Non-Parametric Methods; Simulations

March 6, 2020
Data Science CSCI 1951A
Brown University

Instructor: Ellie Pavlick
HTAs: Josh Levin, Diane Mutako, Sol Zitter

Announcements

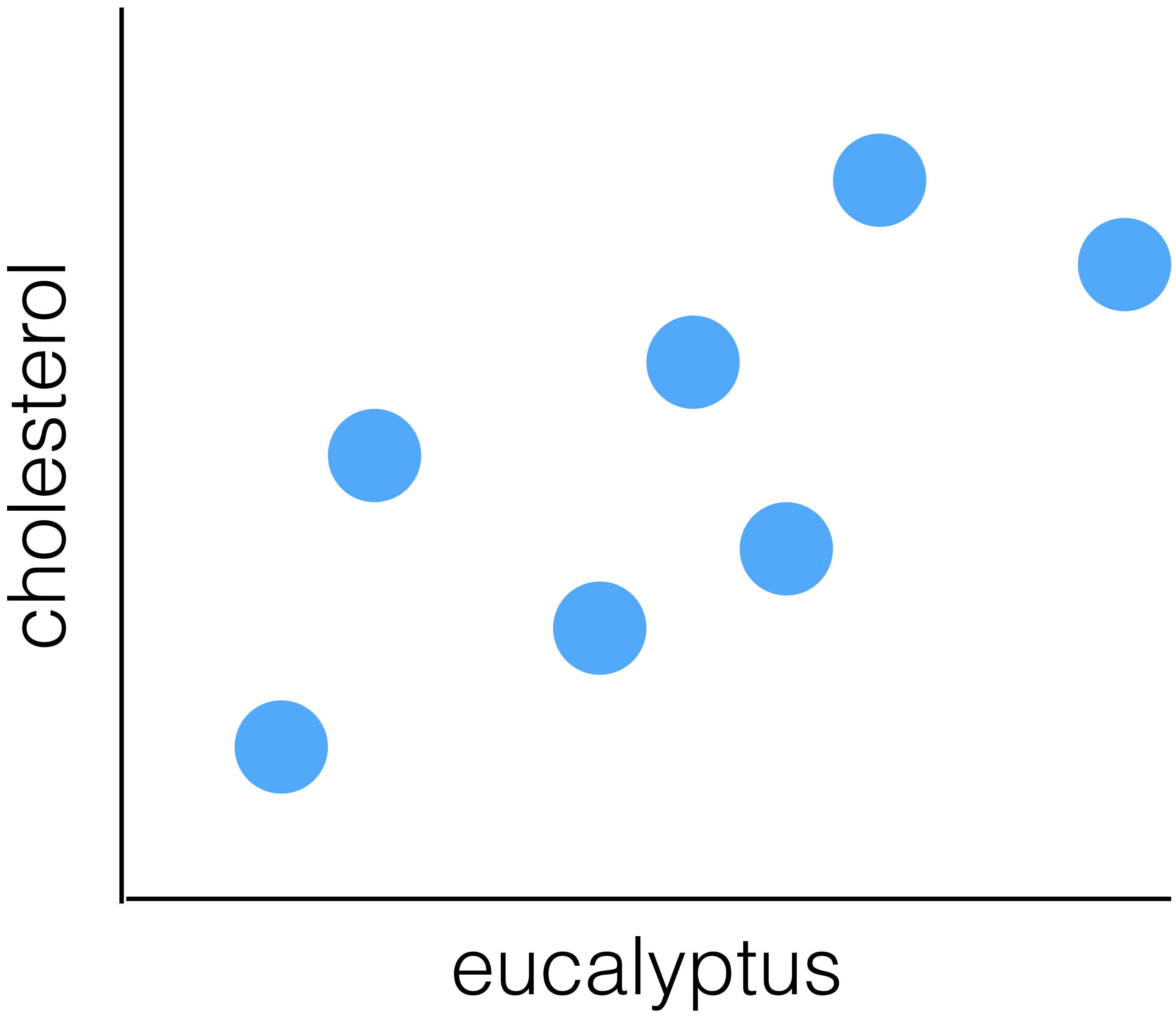
Today

- Non-Parametric Methods
- Simulations (example using Gaussian Mixture Models)

Today

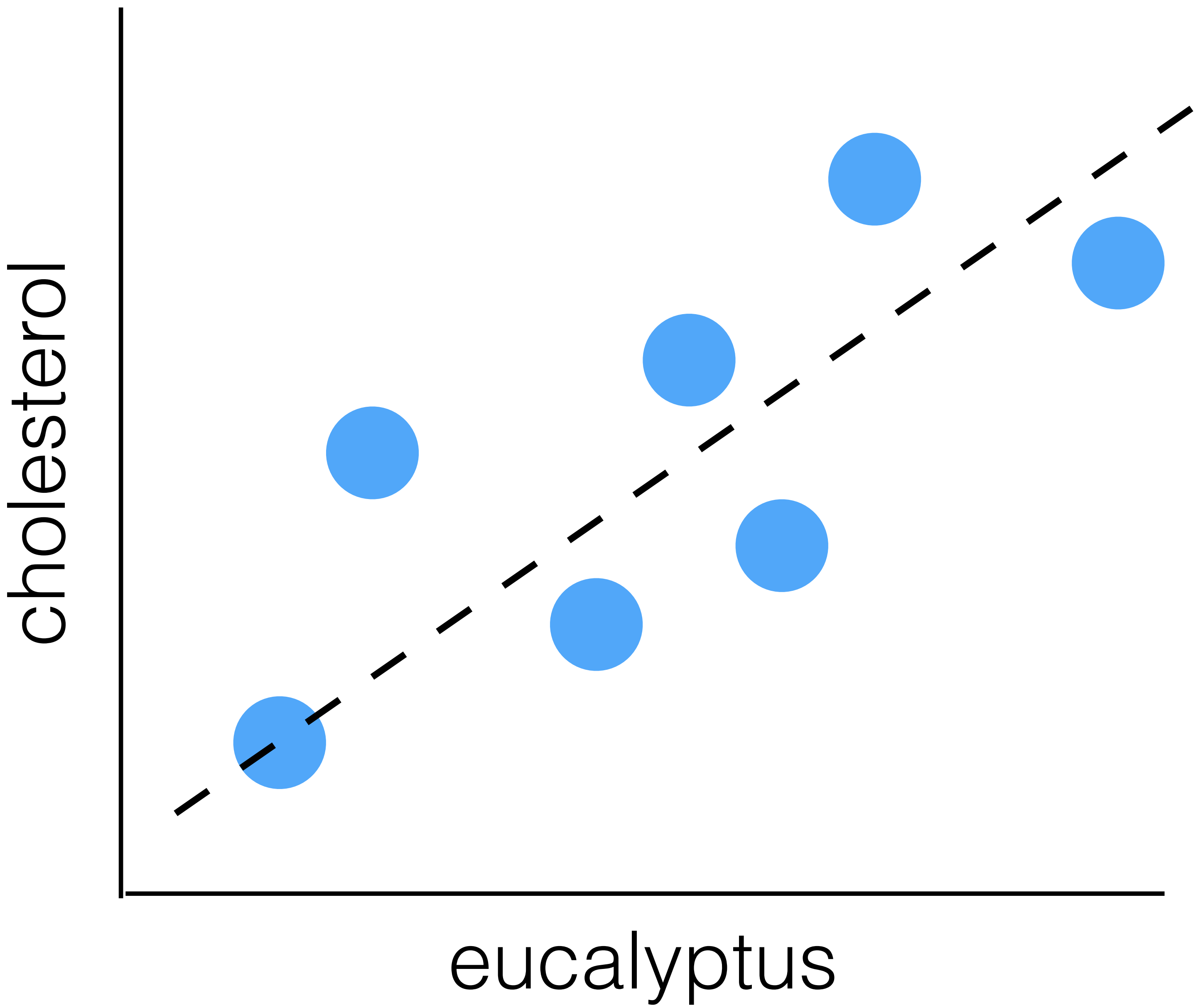
- **Non-Parametric Methods**
- Simulations (example using Gaussian Mixture Models)

