CS 260: Foundations of Data Science

Prof. Sara Mathieson

Fall 2023

HAVERFORD
COLLEGE

Admin

- **Project proposal** due last night
- **Lab 7** due 5pm today
- **Lab 8 posted**, due Wednesday Nov 15
- Next week:
  - Finish statistics, then midterm review
  - Midterm will go out Tuesday Nov 14

Lab 7 notes

cost

$$J(\vec{\omega}) = -\sum_{i=1}^{n} y_i \log h(\vec{x}_i) + (1-y_i) \log(1-h(\vec{x}_i))$$

if $h(\vec{x}_i) = 0$ or $1 - h(\vec{x}_i) = 0$ $\Rightarrow$ ~~$\log(0)$~~ (skip or add 0)

Outline for November 9

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

Outline for November 9

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

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

- However, this only approximates a p-value

rolls

die

the

mean

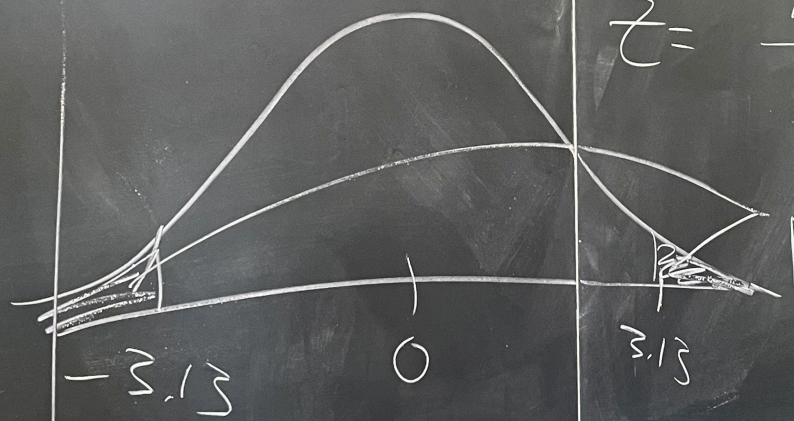
Handout 17

fair coin: $\mu = \frac{1}{2}$, $\sigma^2 = \frac{1}{4}$

reject null hypothesis

our data: $n = 80$, $\bar{X}_n = \frac{54}{80} = 0.675$

$$Z = \frac{\bar{X}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}} = \frac{0.675 - 0.5}{\sqrt{\frac{0.25}{80}}} \approx 3.13$$



p-value $\approx 0.001745 \leq 0.05$

Better Way? randomized trials

simulate the distribution under the null hypothesis

die example | $n = 20$ rolls, $\bar{X}_n = 4.2$ ← our data

H_0 : null hypothesis (fair die) ($\mu = 3.5$)

H_1 : is the die weighted toward higher values?

General Idea

1. run T trials that mimic our data under the null hyp.
2. record relevant info for each trial
3. count how many times you observe a result as or more extreme than original data } N_e
4. $p\text{-value} = \frac{N_e}{T}$

die example

1 trial : 20 rolls of a fair die

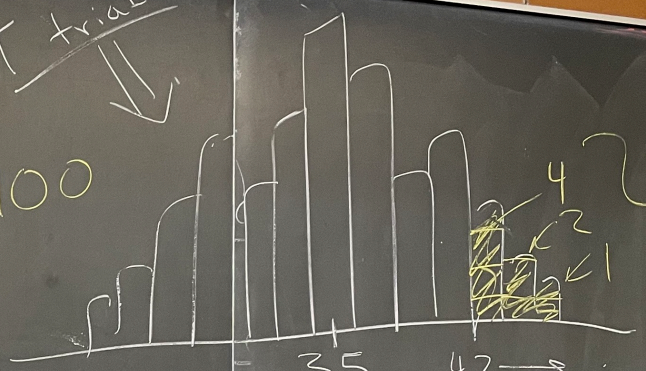
(i.e. mean of the rolls)

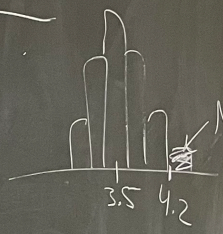
any trial with mean ≥ 4.2

$CLT = 0.033 < 0.05$

T trials

$T = 100$





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

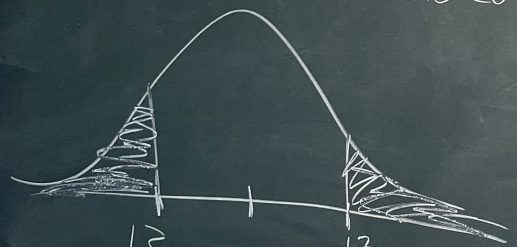
☆ die not unfair

Handout 18

[19, 14, 12, 16, 12, 16, 13, 18, 18, 12, 17, 16, ...

