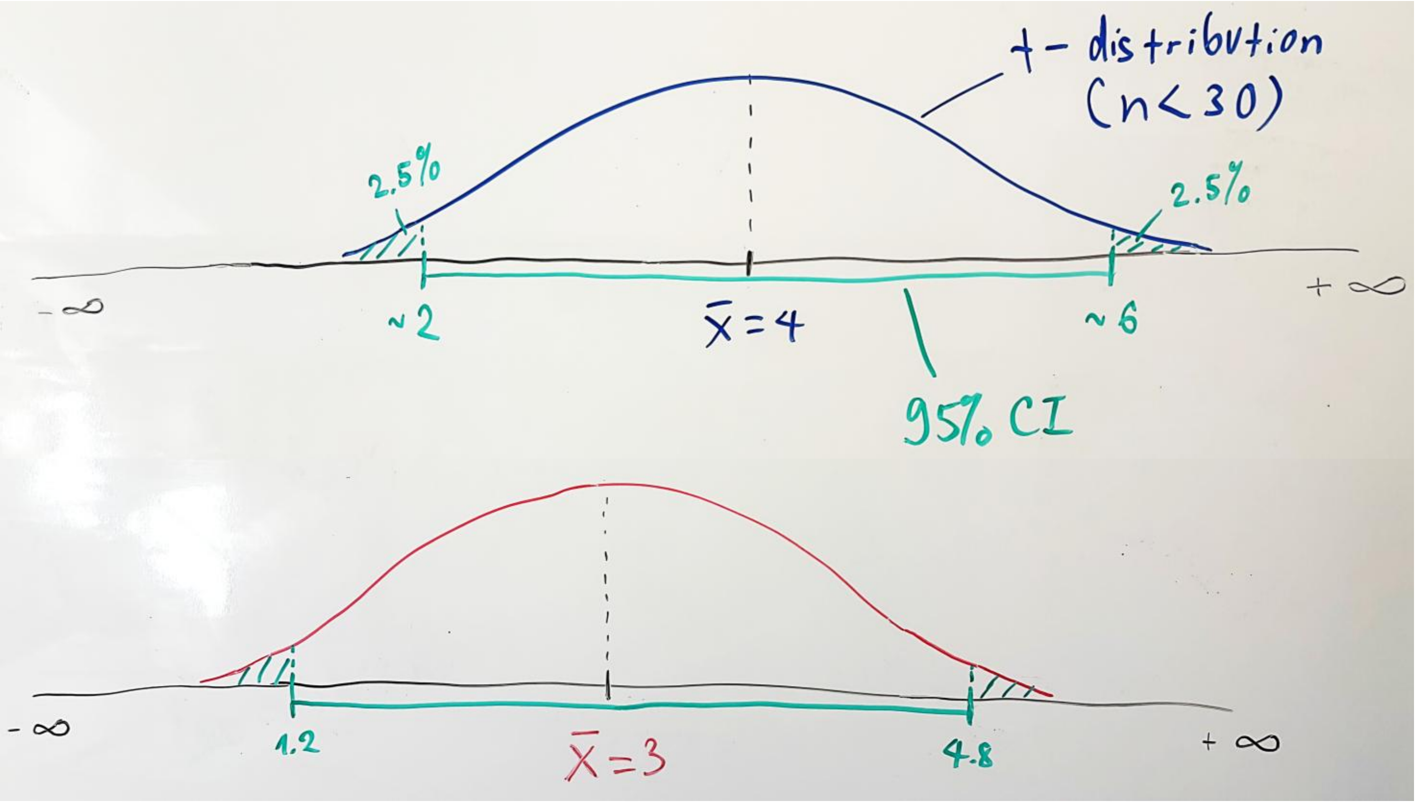
Sample standard deviation $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$ = 3.16

By the central limit theorem, we know that $\bar{X}$ is approximately normally distributed with variance $\frac{s^2}{n}$, so we can construct confidence intervals and p-values for μ etc.

Slide: Iain Mathieson

Confidence Intervals

- Range of most common values from the distribution



- 95% of the time, the 95% CI will contain the true value

The Bootstrap

In an 18th century story by Rudolph Erich Raspe, Baron Munchausen falls to the bottom of a deep lake.