Parametric vs. Non-Parametric



Given x, predict y

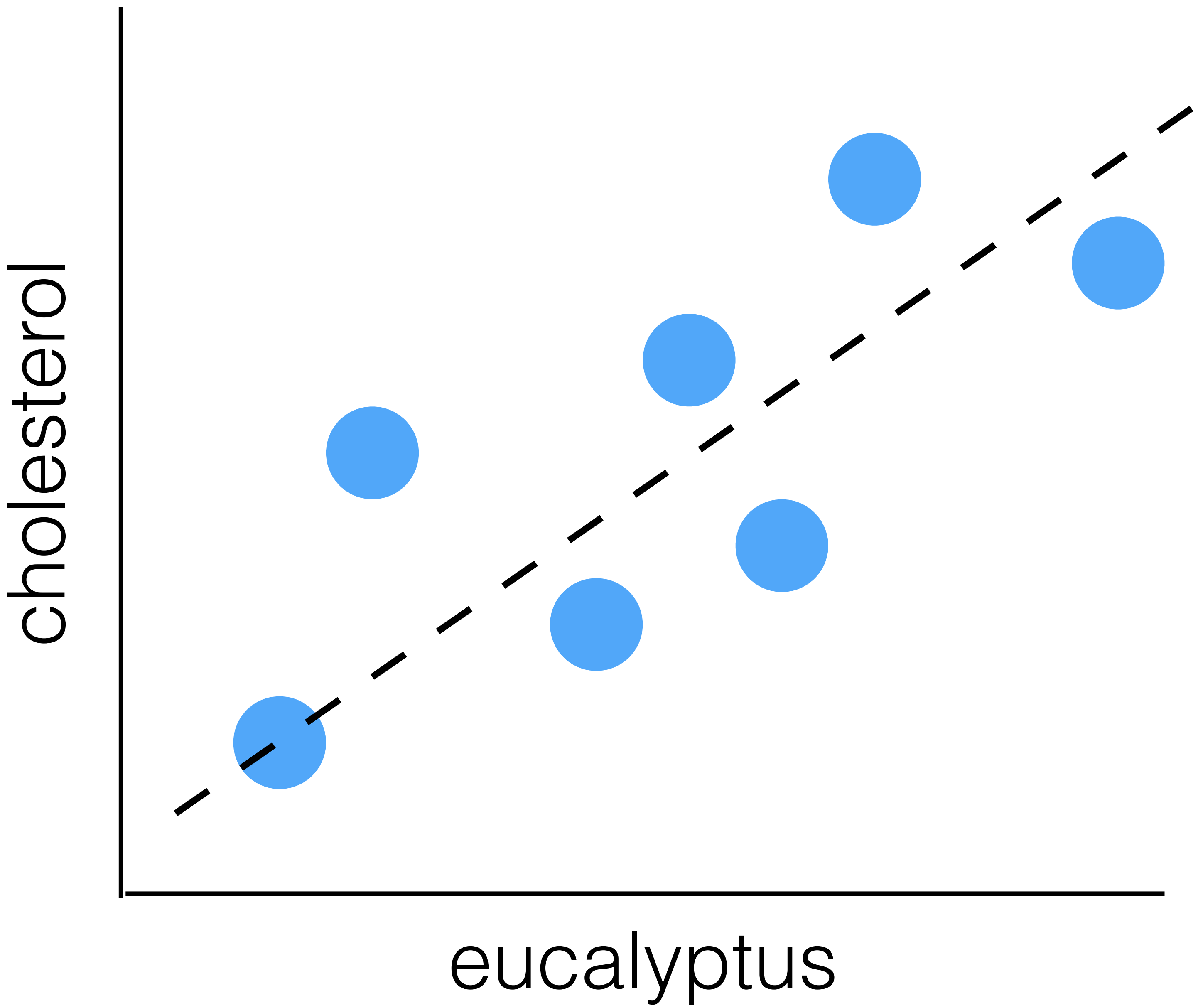
Parametric vs. Non-Parametric



Given x, predict y

$$y = mx + b + e$$

Parametric vs. Non-Parametric

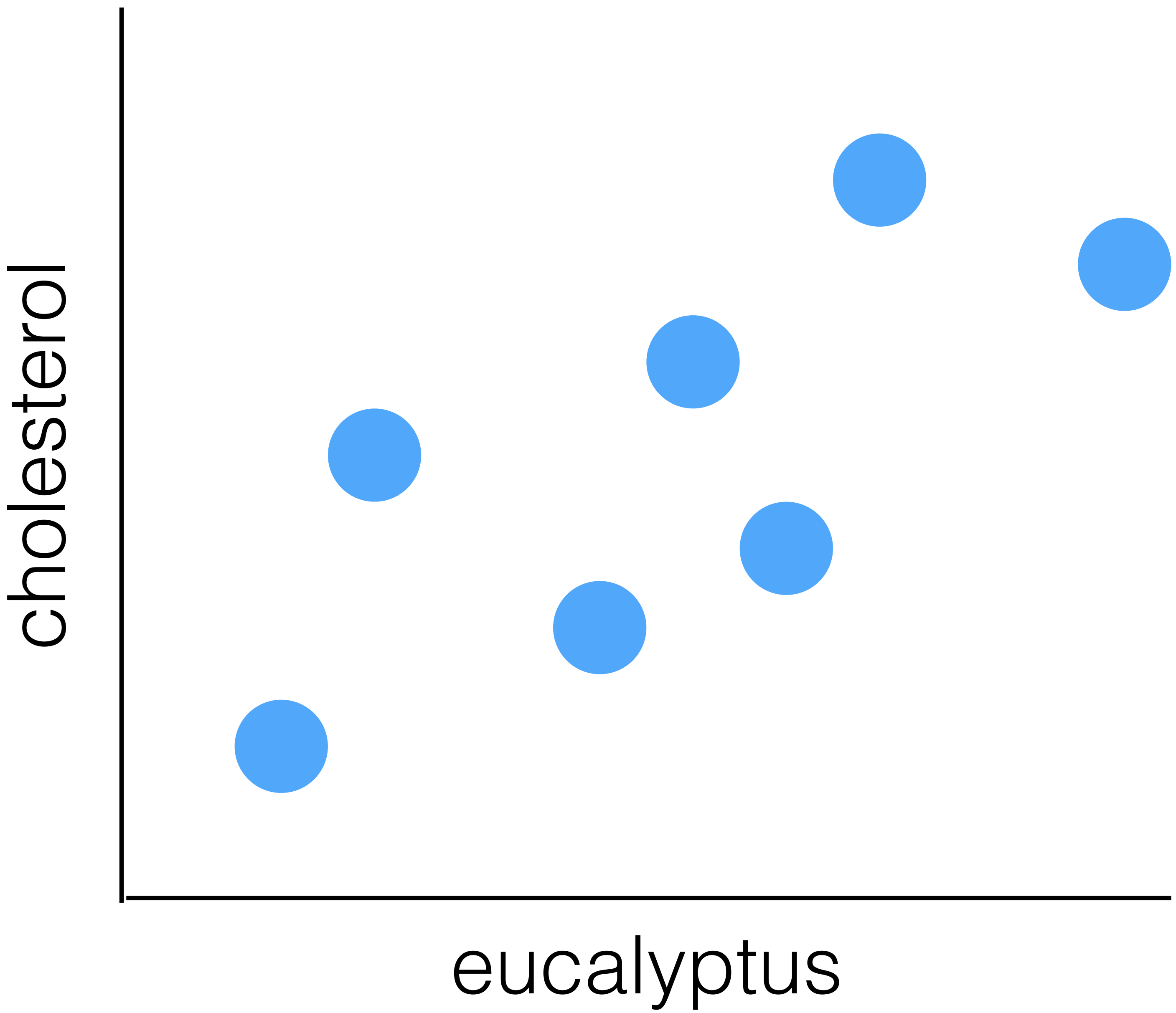


Given x, predict y

$$y = mx + b + e$$

Clicker Question!

Parametric vs. Non-Parametric

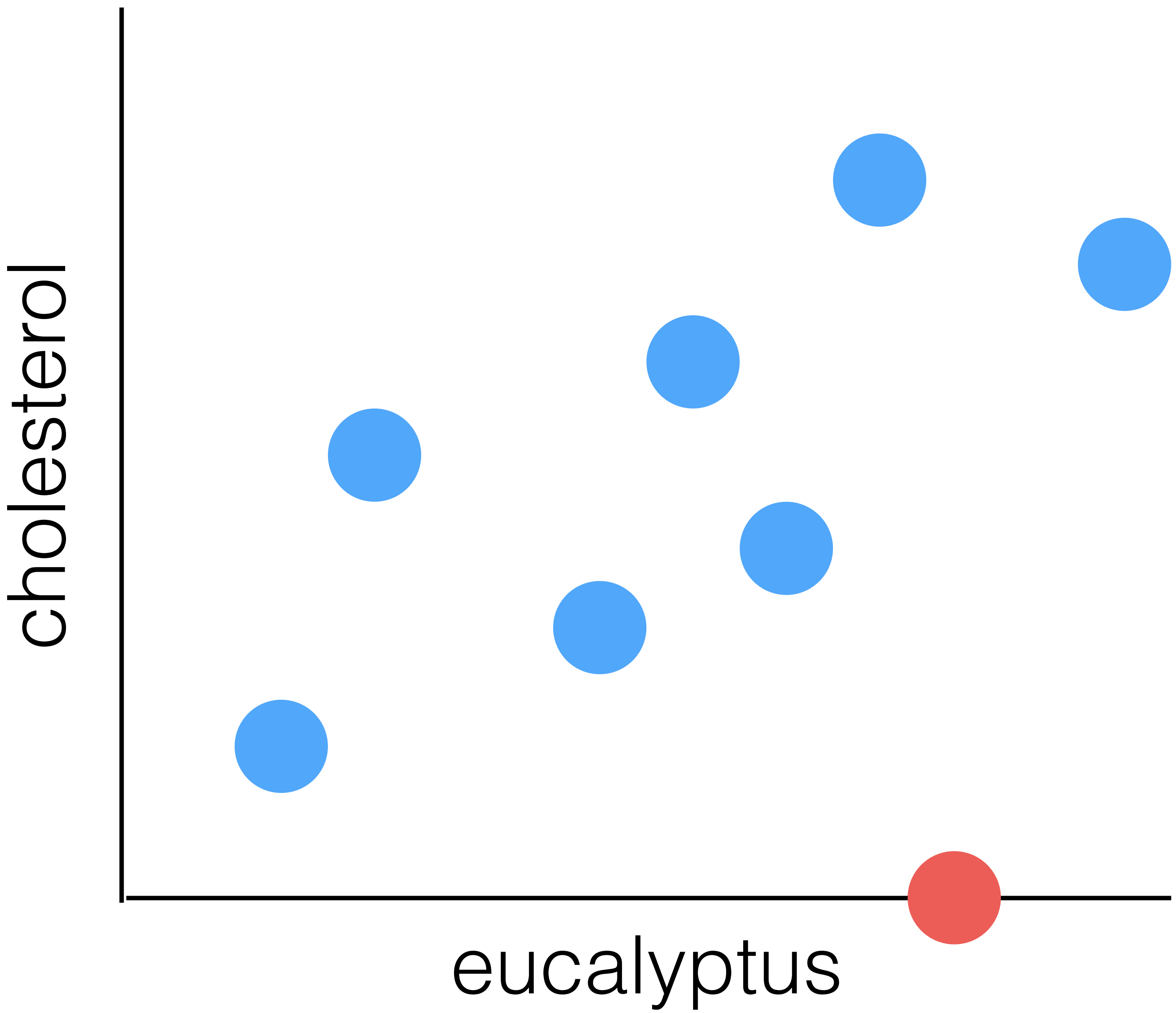


Given x, predict y

~~$y = mx + b + e$~~

Thoughts?

Parametric vs. Non-Parametric

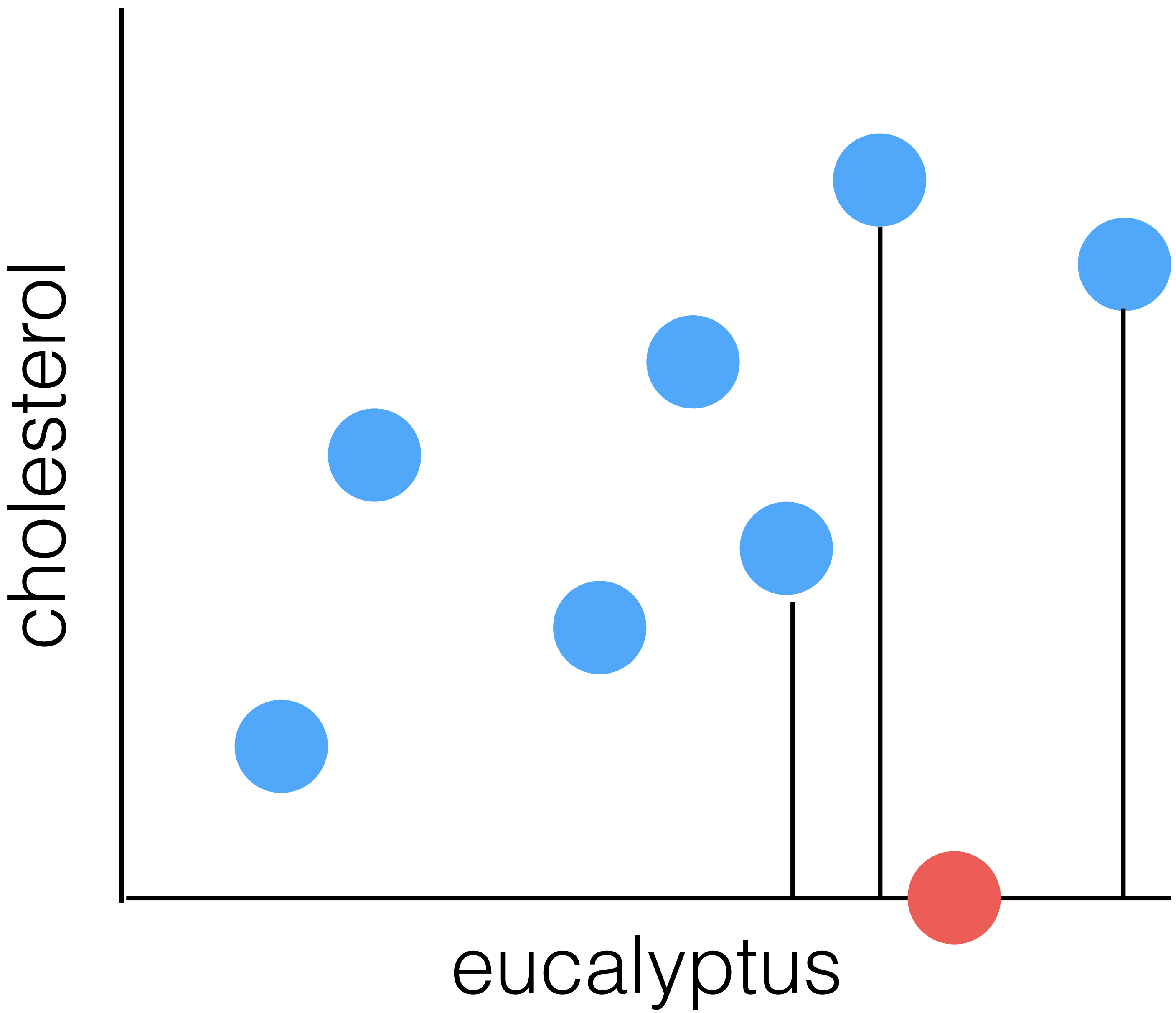


Given x, predict y

~~$y = mx + b + e$~~

Nearest Neighbors!

