

CS 260 – Foundations of Data Science

Lecture 20– Bootstrap

11/17/2026

Adam Poliak

BRYN MAWR COLLEGE

Announcements

HW08 due tonight 11/17

Midterm 2 – Tuesday 11/24

Definition of the P-value

Formal name: **observed significance level**

The *P*-value is the chance,

- Under the null hypothesis,
- That the test statistic
- Is equal to the value that was observed in the data
- Or is even further in the direction of the tail

How far from the average

Standard deviation (SD) measures roughly
how far the data are from their average

SD = root mean square of deviations from average

Steps: 5 4 3 2 1

SD has the same units as the data

Why use Standard Deviation?

There are two main reasons.

The first reason:

- No matter what the shape of the distribution, the bulk of the data are in the range “average plus or minus a few SDs”

The second reason:

- Relation with the bellshaped curve



Chebyshev's Inequality

BRYN MAWR
COLLEGE

How big are most values?

No matter what the shape of the distribution,
the bulk of the data are in the range "average ± a few SDs"

Chebyshev's Inequality

No matter what the shape of the distribution,
the proportion of values in the range "average ± *z* SDs" is
at least 1 - 1/*z*^2

Chebyshev's Bounds

the proportion of values in the range "average ± *z* SDs" is at least 1 - 1/*z*2

Range	Proportion

Chebyshev's Bounds

the proportion of values in the range "average ± z SDs" is at least 1 - 1/z2

Range	Proportion
average ± 2 SDs	at least 1 - 1/4 (75%)

Chebyshev's Bounds

the proportion of values in the range "average ± z SDs" is at least 1 - 1/z2

Range	Proportion
average ± 2 SDs	at least 1 - 1/4 (75%)
average ± 3 SDs	at least 1 - 1/9 (88.888...%)

Chebyshev's Bounds

the proportion of values in the range "average ± z SDs" is at least 1 - 1/z2

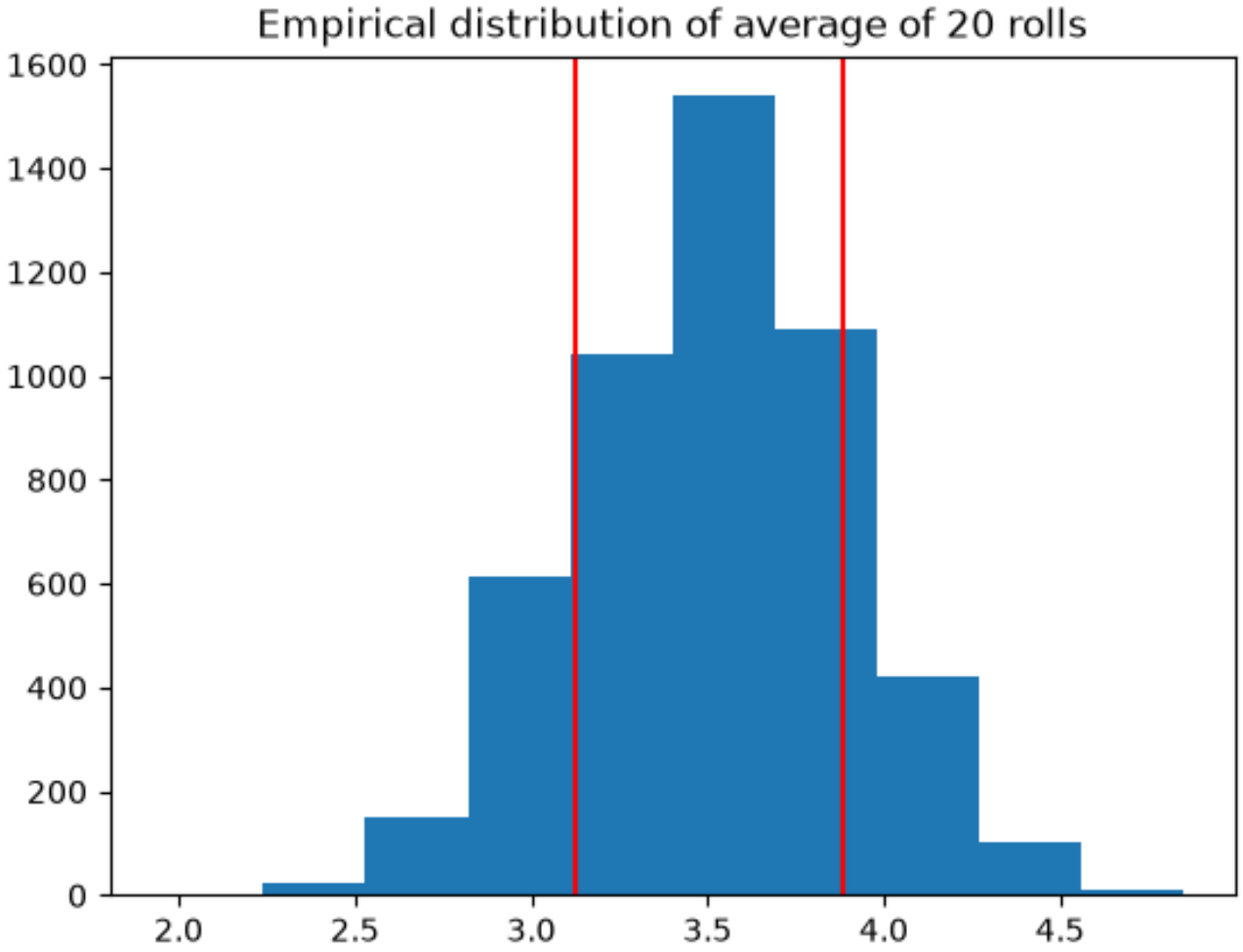
Range	Proportion
average ± 2 SDs	at least 1 - 1/4 (75%)
average ± 3 SDs	at least 1 - 1/9 (88.888...%)
average ± 4 SDs	at least 1 - 1/16 (93.75%)