About to drown, he has the idea to lift himself up by pulling on his bootstraps

(In the original German version, he pulls himself up by his hair, left).

Obviously impossible, this story gave its name to a statistical technique (Efron, 1795) that seems magical, in the sense that you can get something (estimates of uncertainty) for nothing!

In general, the bootstrap is an incredibly useful statistical technique – perhaps one of the most useful in all of modern statistics. You should use it <u>everywhere</u>.

Slide: Iain Mathieson

The bootstrap: Resampling

Data: ***X*** = [2, 3, 4, 8, 0, 6, 1, 10, 2, 4]

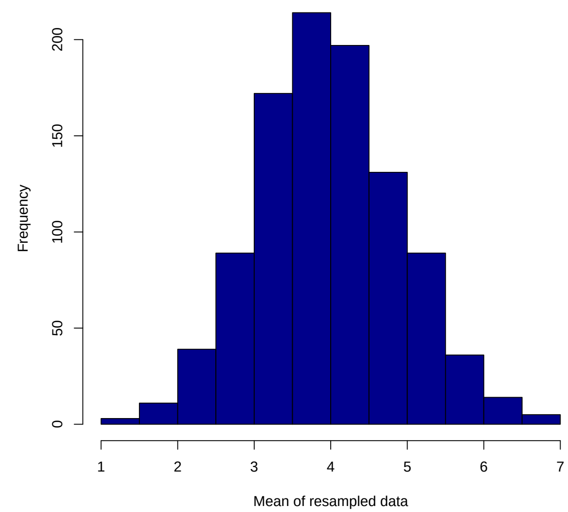
Compute Mean

Resample, with replacement, T times

1 8 2 4 6 10 1 1 1 8	4.2
1 0 1 6 4 1 4 2 1 2	2.2
8 1 6 2 6 4 2 4 10 2	4.5
8 3 4 2 10 8 10 8 6 3	6.2
6 4 6 4 6 4 2 4 3 4	4.3
…	…
…	…

Use the means from the resampled data to estimate the distribution!

95% of the means are between 2.3 and 5.9 (T=1000)



Slide: Iain Mathieson

The bootstrap: Resampling

"Estimate the range (Max—Min)"

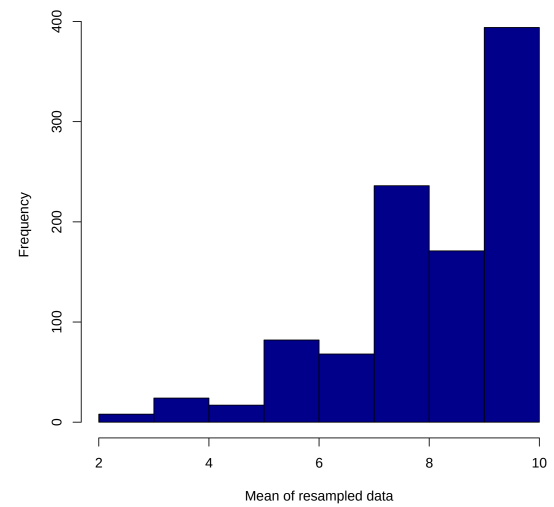
Data: ***X*** = [2, 3, 4, 8, 0, 6, 1, 10, 2, 4]

Compute Range

Resample, with replacement, T times

Resampled data	Range
1 8 2 4 6 10 1 1 1 8	9
1 0 1 6 4 1 4 2 1 2	6
8 1 6 2 6 4 2 4 10 2	9
8 3 4 2 10 8 10 8 6 3	8
6 4 6 4 6 4 2 4 3 4	4
…	…
…	…

Use the ranges from the resampled data to estimate the distribution!



Slide: Iain Mathieson

The bootstrap: Resampling

- The key point is that as long as we can resample our data (which we can always do).
- And calculate the thing we want to estimate (which we can almost always do).
- We can bootstrap anything, and get a sense of how good our estimate is.
- We do not need to make any assumptions about the underlying distribution. For example, to apply the central limit theorem.