Parametric vs. Non-Parametric

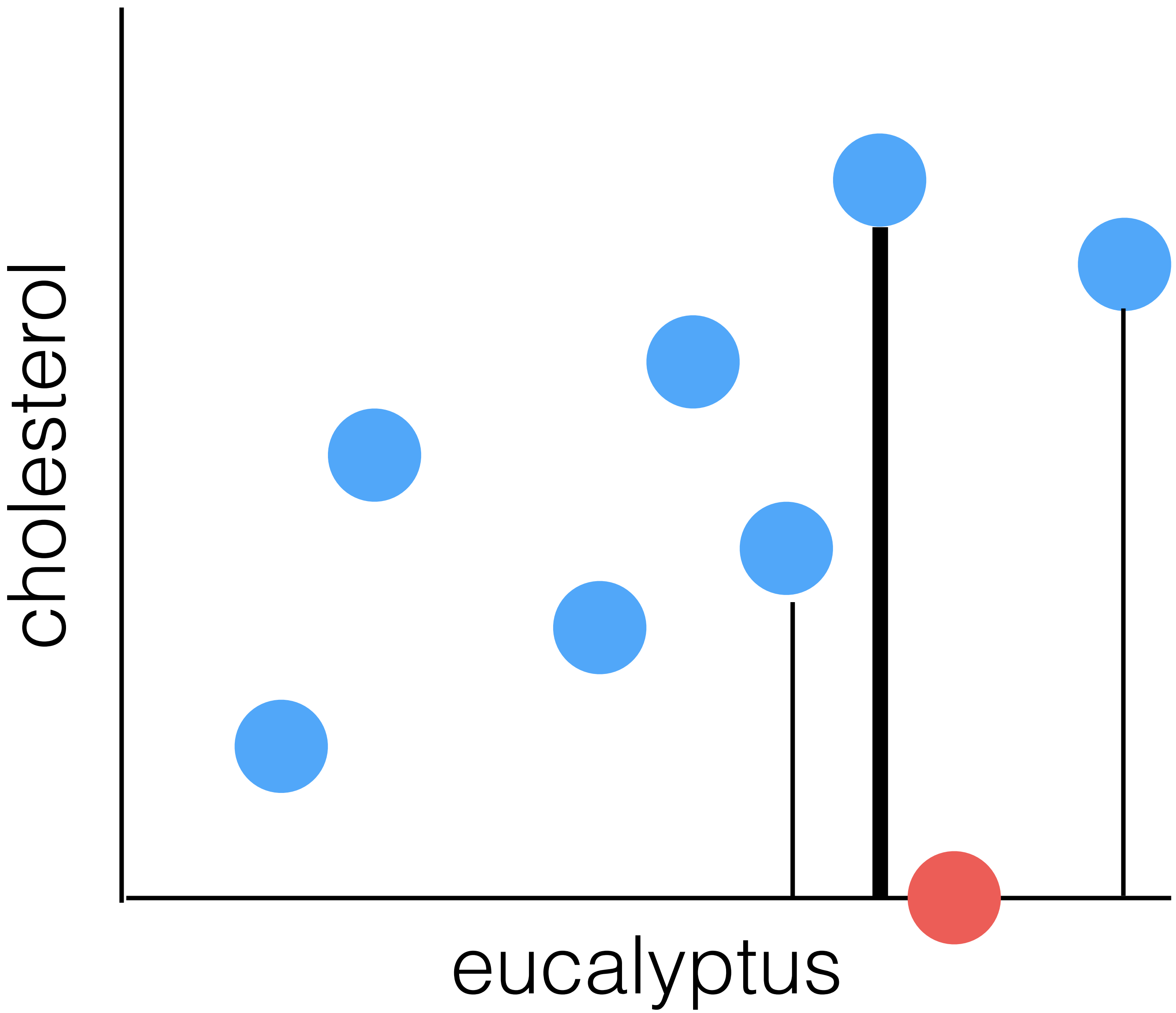


Given x, predict y

~~$y = mx + b + e$~~

Nearest Neighbors!

Parametric vs. Non-Parametric

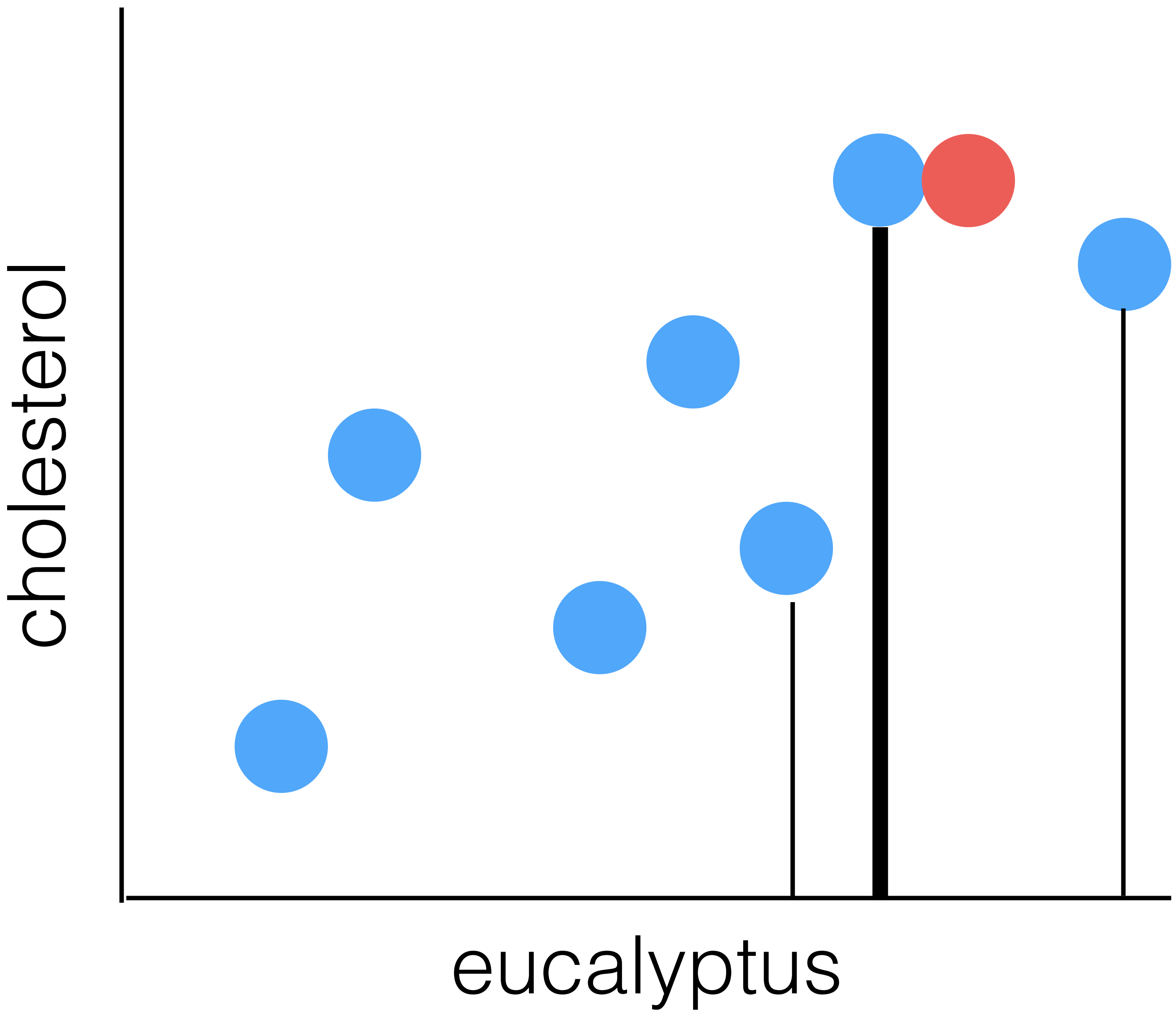


Given x, predict y

~~y = mx + b + e~~

Nearest Neighbors!

Parametric vs. Non-Parametric



Given x, predict y

~~$y = mx + b + e$~~

Nearest Neighbors!

Clicker Question!

Non-Parametric Models

Non-Parametric Models

- “Non-parametric” models: No assumptions about the number of parameters in the model or the particular form of the model

Non-Parametric Models

- “Non-parametric” models: No assumptions about the number of parameters in the model or the particular form of the model
- Pros:
  - Can work well with small data
  - Or when you have very complex distributions and you aren’t sure what assumptions can be made

Non-Parametric Models

- “Non-parametric” models: No assumptions about the number of parameters in the model or the particular form of the model
- Pros:
  - Can work well with small data
  - Or when you have very complex distributions and you aren’t sure what assumptions can be made
- Cons:
  - Size of model can increase with size of data
  - Slow to compute (randomized/iterative processes)
  - Fewer assumptions -> weaker conclusions (higher p-values)

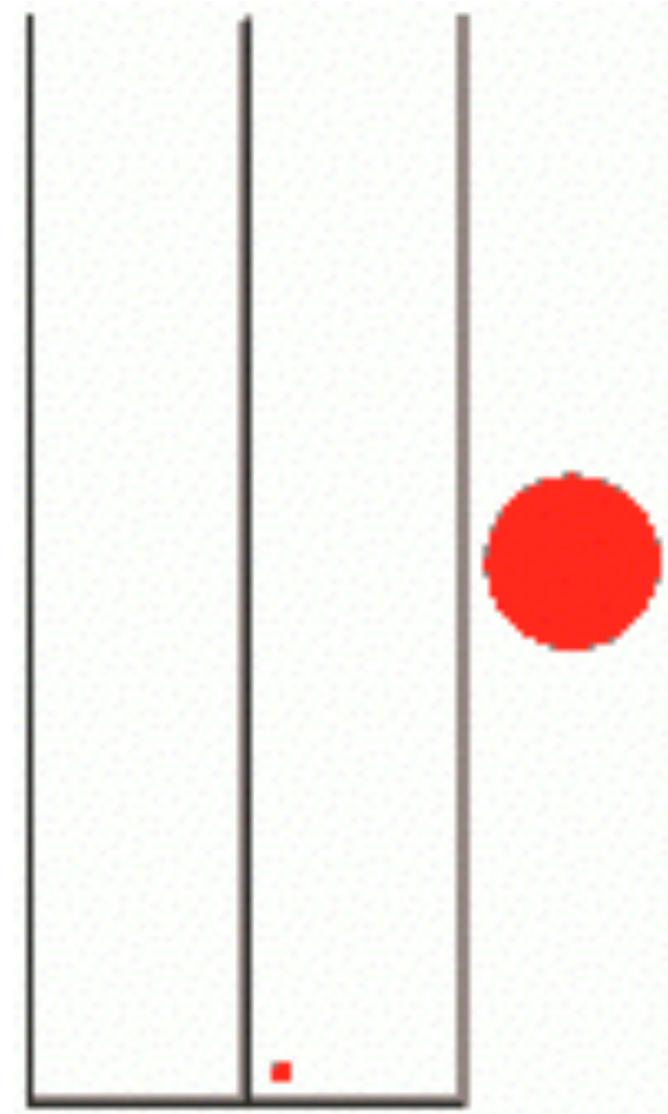
Law of Large Numbers

- If you perform the same experiment a large number times, the *average* will converge to the expected value

