CS 260: Foundations of Data Science

Prof. Thao Nguyen

Fall 2024

HAVERFORD
COLLEGE

Guest lecture: Sara Mathieson

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

Recap Handout 17

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

(2) $Var(X) = E[(X-\mu)^2]$

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

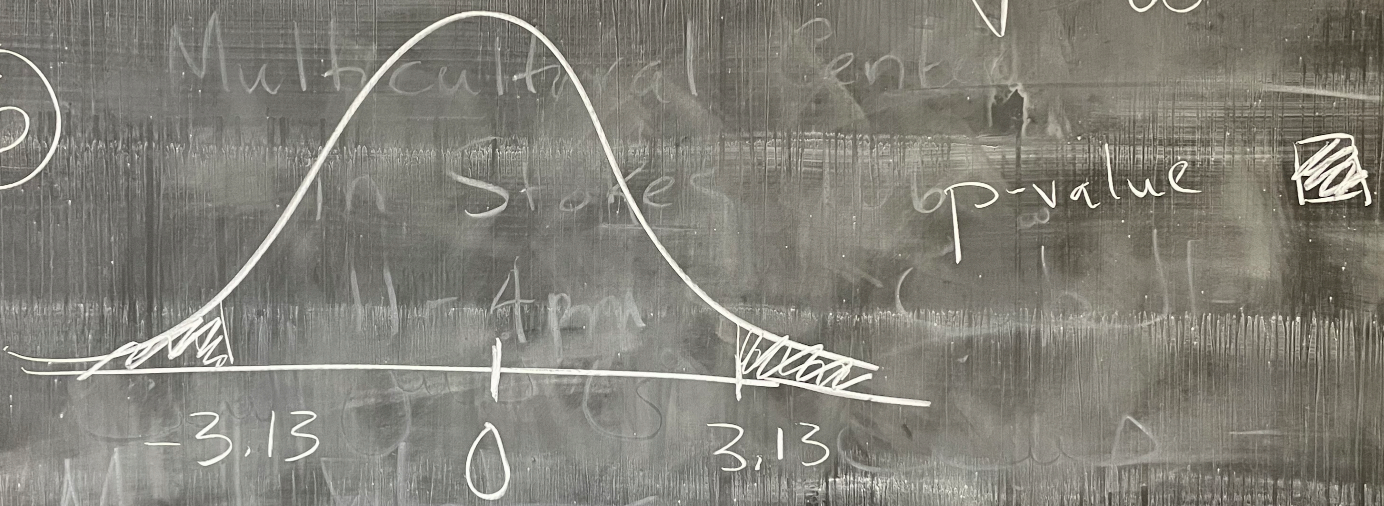
$= \boxed{\frac{1}{4}} = \sigma^2$

$\boxed{\frac{1}{6}(1+2 \ldots 6) = 3.5}$

(3) 54 heads $\Rightarrow \bar{X}_n = \frac{54}{80} \approx \boxed{0.675}$

(4) $$z = \frac{\bar{X}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}} = \frac{0.675 - 0.5}{\sqrt{\frac{0.25}{80}}} \approx \boxed{3.13}$$

(5)



$$\text{p-value} = \boxed{0.001745} \le 0.05$$

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

   roll a fair die

   Right now: each group does 1 trial!

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

3.5	2.4	4.1
4.5	3.4	
3.5	3.3	3.6
3.2	3.1	3.9
3.6	3.7	2.6
4.0	3.6	
3.8	3.9	
3.9	3.2	

$T = 20$

$N_e = 1$

$$\text{p-value} = \frac{1}{20} = \boxed{0.05}$$

$T = 1000$

$N_e = 30$

$$\text{p-value} = \frac{30}{1000} = \boxed{0.03}$$

1	2	3	4	5	6	
0.2	0.2	0.3	0.1	0.1	0.1	} Sum = 1

Handout 18

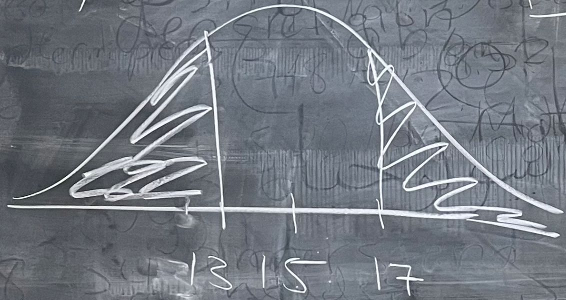
[19, 14, 12, 16, 12, 16, 13, 18, 18, 12, 17, 16, 11, 12, 12, 15,
15, 16, 15, 13]

$\boxed{n = 20}$

$\mu = 15$

$\boxed{N_e = 12}$ (2-sided)