Chebyshev's Bounds

the proportion of values in the range "average ± *z* SDs" is at least 1 - 1/*z*2

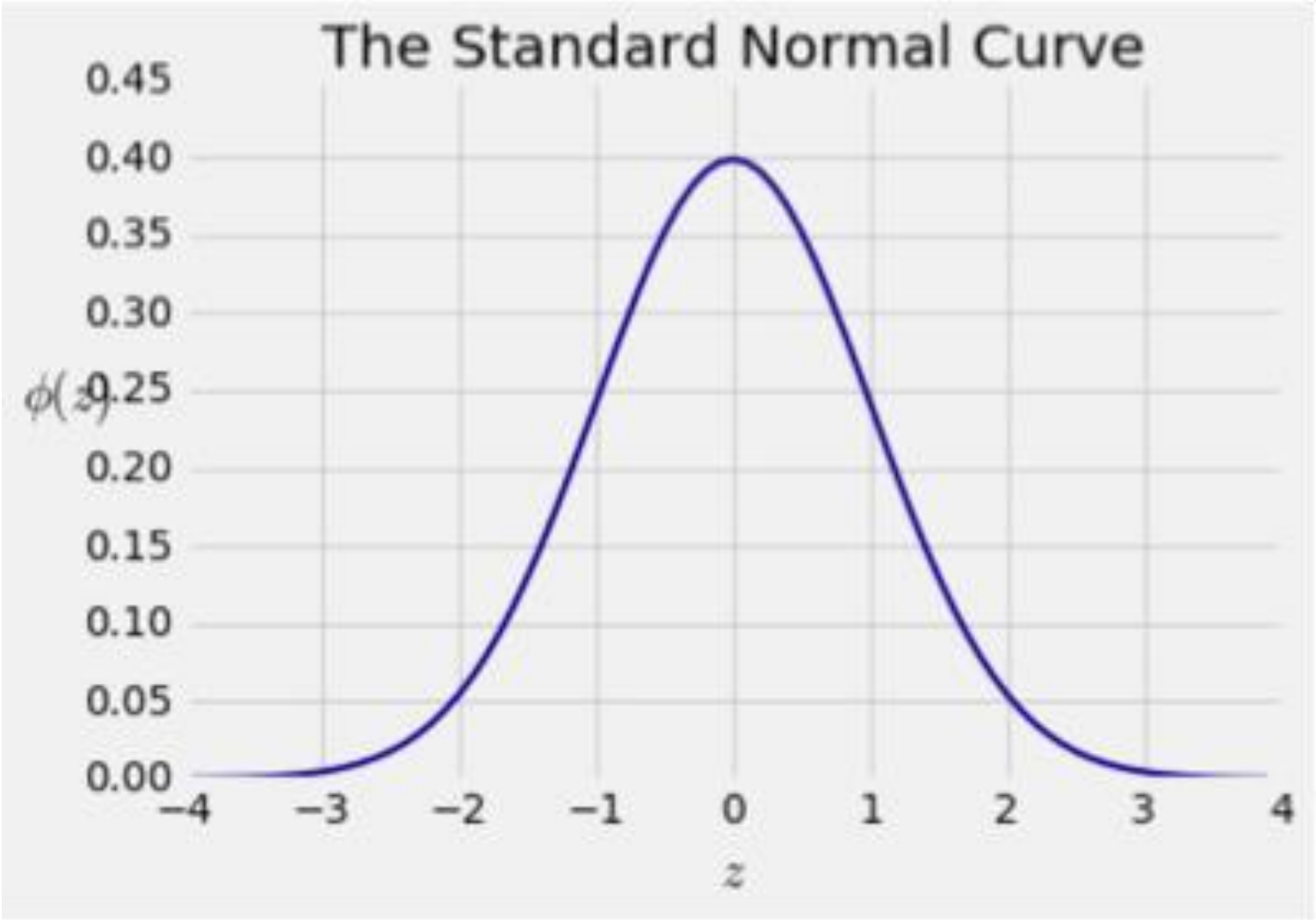
Range	Proportion
average ± 2 SDs	at least 1 - 1/4 (75%)
average ± 3 SDs	at least 1 - 1/9 (88.888...%)
average ± 4 SDs	at least 1 - 1/16 (93.75%)
average ± 5 SDs	at least 1 - 1/25 (96%)