- Assumes that errors are "random" and uncorrelated, so will balance out over time

$$\bar{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$$

$$\bar{X}_n \to \mu \text{ as } n \to \infty$$

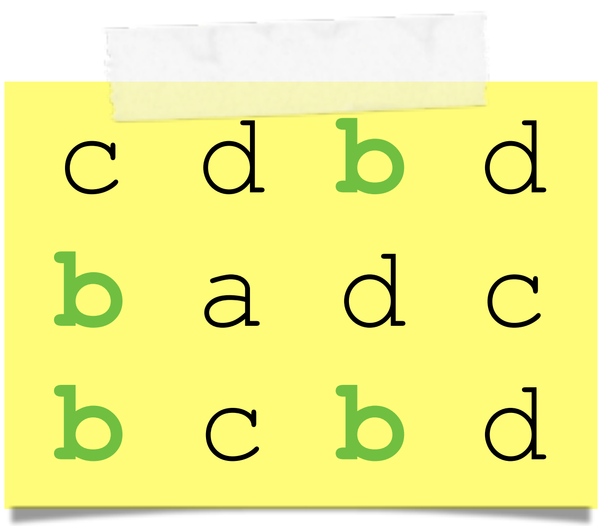


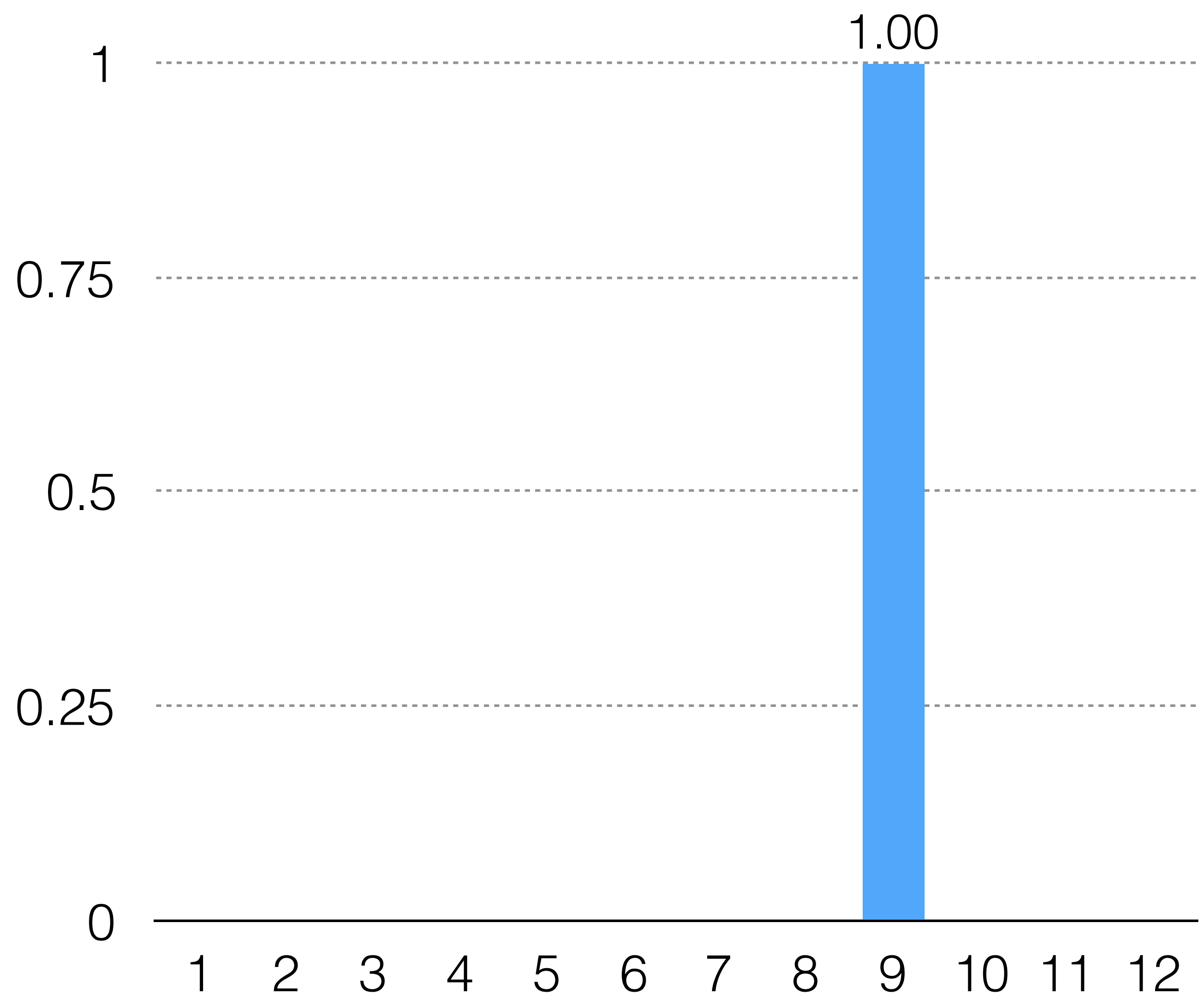
https://en.wikipedia.org/wiki/Law_of_large_numbers

Central Limit Theorem

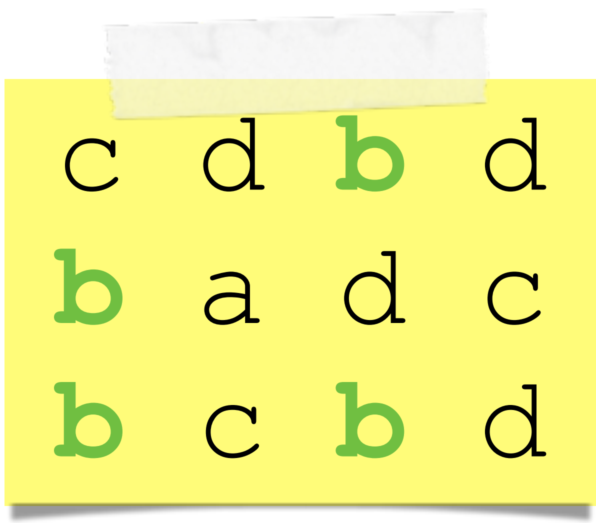
- Given $X_1 \ldots X_n$
- Not only does a $\bar{X}_n \to \mu \text{ as } n \to \infty$
- But the distribution approaches a normal distribution