$N_e = 8$ (1-sided)



$$p\text{-value} = \frac{12}{20} = 0.6$$

not surprising

not statistically significant
at $\alpha = 0.05$ confidence level

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

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

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

after drug: [72, 98, 105, 82, ...], $\bar{X}_m = 96$

m examples

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

H_1: after the drug, blood pressure went down (one-sided)

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

=96

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 ($T \approx 1000 - 100{,}000$)

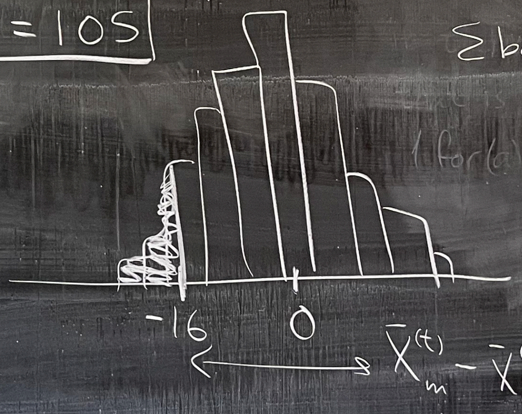
compute $\boxed{\bar{X}_m^{(t)} - \bar{X}_n^{(t)}}$

p-value?

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

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

Σ bars $= T$



t-

t-tests (don't know σ^2?)

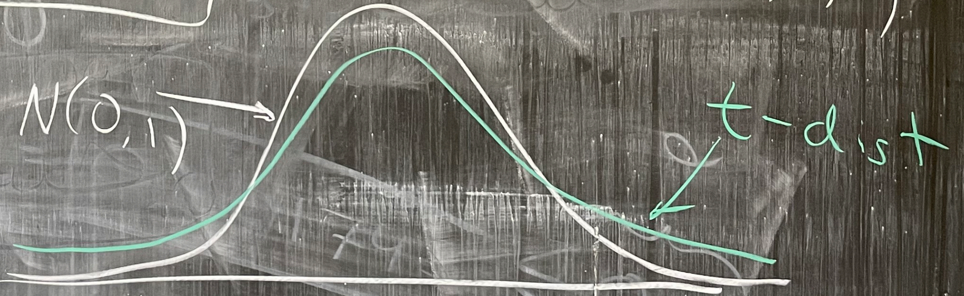
Sample variance

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

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

(in some cases $N(0,1)$)

like z but when you don't know variance



difference in means

example pops (Khan)

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

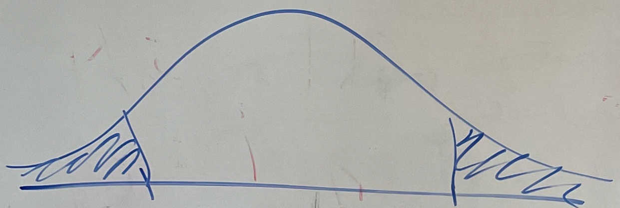
sample stddev

$H_0: \mu_A = \mu_B$

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

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

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



-2.44 2.44

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

reject null!

Outline for today

- 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 everywhere.

Slide: Iain Mathieson

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… “95% of the time, the 95% CI will contain the true value”.

The bootstrap: Resampling

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

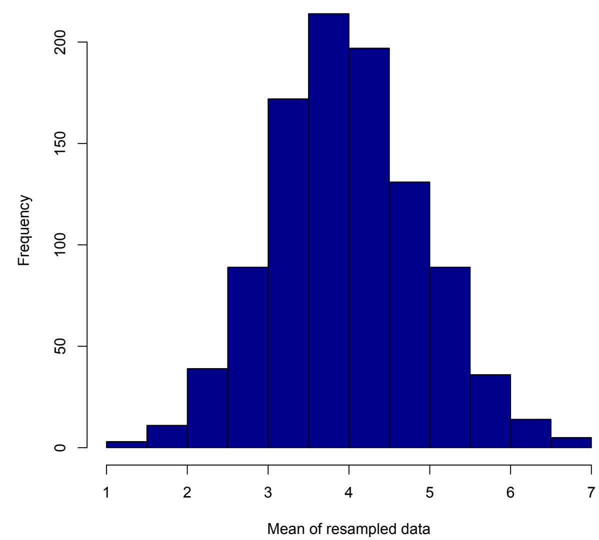
Resample, with replacement, T times

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 (T=1000)