① T = 20 (should be ≈ 1 million)

② N_e = one-sided? $N_e = 8$
two-sided? $N_e = 12$



95% or more extreme

→ two-sided

③ $\frac{12}{20} = 0.6 \gg 0.05$

Difference in mea...

Example: before drug ...
after drug ...

H_0 : all #'s are ...

H_1 : after the drug, ...

Outline for November 9

- Randomized trials for the null distribution

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

- Bootstrapping

13, 18, 18, 12, 17, 16, 11, 12, 12, 15, 15, 16, 15, 13]

Difference in means

blood pressure data

example

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

~~same person~~

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

n examples $\bar{X}_n = 112$

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)

[72 54 105 ...]

[117 98 96 ...]

0.05

simu

perm

Simulate the null dist!

permute the "labels" of the data!

before vs after

Shuffle ⇒ not w/ replacement

1 trial

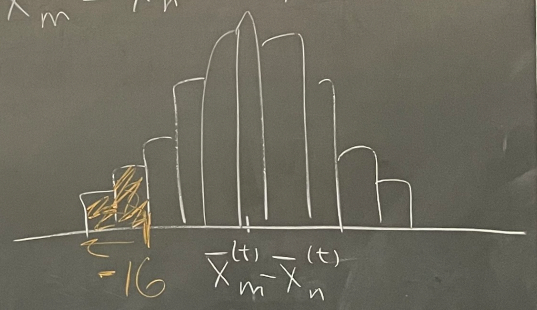
still n
still m

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

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

T times where T is large!

