CS 260: Foundations of Data Science

Prof. Sara Mathieson

Fall 2021

HAVERFORD
COLLEGE

Admin

- **Note-taker**: Bailey
- **Lab 7 + project proposal** due TONIGHT
  - During lab this week: project meetings with all groups
  - Try to come to the same lab section as your partner
  - We can also discuss Lab 7
- **Lab 8 posted**, due next Thursday Nov 11
- Next week: on **Zoom** (I'll send out info)
  - Finish statistics, then midterm review
  - Midterm will go out Thursday Nov 11

Outline for November 4

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

Outline for November 4

- Randomized trials for the null distribution

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

- Bootstrapping

- Random forests

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

Better way? randomized trials

simulate the distribution under the null hypothesis

die example : 20 rolls, mean: 4.2 our observed data

"true"

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

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

General idea:

1 trial: 20 rolls of a fair die

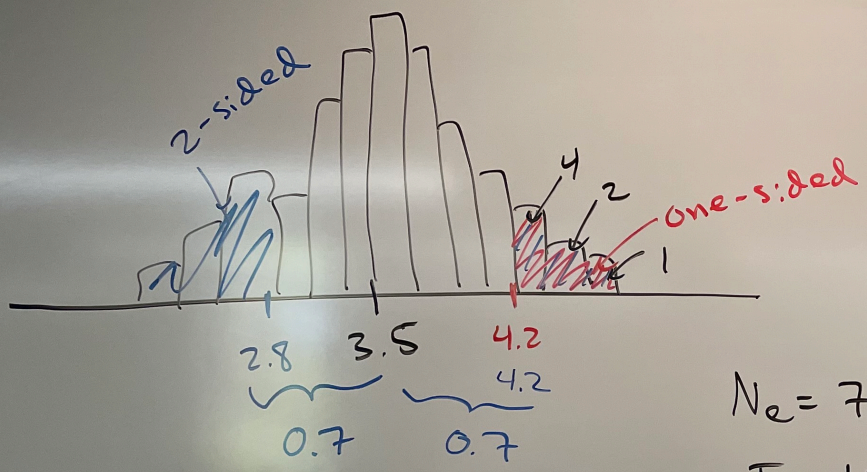
① run T trials that mimic our data under the null hypothesis

② record relevant information for each trial (i.e. mean of the rolls)

③ count how many times you observe a result as or more extreme than your data (any trial w/ mean ≥ 4.2)

N_e # extreme

if 2-sided => also count mean ≤ 2.8



difference = $|4.2 - 3.5| = 0.7$

(4) p-value: $\frac{N_e}{T}$ ☆

$N_e = 7$

$T = 100$

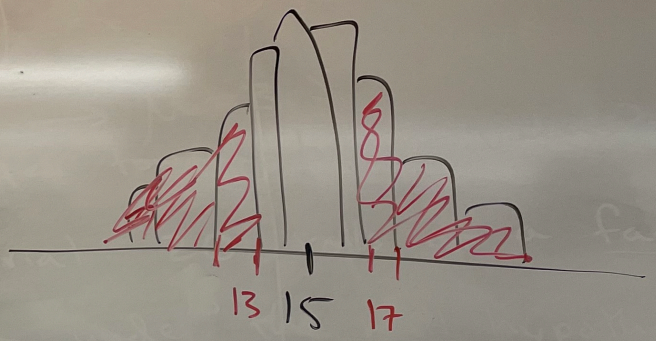
p-value: $0.07 > 0.05$ (= significance threshold)

$\Rightarrow$ fail to reject H_0

Handout 18

[19, 14, 12, 16, 12, 16, 13, 18, 18, 12, 17, 16, 11, 12, 12, 15, 15, 16, 15, 13]

$T = 20$, H_0: expect: 15 Heads



N_e: # heads ≥ 17
" ≤ 13
} $N_e = 12$

p-value $= \frac{12}{20} = 0.6$ fail to reject null hypothesis

Outline for November 4

- Randomized trials for the null distribution

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

- Bootstrapping

- Random forests

Difference in means

blood sugar

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 ??