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

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

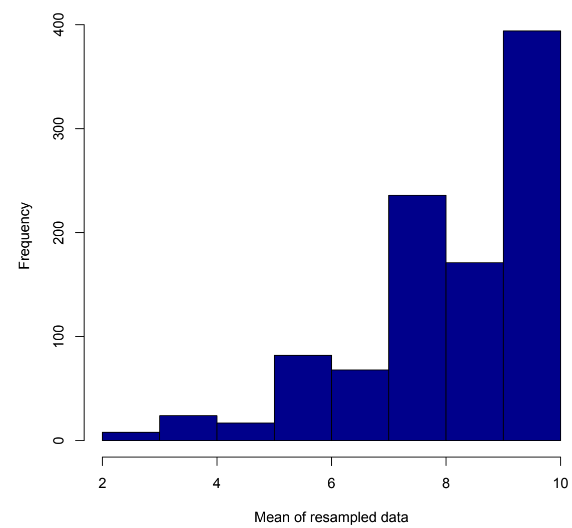
8 3 4 2 10 8 10 8 8 1 → 8

6 4 6 4 6 4 2 4 3 4 0 → 6

… → …

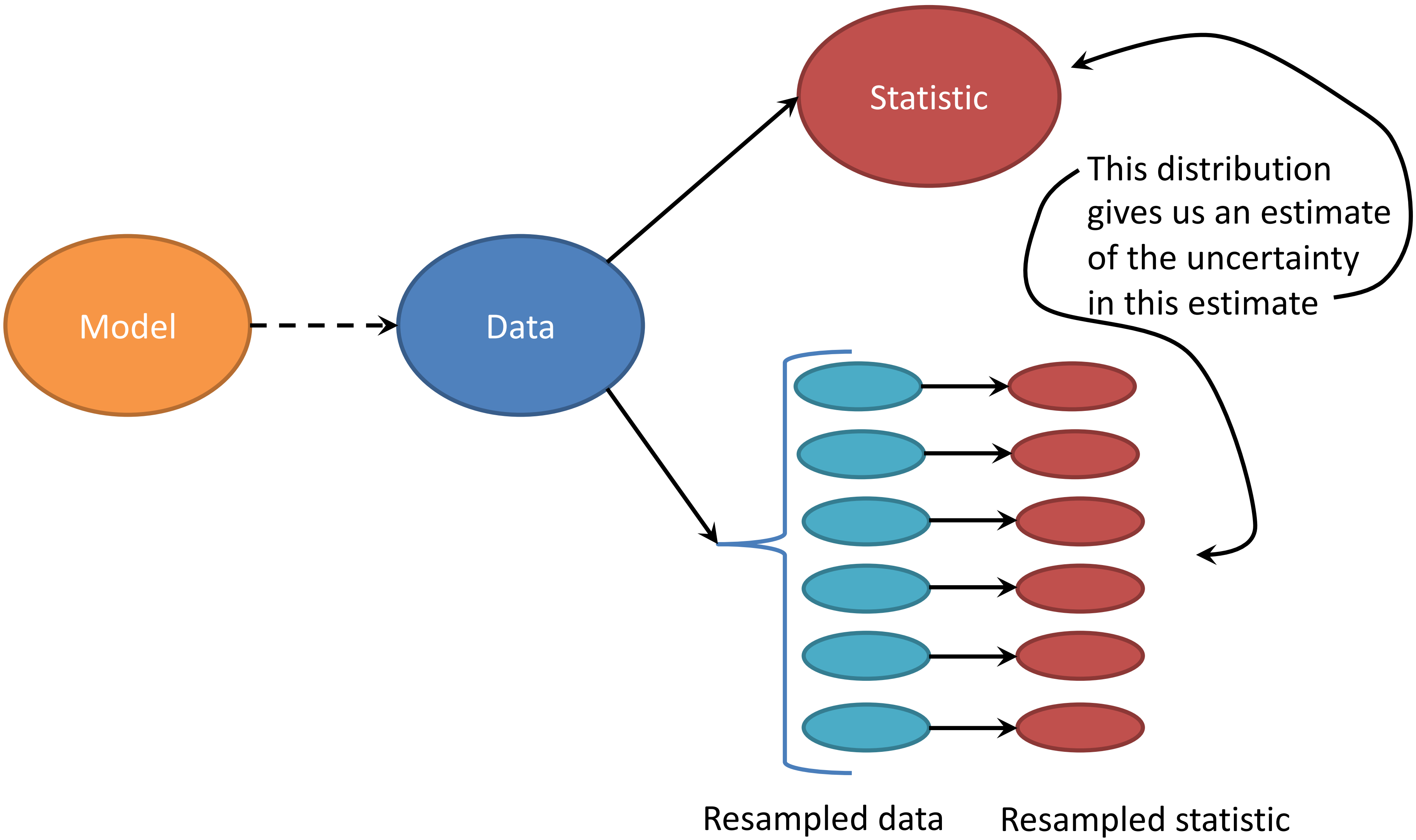
… → …

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



Slide: Iain Mathieson

The bootstrap: Resampling



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 permutation tests and bootstrapping?

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

Slide: Iain Mathieson

Bootstrap example

<u>Setup</u>: you obtain 0.87 accuracy on a test dataset using a new algorithm

<u>Goal</u>: find a 95% confidence interval for your estimate