$\bar{X}_m - \bar{X}_n = 96 - 112 = -16$

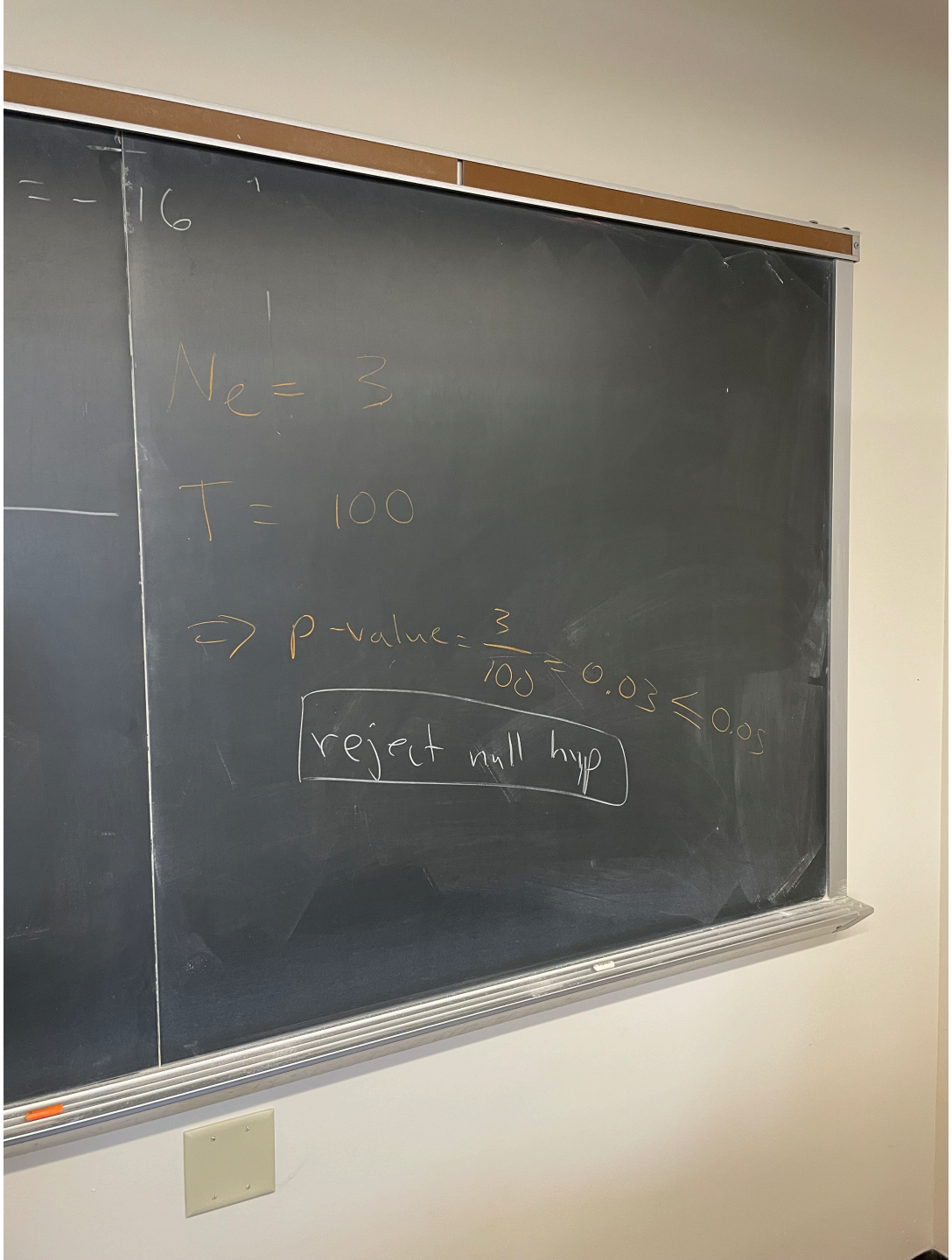


$N_e = 3$

$T = 10$

⇒ p-val

rej


= - 16
Ne = 3
T = 100
⇒ p-value = 3/100 = 0.03 ≤ 0.05
reject null hyp

Outline for November 9

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

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

Example: estimating the mean

Data, X_i = 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?

Standard deviation $\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2}$ = 3.16

By the central limit theorem, we know that $\bar{X}$ is approximately normally distributed with variance $\frac{\sigma^2}{N}$ so we can construct confidence intervals and P-values for μ etc… "95% of the time, the 95% CI will contain the true value". In this case, the 95% CI is 2.1-5.9

Slide: Iain Mathieson

The bootstrap: Resampling

Data, X_i = 2 3 4 8 0 6 1 10 2 4

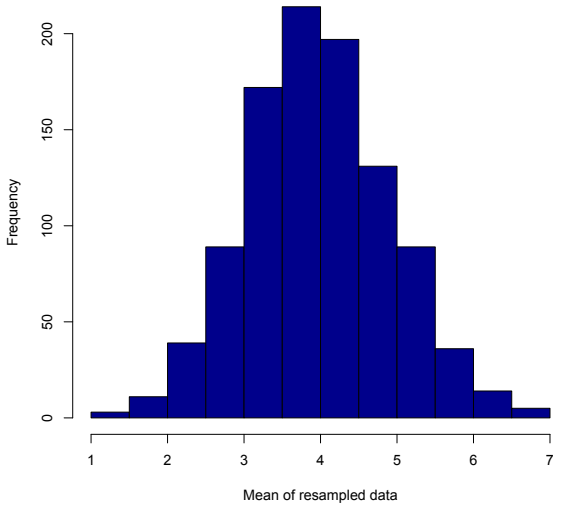
Resample, with replacement, K times

Compute Mean

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 8 1	6.2
6 4 6 4 6 4 2 4 3 4 0	4.3
…	…
…	…

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

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



Slide: Iain Mathieson

The bootstrap: Resampling

"Estimate the range (Max—Min)"

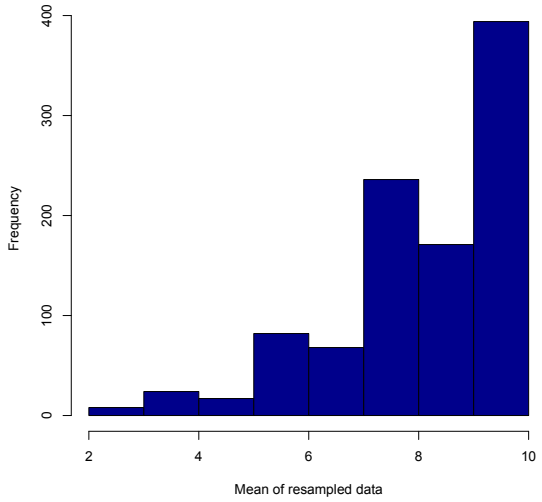
Data, X_i = 2 3 4 8 0 6 1 10 2 4

Compute Range

Resample, with replacement, K times

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 8 1	8
6 4 6 4 6 4 2 4 3 4 0	6
…	…
…	…