True no matter what the distribution looks like



Mean is 3.5,
Std 0.379
67% of the empirical distribution lies between the red lines

Bell Curve


The Standard Normal Curve
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
$\phi(z)$
−4
−3
−2
−1
0
1
2
3
4
z

How Big are Most of the Values

No matter what the shape of the distribution,
the bulk of the data are in the range "average ± a few SDs"

If a histogram is bell-shaped, then
Almost all of the data are in the range "average ± 3 SDs

Bounds and Approximations

Percent in Range	All Distributions	Normal Distributions
Average +- 1 SD	At least 0%	About 68%
Average +- 2 SDs	At least 75%	About 95%
Average +- 3 SDs	At least 88.888...%	About 99.73%

Standard Units



brynmawr.edu

BRYN MAWR COLLEGE

Standard Units

How many SDs above average?

***z* = (value - average)/SD**

- Negative z: value below average
- Positive z: value above average
- z = 0: value equal to average

When values are in standard units:

average = 0, SD = 1

Chebyshev: At least 96% of the values of *z* are between -5 and 5

Questions

What whole numbers are closest to

(1) Average age

(2) The SD of ages

Age in Years	Age in Standard Units
27	-0.0392546
33	0.992496
28	0.132704
23	-0.727088
25	-0.383171
33	0.992496
23	-0.727088
25	-0.383171
30	0.476621
27	-0.0392546

Answers

What whole numbers are closest to

(1) Average age is close to 27
(standard unit here is closest to 0)

(1) The SD of ages is about 6 years
(standard unit at 33 is closest to 1)

Age in Years	Age in Standard Units
27	-0.0392546
33	0.992496
28	0.132704
23	-0.727088
25	-0.383171
33	0.992496
23	-0.727088
25	-0.383171
30	0.476621
27	-0.0392546

Where to Get Another Sample?

We want to understand errors of our estimate

Given the **population**, we could simulate

- ...but we only have the **sample**!

To get many values of the estimate, we needed many random samples

Can't go back and sample again from the population:

- No time, no money

Stuck?

Boostrap



BRYN MAWR
COLLEGE

The Bootstrap

A technique for simulating repeated random sampling

All that we have is the original sample

- … which is large and random
- Therefore, it probably resembles the population

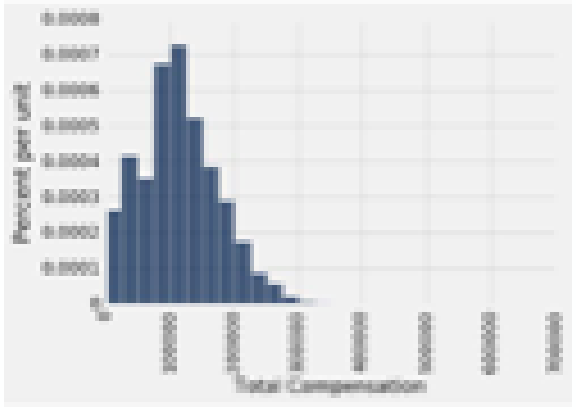
So we sample at random from the original sample!