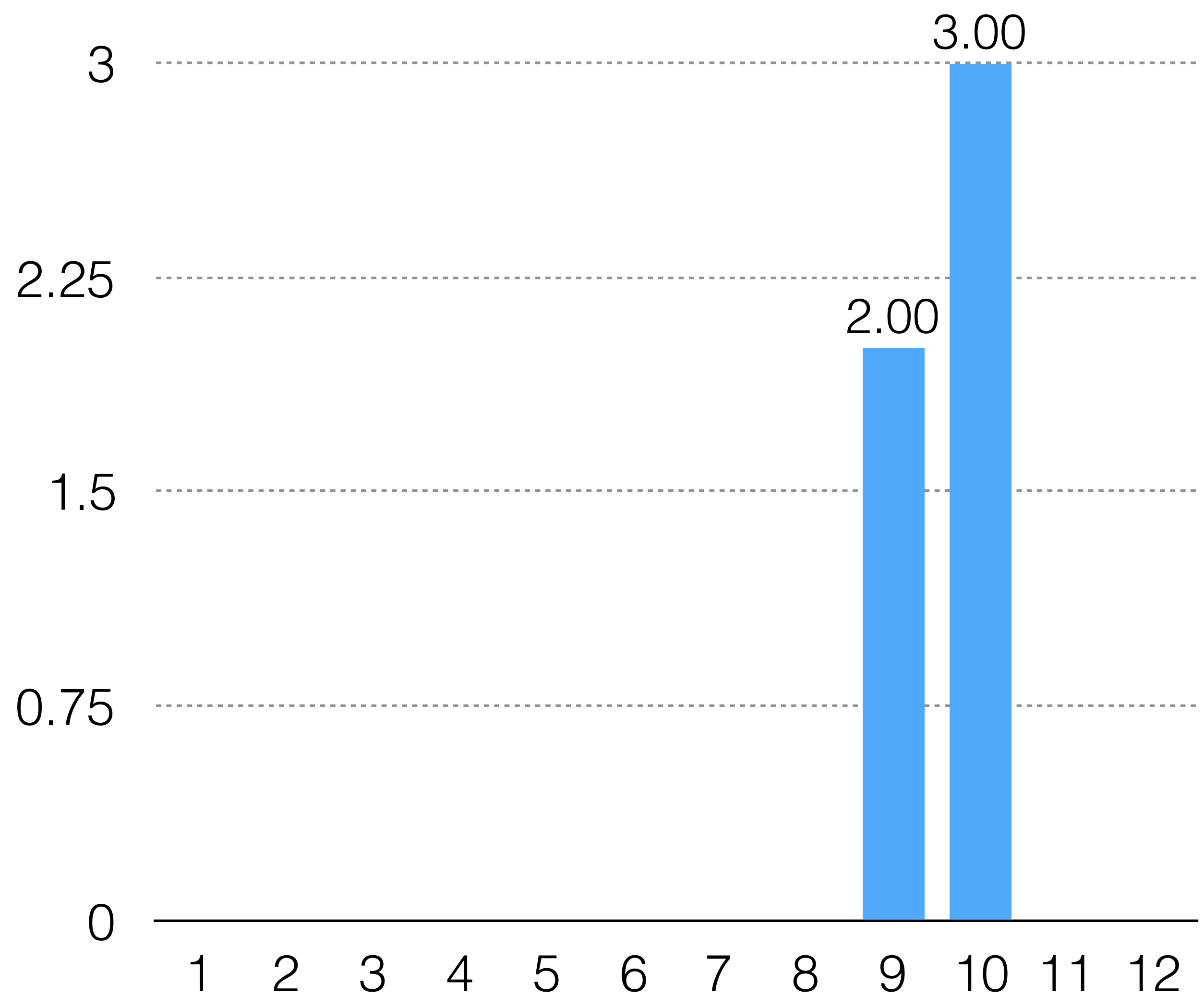
Central Limit Theorem



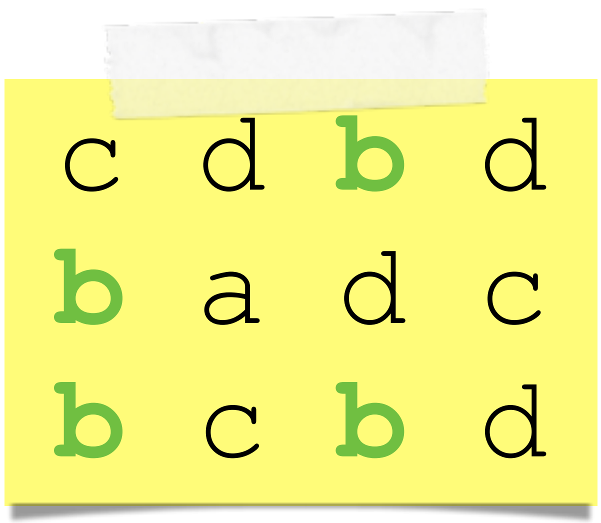


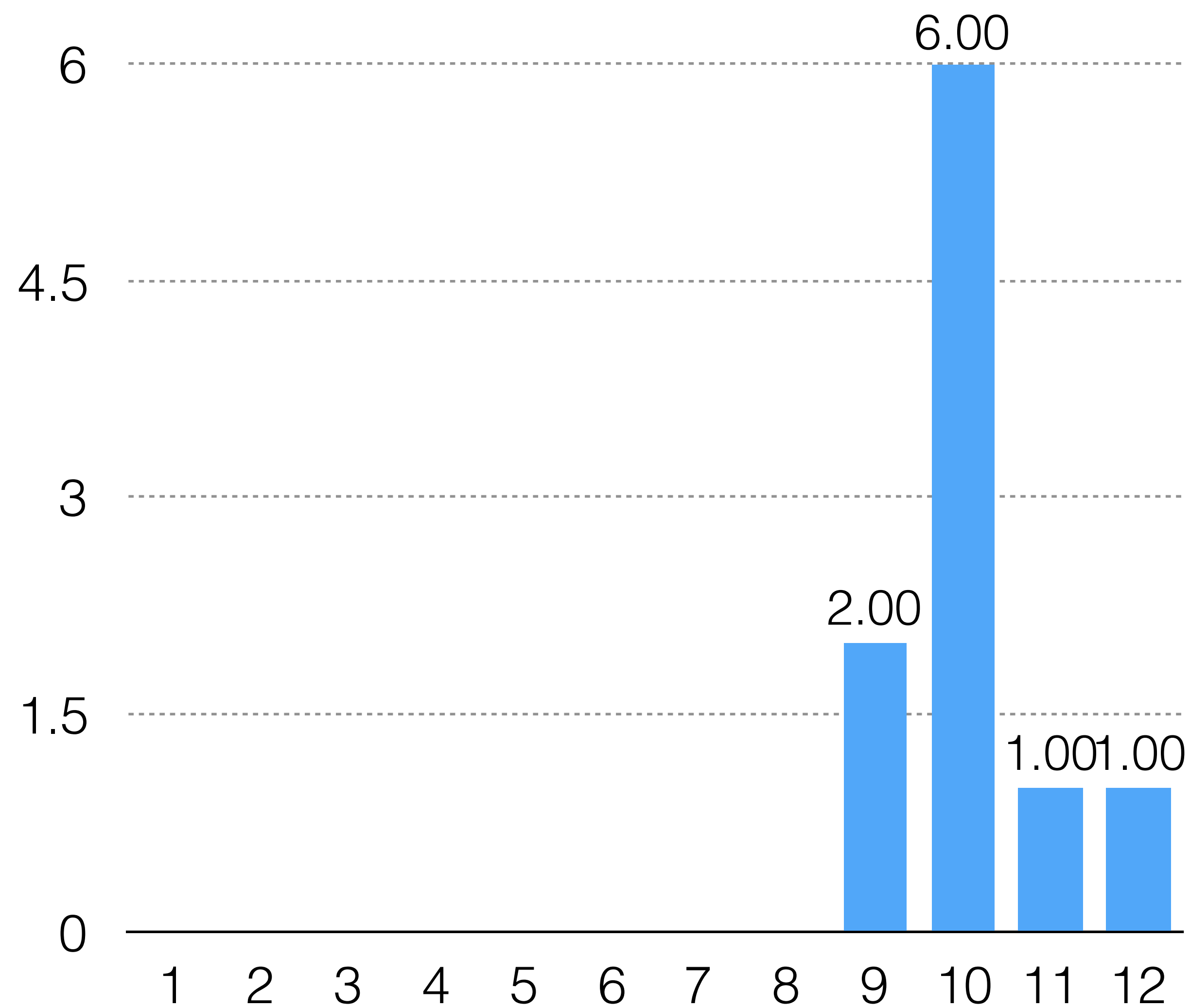
Central Limit Theorem



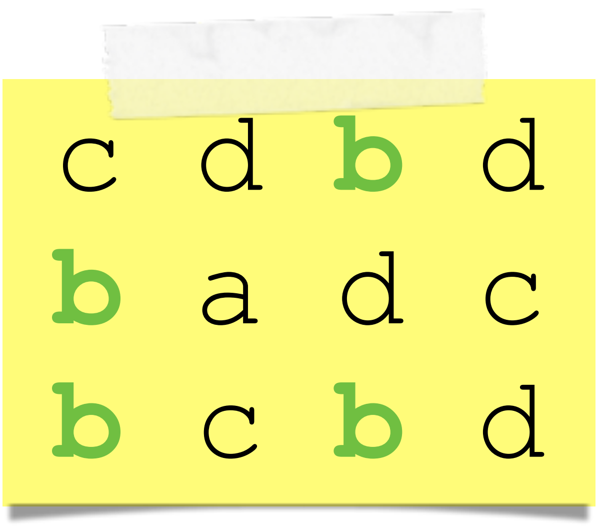


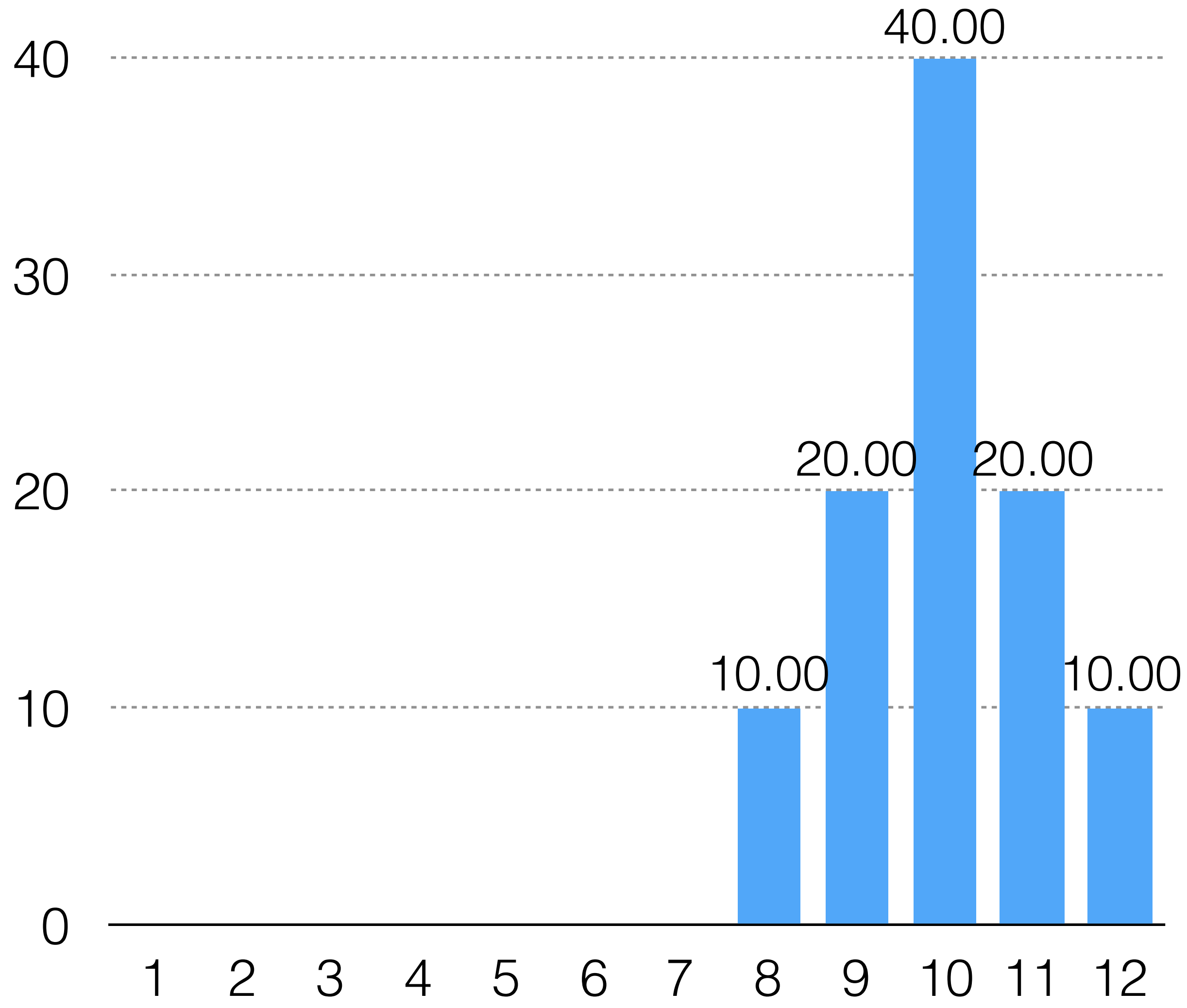
Central Limit Theorem



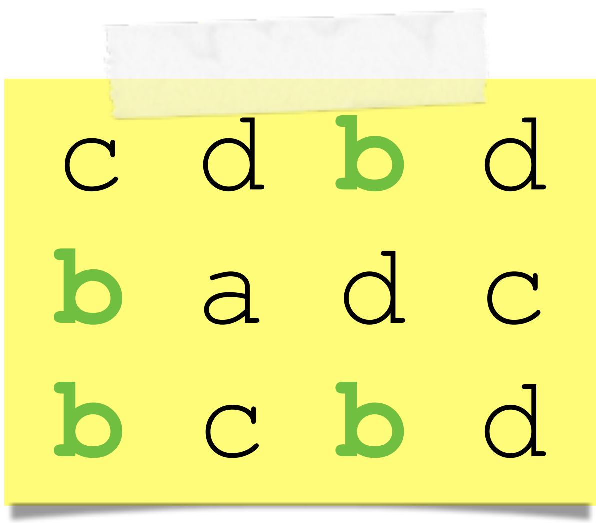


Central Limit Theorem

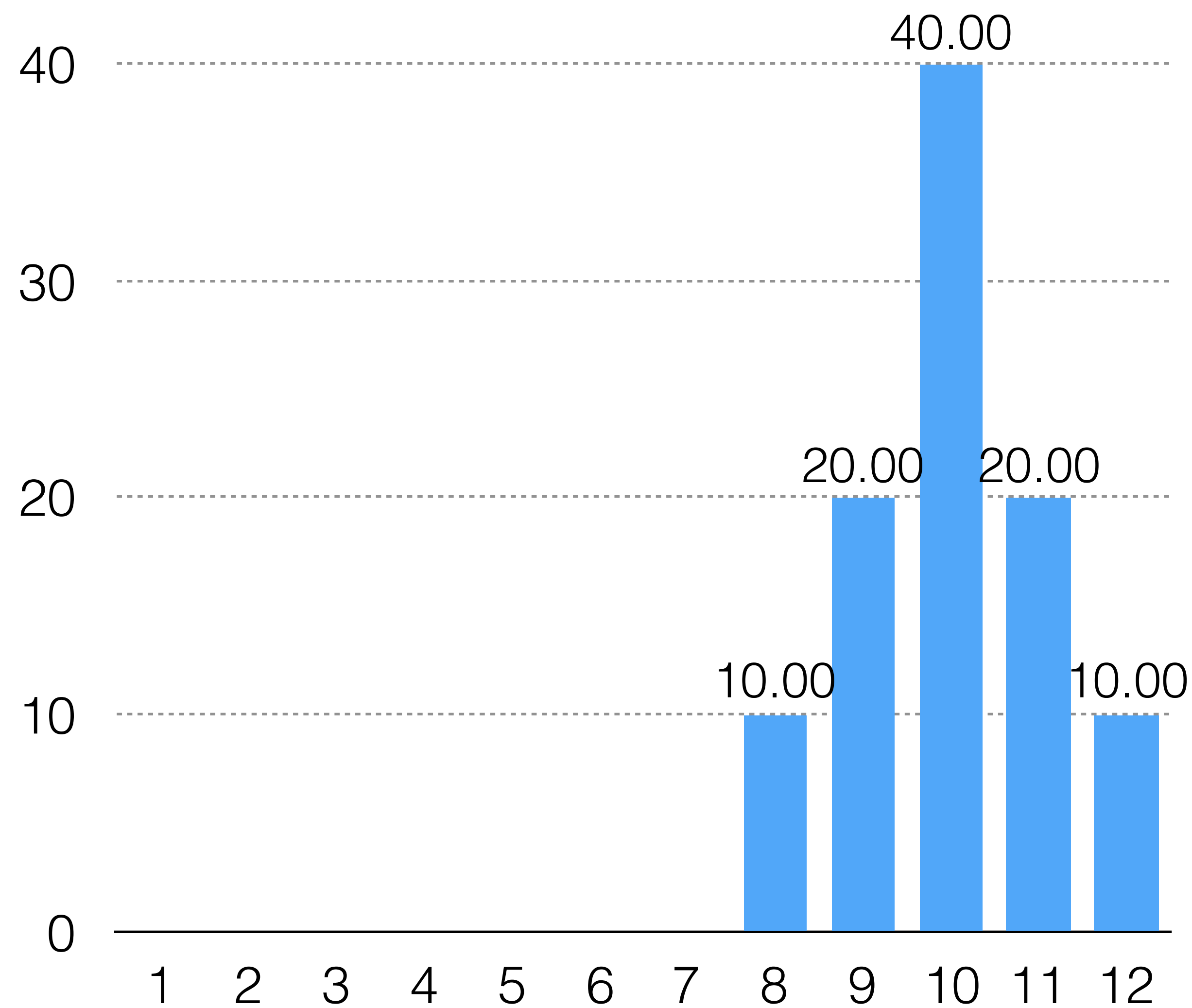




Central Limit Theorem



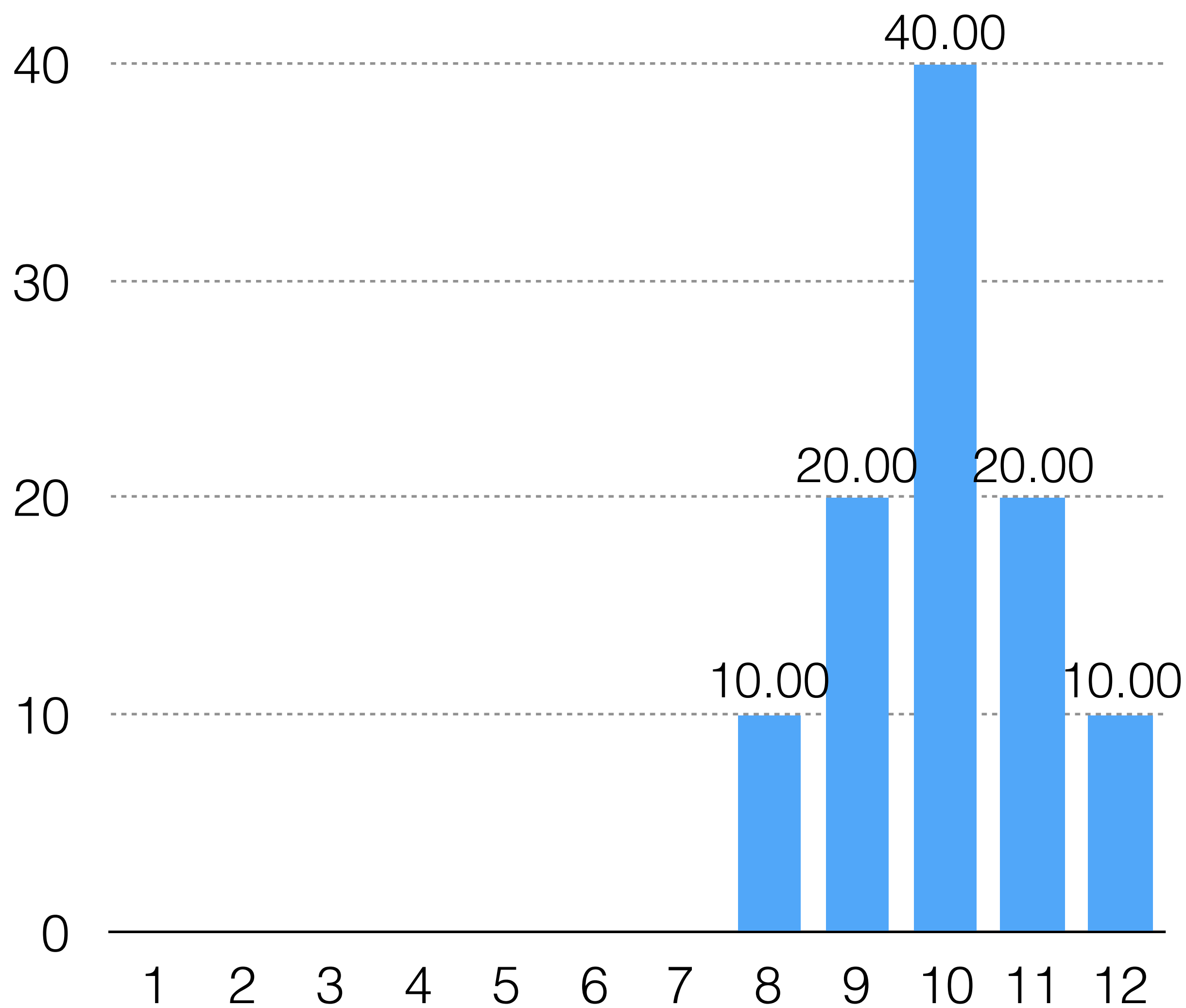
I.e. test statistics are often normally distributed...



Central Limit Theorem

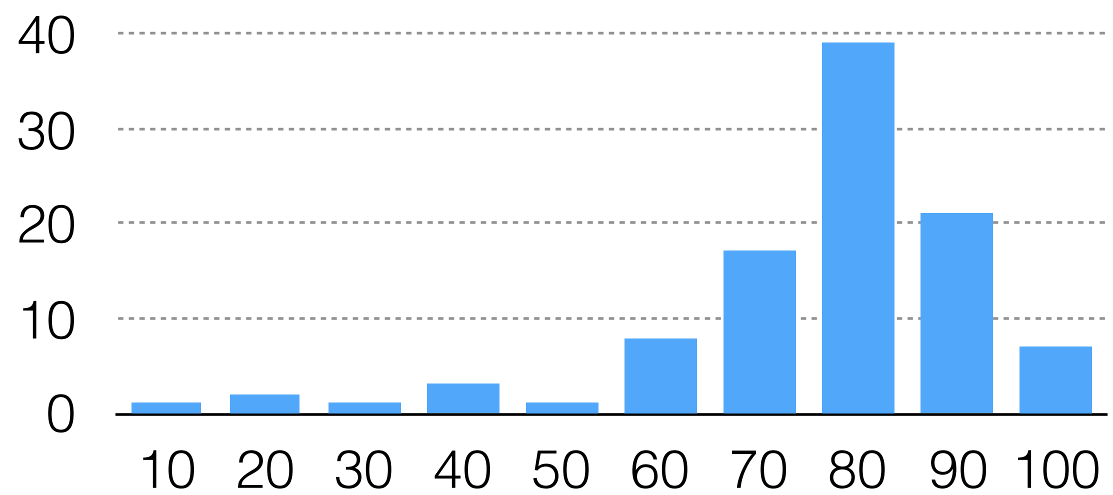
c d **b** d
b a d c
b c **b** d

Can apply statistical methods designed for normal distributions even when underlying distribution is not normal

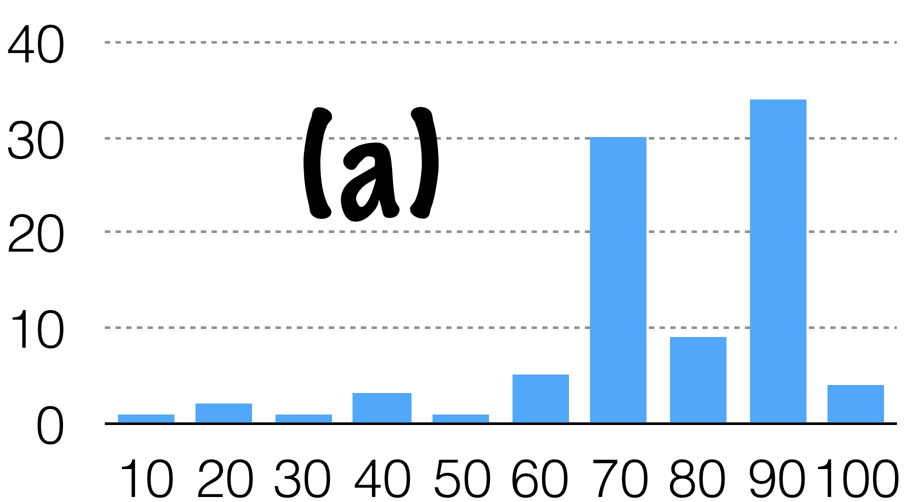


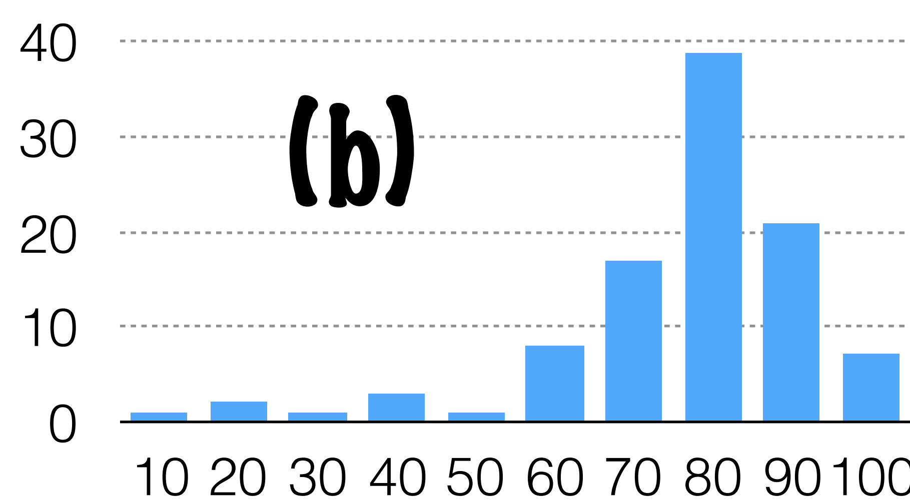