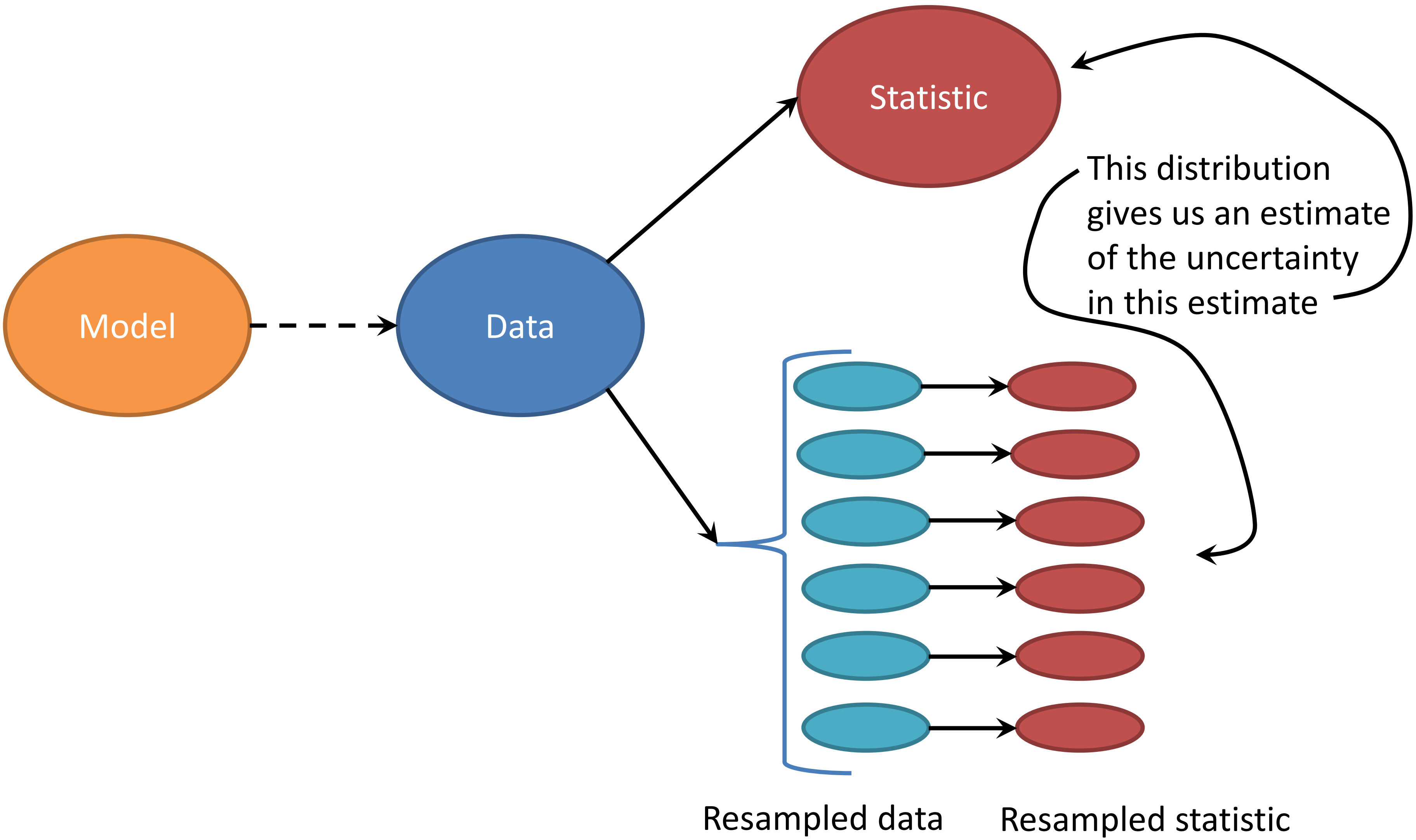
The bootstrap: Resampling

- In general, resampling or permutation method can answer most of the statistical questions that we are interested in (is the mean zero? are these distributions the same?)

- Why then in intro stats did we learn about t-tests, z-scores, and the central limit theorem instead of randomized trials, permutation tests, and bootstrapping?

- Because when statistics was invented in the 1920s, people didn’t have computers!

Slide: Iain Mathieson



Slide: Iain Mathieson