How the Bootstrap Works



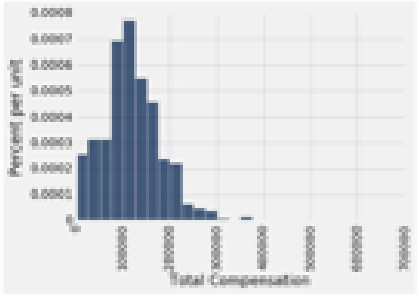
Why the Bootstrap Works

Population

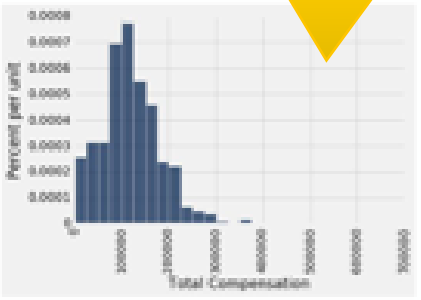


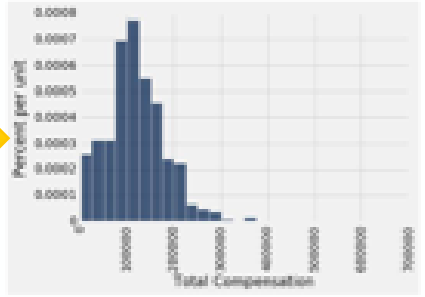
What we wish we could get

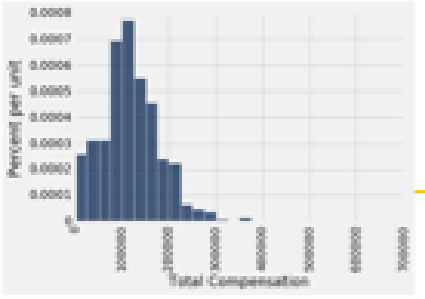
Sample



Resamples







What we actually can get

Real World vs Bootstrap World

Real World

- True probability distribution (population)
  - Random sample 1
    - Estimate 1
  - Random sample 2
    - Estimate 2
  - …
  - Random sample 1000
    - Estimate 1000

Bootstrap World

- Empirical distribution of original sample (“population”)
  - Bootstrap sample 1
    - Estimate 1
  - Bootstrap sample 2
    - Estimate 2
  - …
  - Bootstrap sample 1000
    - Estimate 1000

Hope: these two scenarios are analogous

The Bootstrap Principle

The bootstrap principle:

- **Bootstrap-world** sampling ≈ **Real-world** sampling

Not always true!

- … but reasonable if sample is large enough

We hope that:

a) Variability of bootstrap estimate

b) Distribution of bootstrap errors

…are similar to what they are in the real world

Key to Resampling

From the original sample,

- draw at random
- with replacement
- as many values as the original sample contained

The size of the new sample has to be the same as the original one, so that the two estimates are comparable

Confidence Intervals

95% Confidence Interval

Interval of **estimates of a parameter**

Based on random sampling

95% is called the confidence level

- Could be any percent between 0 and 100
- Higher level means wider intervals

The **confidence is in the process** that gives the interval:

- It generates a “good” interval about 95% of the time

Confidence Interval without simulation

$$\text{estimate} \pm \text{critical value} \times \text{standard deviation}$$

Use Methods Appropriately

Can You Use a CI Like This?

By our calculation, an approximate 95% confidence interval for the average age of the mothers in the population is (26.9, 27.6) years.

True or False:

- About 95% of the mothers in the population were between 26.9 years and 27.6 years old.

Answer:

- **False.** We're estimating that their **average age** is in this interval.

Is This What a CI Means?

An approximate 95% confidence interval for the average age of the mothers in the population is (26.9, 27.6) years.

True or False:

There is a 0.95 probability that the average age of mothers in the population is in the range 26.9 to 27.6 years.