don't know μ or σ^2

H_1 : after the drug, mean was lower (one-sided)

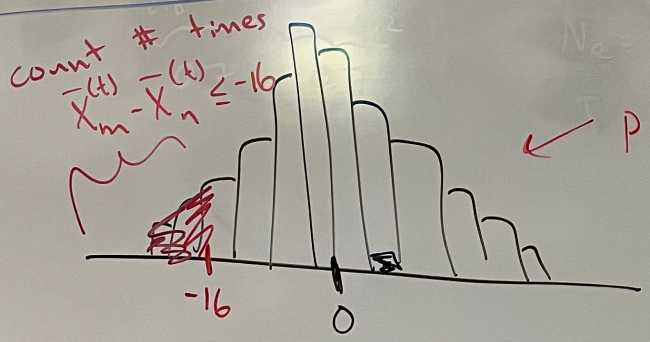
Simulate null distribution! => permute "labels" of data (i.e. before or after)

1 trial

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

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

do for T trials => plot $\bar{X}_m - \bar{X}_n$



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

Can also use a CLT-inspired test

$\Rightarrow$ t-test

Sample variance

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

CLT:

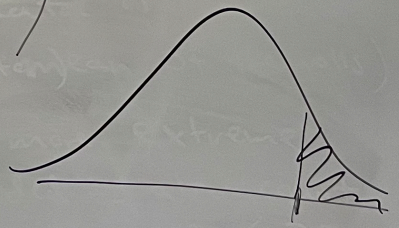
$$\sqrt{n}\left(\frac{\bar{X} - \mu}{\sigma}\right) \sim \underbrace{N(0,1)}_{\text{standard normal}}$$

$$\underbrace{\sqrt{n}\left(\frac{\bar{X} - \mu}{S}\right)}_{\text{z-score}} \sim \text{t-distribution}$$

S → sample variance $= S^2$

Variance

$$\underbrace{\sigma^2 = Var(X) = E[(X - \mu)^2]}_{\text{in theory}}$$



Outline for November 4

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

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

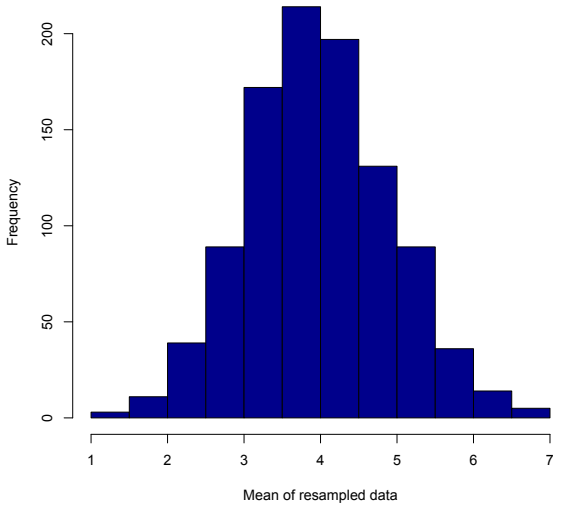
Resample, with replacement, K times

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)"