Central Limit Theorem

Every year, I compute the mean grade in my class. I never change the material or my methods for evaluating because, lazy. Over the 439 years that I have been teaching this class, this has resulted in the below distribution.



Which of these is mostly like the typical distribution on any given year?





(c) can't say, could be either

Central Limit Theorem

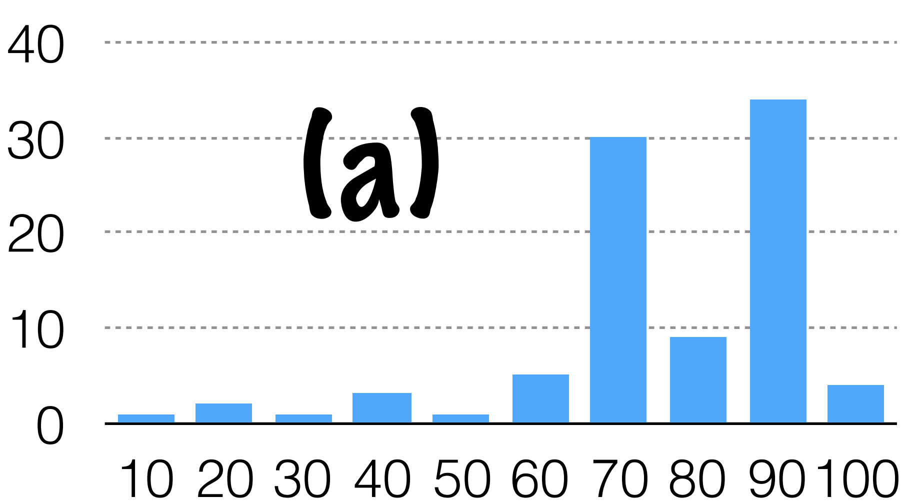
Every year, I compute the mean grade in my class. I never change the

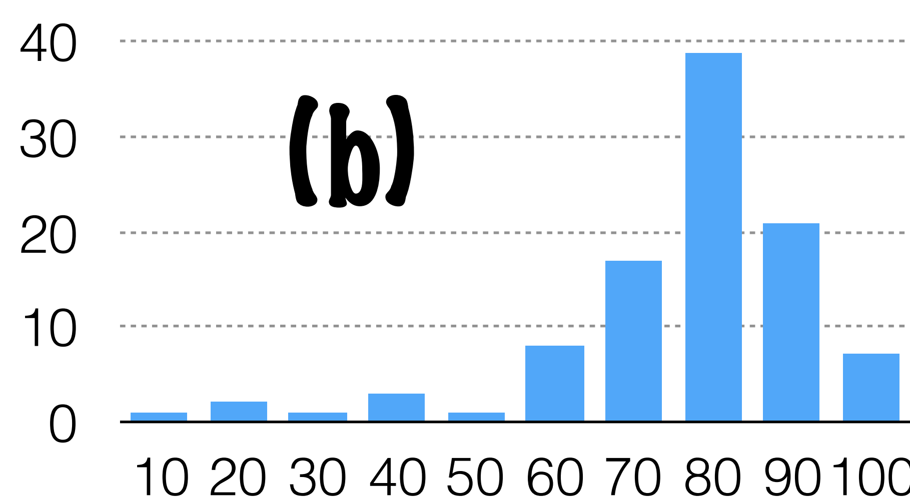
mater … that I

hav … on.

Central Limit Theorem: repeated measures of mean will be normally distributed, doesnt assume the population over which you are taking the mean is normally distributed.

Which of these is mostly like the typical distribution on any given year?