Answer:

False. The average age of the mothers in the population is unknown but it's a constant. It's not random. No chances involved

When *Not* to use the Bootstrap

if you're trying to estimate very high or very low percentiles, or min and max

If you're trying to estimate any parameter that's greatly affected by rare elements of the population

If the probability distribution of your statistic is not roughly bell shaped (the shape of the empirical distribution will be a clue)

If the original sample is very small

Using a CI For Hypothesis Testing

Null hypothesis: **Population average = *x***

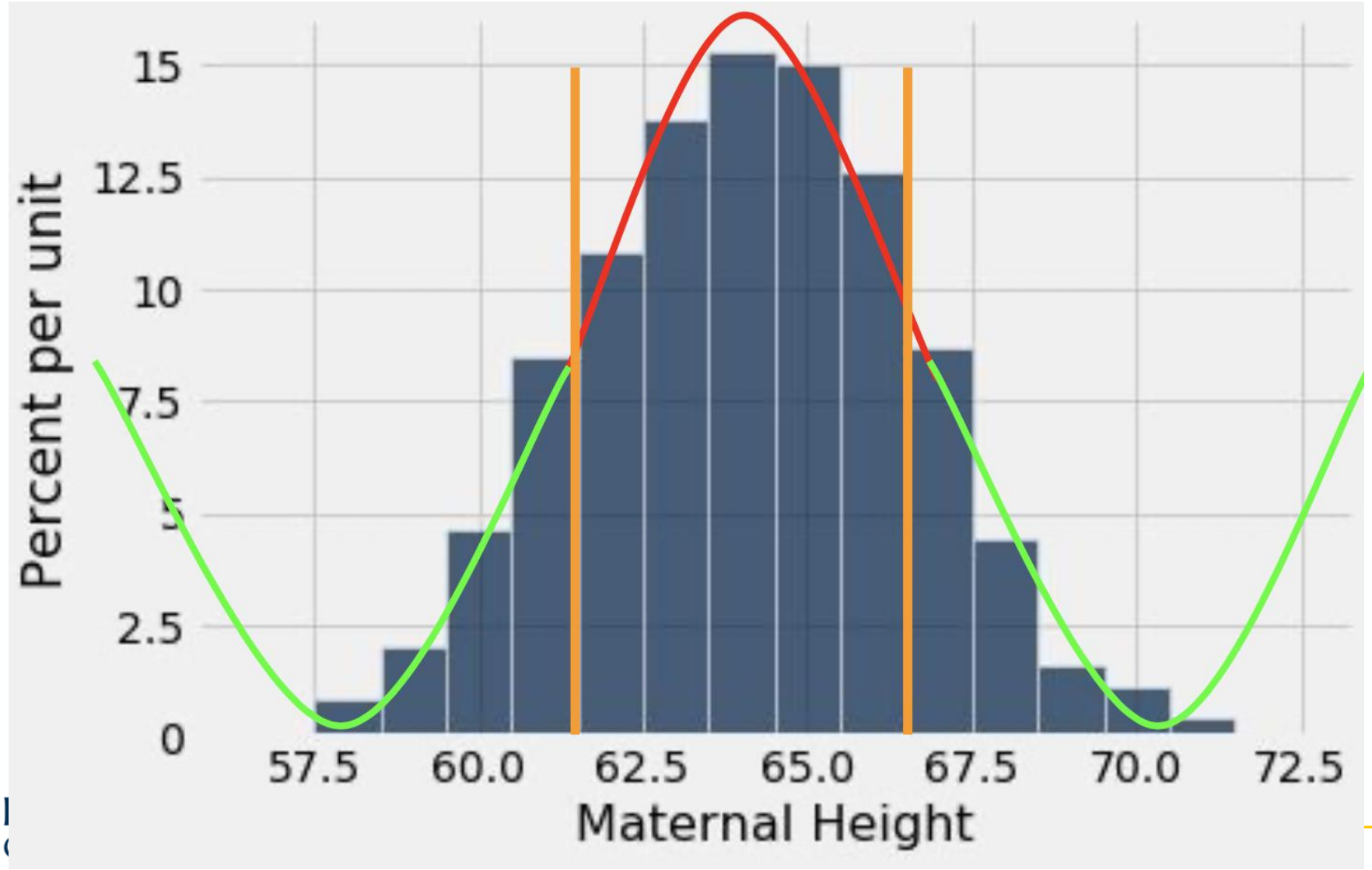
Alternative hypothesis: **Population average = $\neq$ *x***

Cutoff for P-value: *p*%

Method:

- Construct a (100-*p*)% confidence interval for the population average
- If *x* is not in the interval, reject the null
- If *x* is in the interval, can't reject the null

Points of Inflection



Which model is better

On a test set of 200 examples, model A scored 93% accuracy, model B scored 90% accuracy.

Null Hypothesis: This difference is due to change
Alternative: No, it isnt, A is better than B

Test: Bootstrap sample, compute difference in accuracies, repeat
If we rarely see a of difference of 3+, then reject null