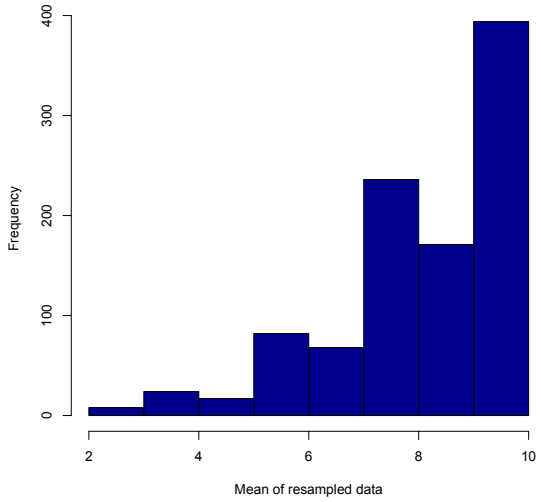
Data, X_i = 2 3 4 8 0 6 1 10 2 4

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

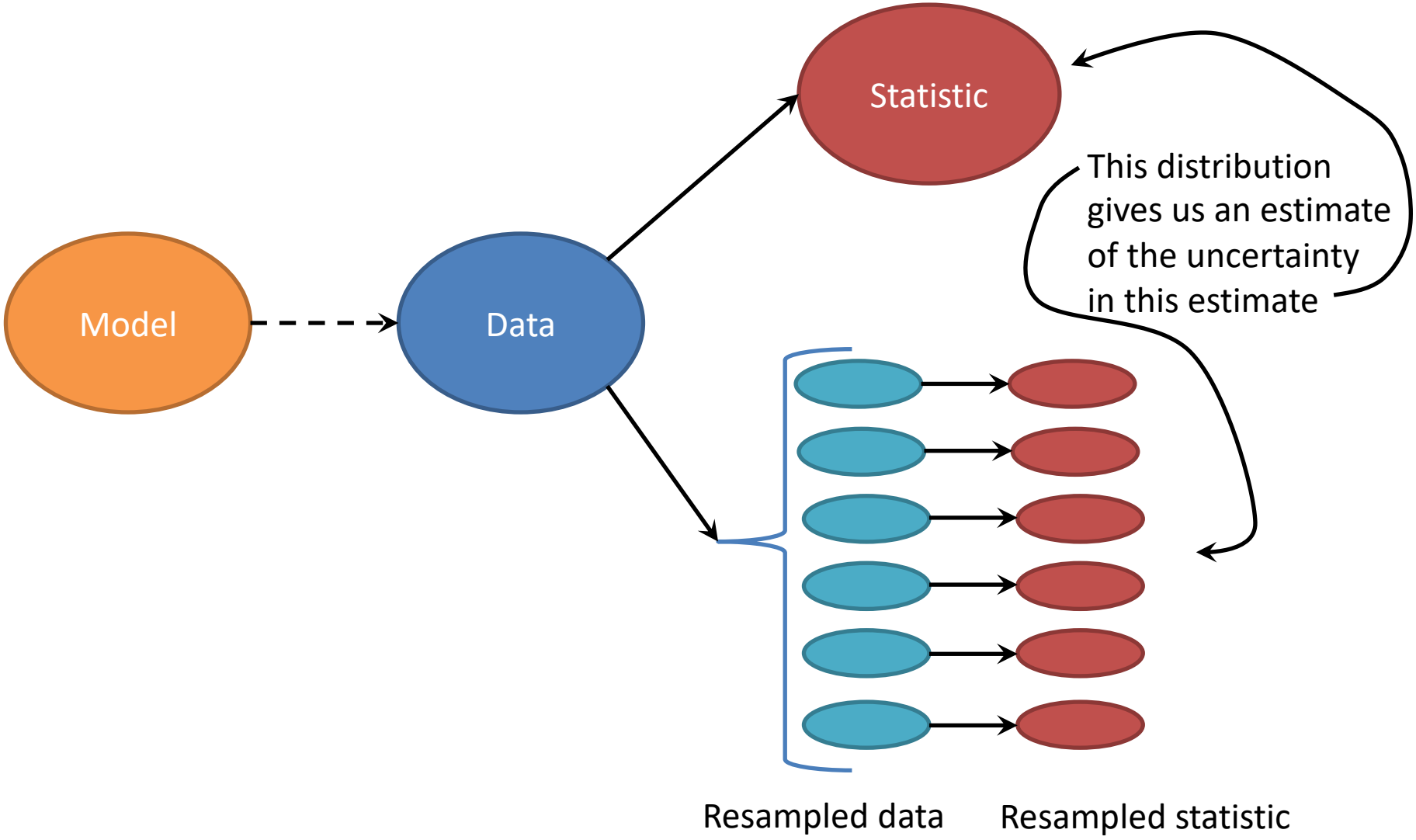
- The key point is that as long as we can resample our data (which we can always do).
- And calculate the thing we want to estimate (which we can almost always do).
- We can bootstrap anything, and get a sense of how good our estimate is.
- We do not need to make any assumptions about the underlying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

Outline for November 4

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

Next time!