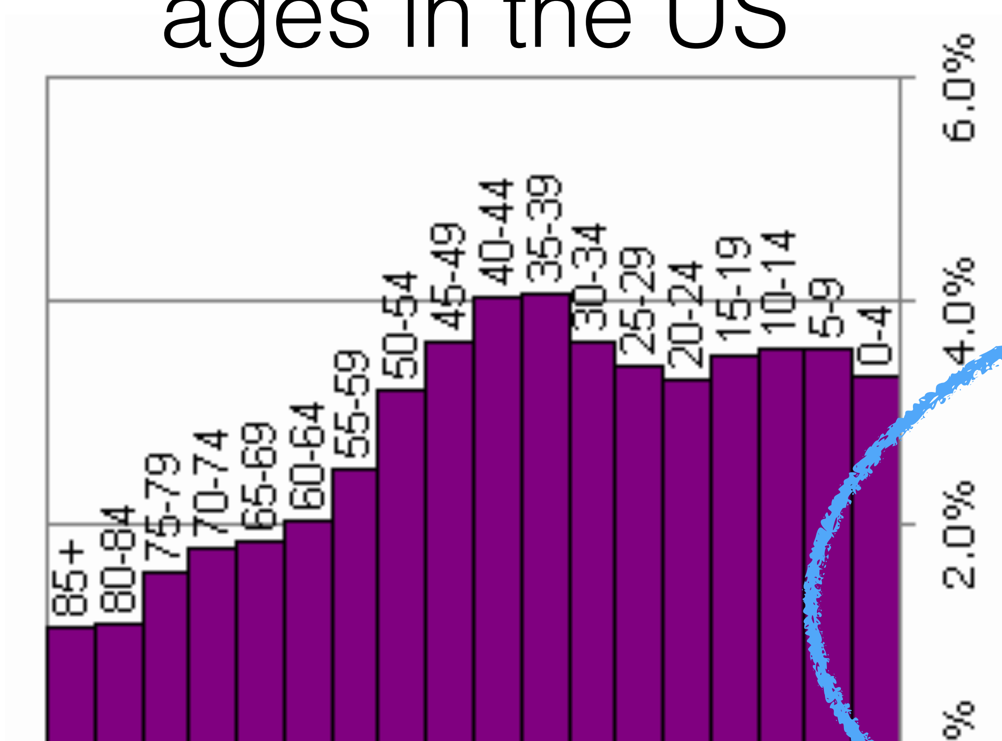




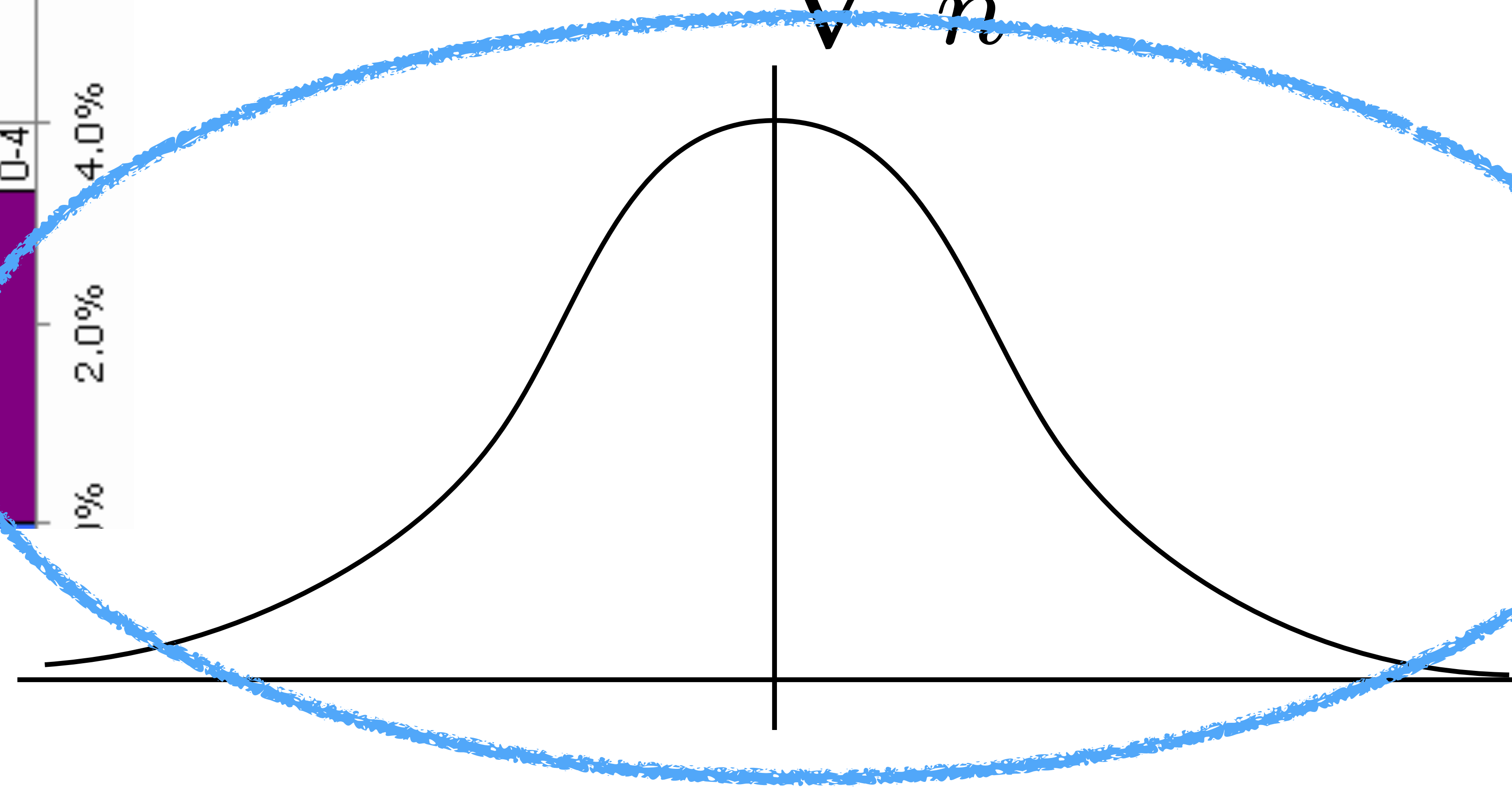
(c) can't say, could be either

Test for population means

Distribution of ages in the US



$$t = \frac{\bar{x} - \mu_0}{\sqrt{\frac{s}{n}}}$$

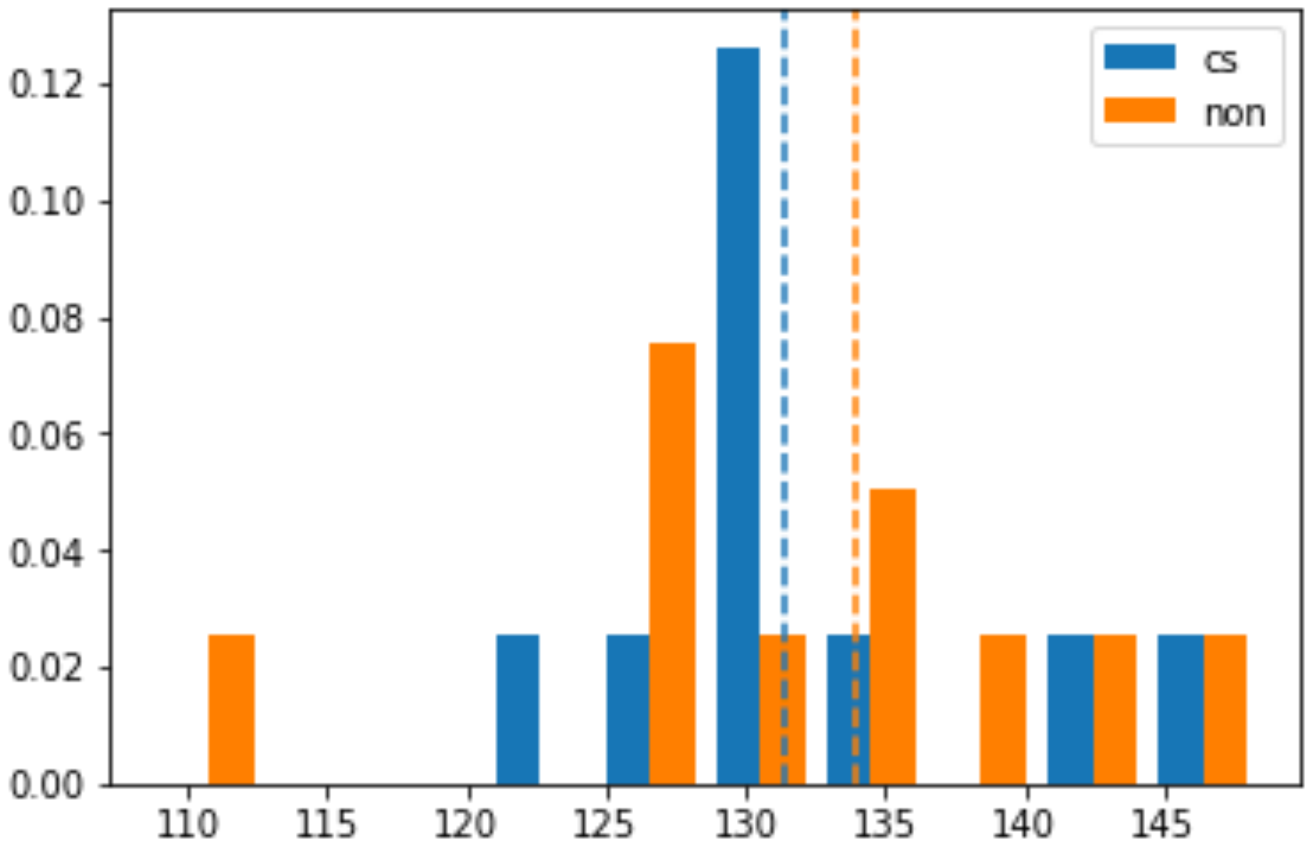


Hypothesis:
Mean age is 35.

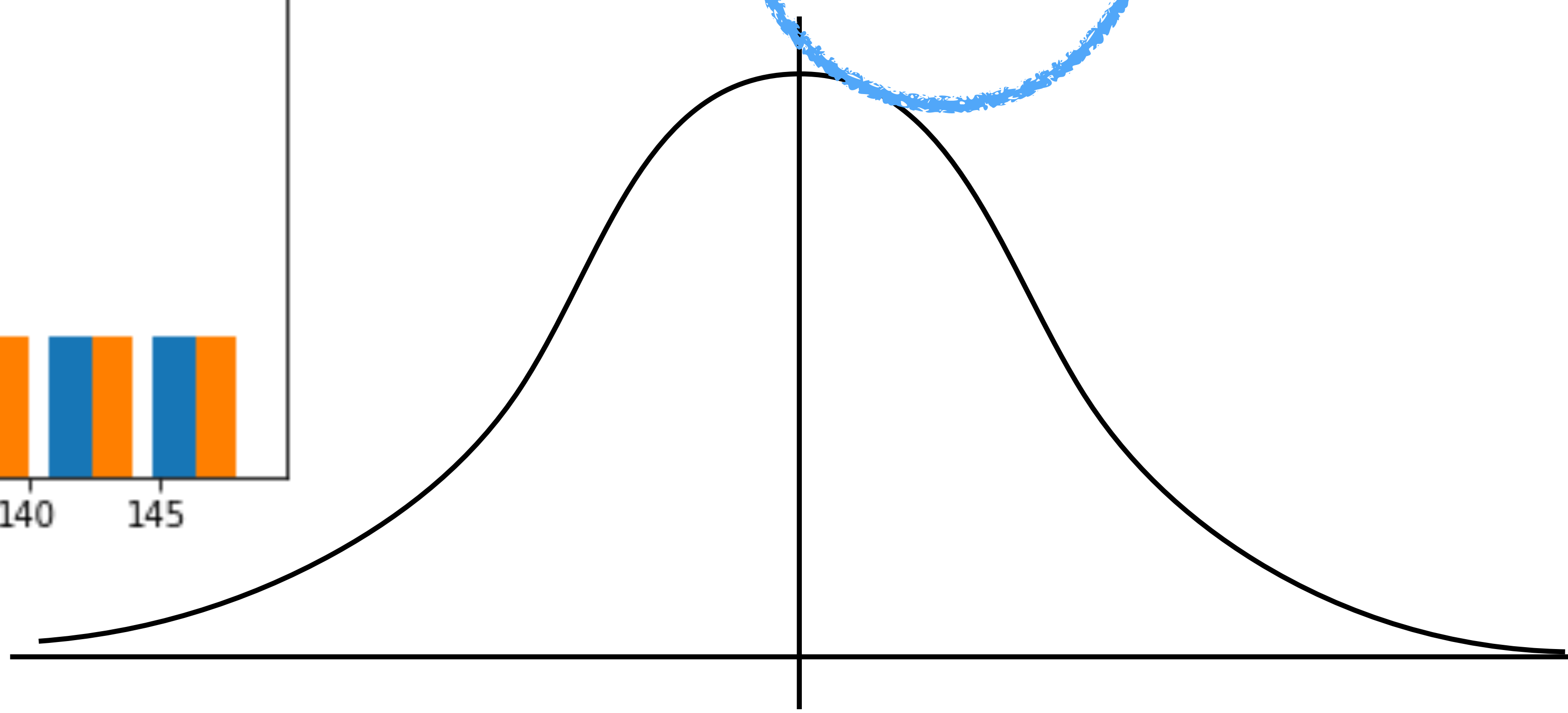
http://www.censusscope.org/us/chart_age.html

Clicker Question!