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



Slide: Iain Mathieson

The bootstrap: Resampling

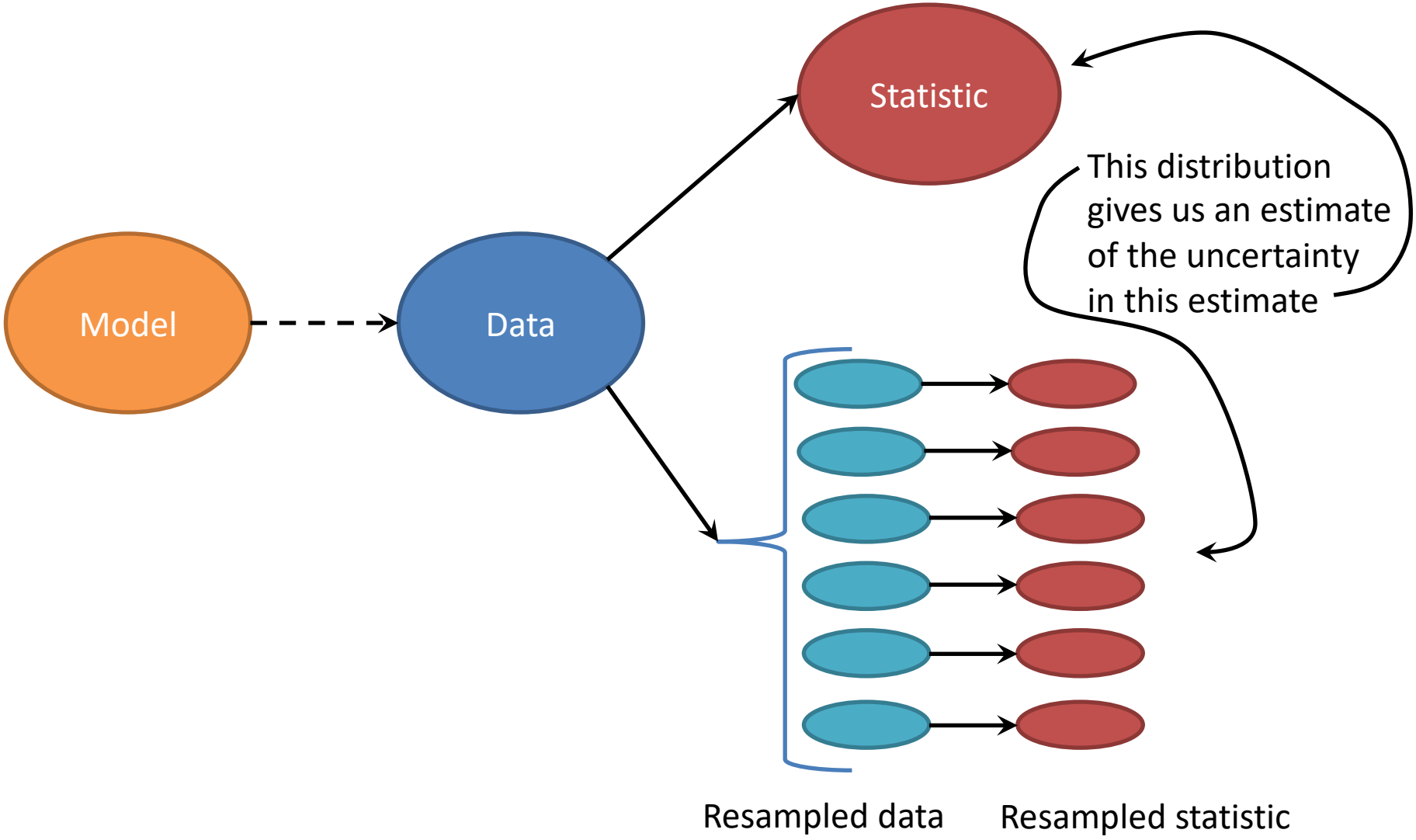
- The key point is that as long as we can resample our data (which we can always do).
- And calculate the thing we want to estimate (which we can almost always do).
- We can bootstrap anything, and get a sense of how good our estimate is.
- We do not need to make any assumptions about the underlying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

Slide: Iain Mathieson

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 permutation tests and bootstrapping?
- Because when statistics was invented in the 1920s, people didn't have computers!

Slide: Iain Mathieson



Slide: Iain Mathieson