Test for population medians?



$$t = \frac{\bar{x} - \mu_0}{\sqrt{\frac{s}{n}}}$$



Non-Parametric Hypothesis Testing

- We still want to determine the probability of the test statistic under the null hypothesis…
- …but we don't have an analytic solution, maybe because
  - Theoretical distribution is unknown, complex, or hard to write down
  - Assumptions about analytic solution are suspect (e.g. sample size not large enough)

Non-Parametric Hypothesis Testing

- We still want to determine the probability of the test statistic under the null hypothesis…
- …but we don't have an analytic solution, maybe because
  - **Theoretical distribution is unknown, complex, or hard to write down**
  - Assumptions about analytic solution are suspect (e.g. sample size not large enough)

Non-Parametric Hypothesis Testing

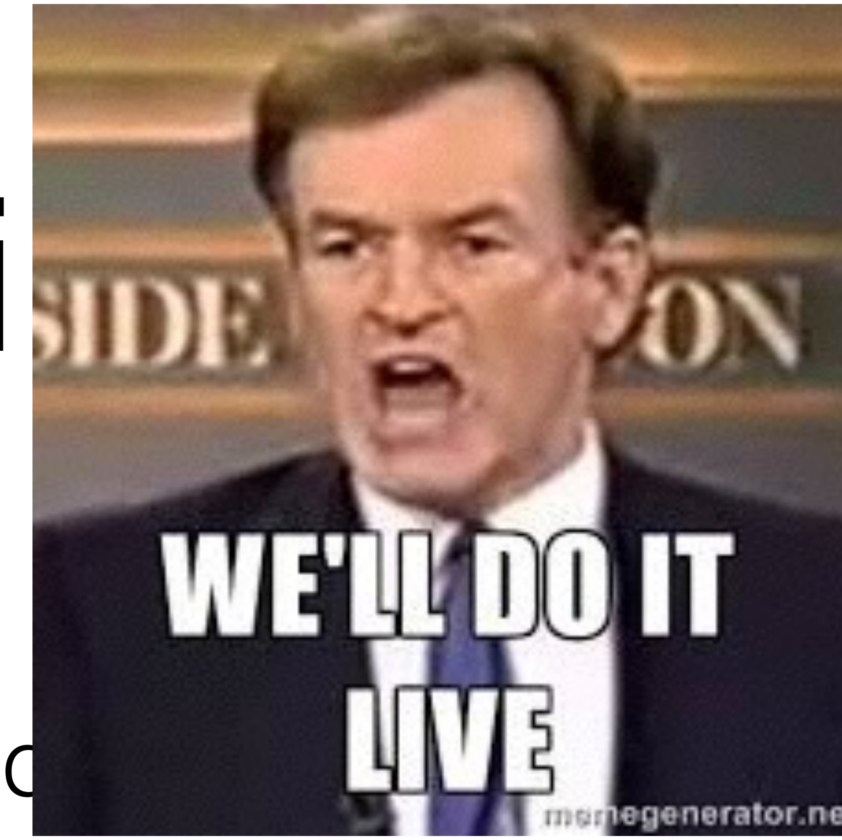
median, 90th percentile, annotator agreement, model accuracy, whatever cool metric you made up that you care about.

- **Theoretical distribution is unknown, complex, or hard to write down**
- Assumptions about analytic solution are suspect (e.g. sample size not large enough)

Bootstrapping

- Resample (with replacement) in order to approximate the distribution of the test statistic
- Compute the test statistic over each sample
- Repeat some large number of times (say 10,000)
- View distribution of computed test statistics

Bootstrappi



- Resample (with replacement) in ord approximate the distribution of the test statistic
- Compute the test statistic over each sample
- Repeat some large number of times (say 10,000)
- View distribution of computed test statistics

Permutation Test

H_a: CS students sleep less than the rest of Brown students

Permutation Test 

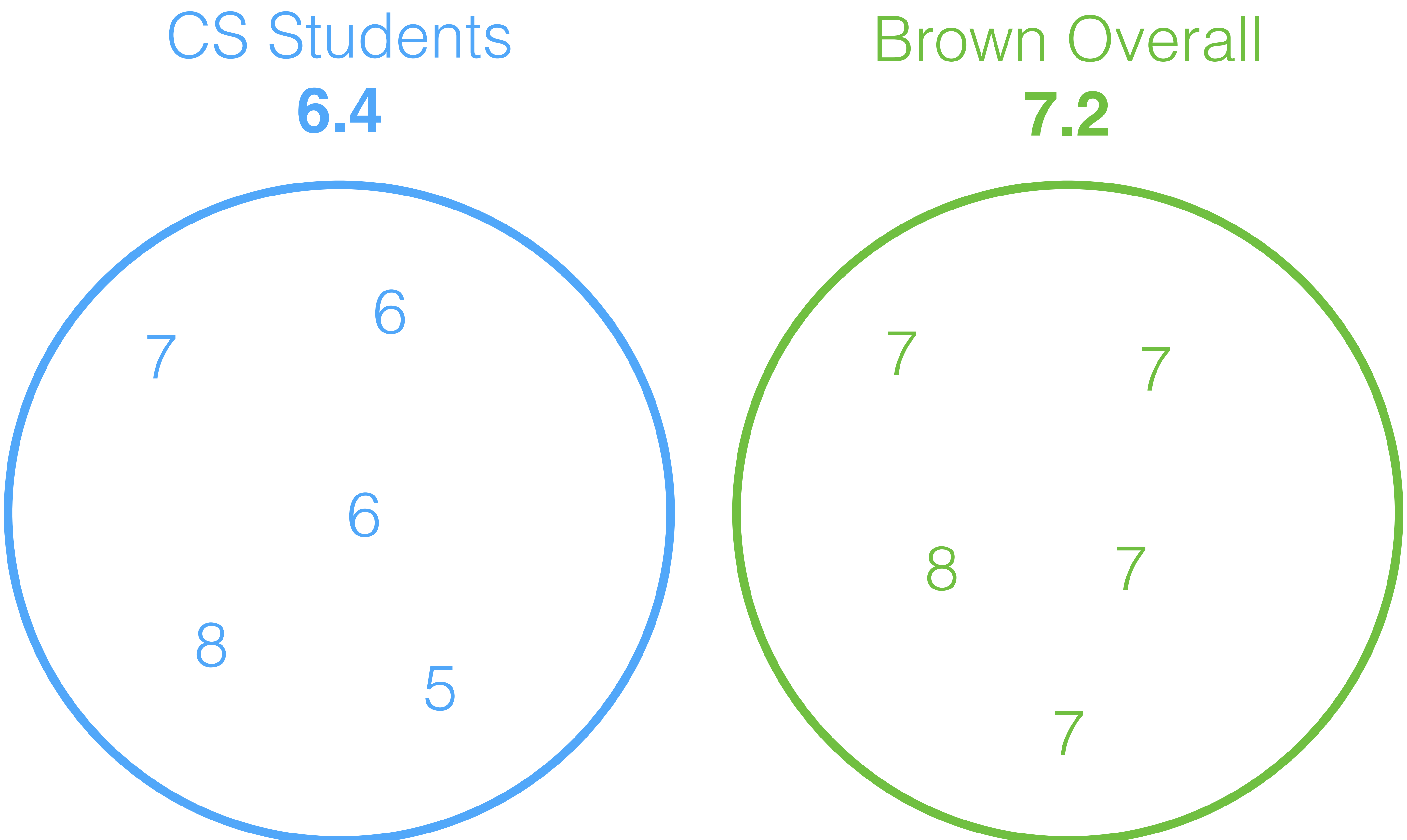
H_a: CS students sleep less than the rest of Brown students

Permutation Test

H_0: CS students sleep the same amount as everyone else
H_a: CS students sleep less than the rest of Brown students

Permutation Test

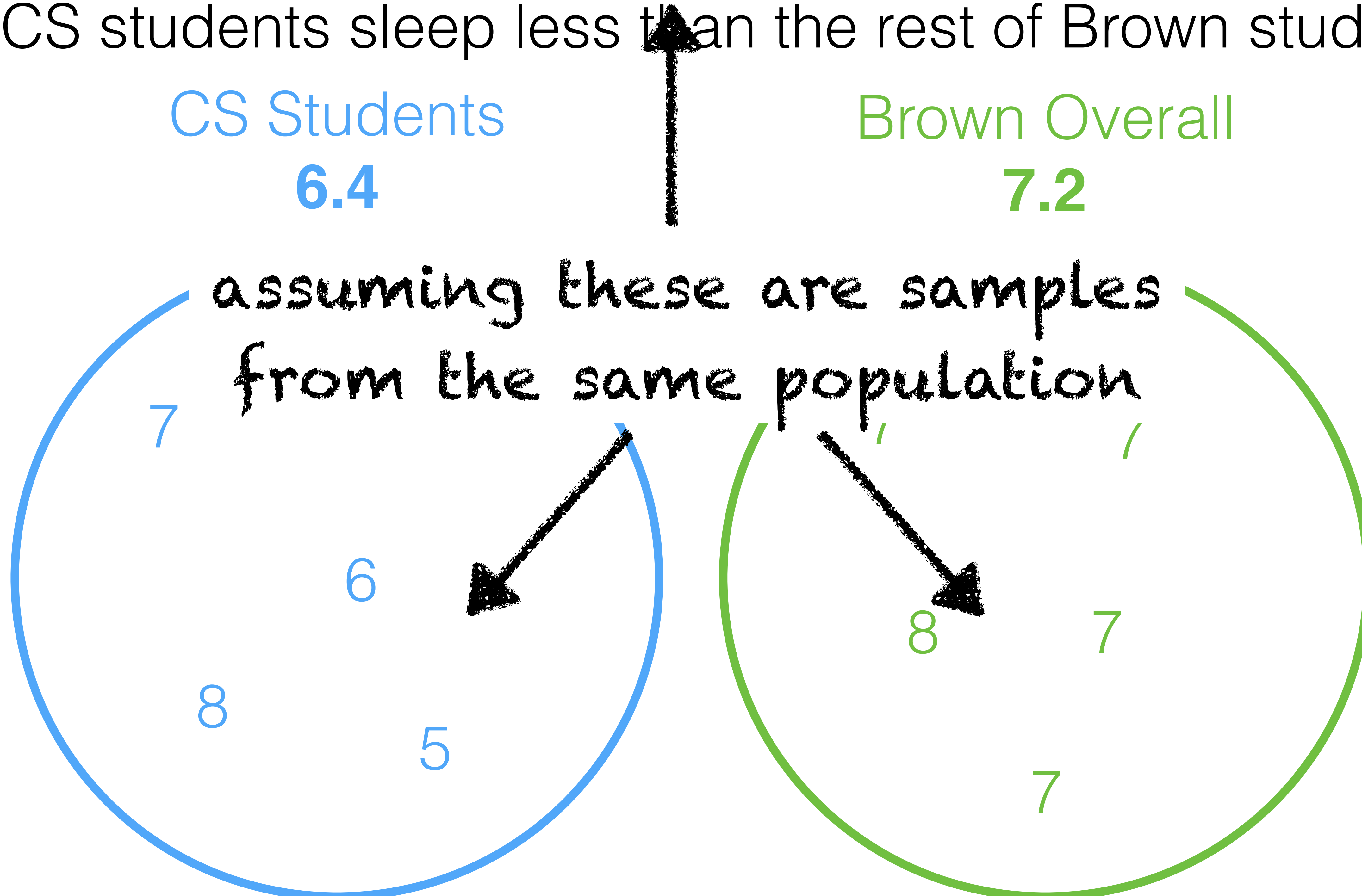
H_0: CS students sleep the same amount as everyone else
H_a: CS students sleep less than the rest of Brown students



Permutation Test

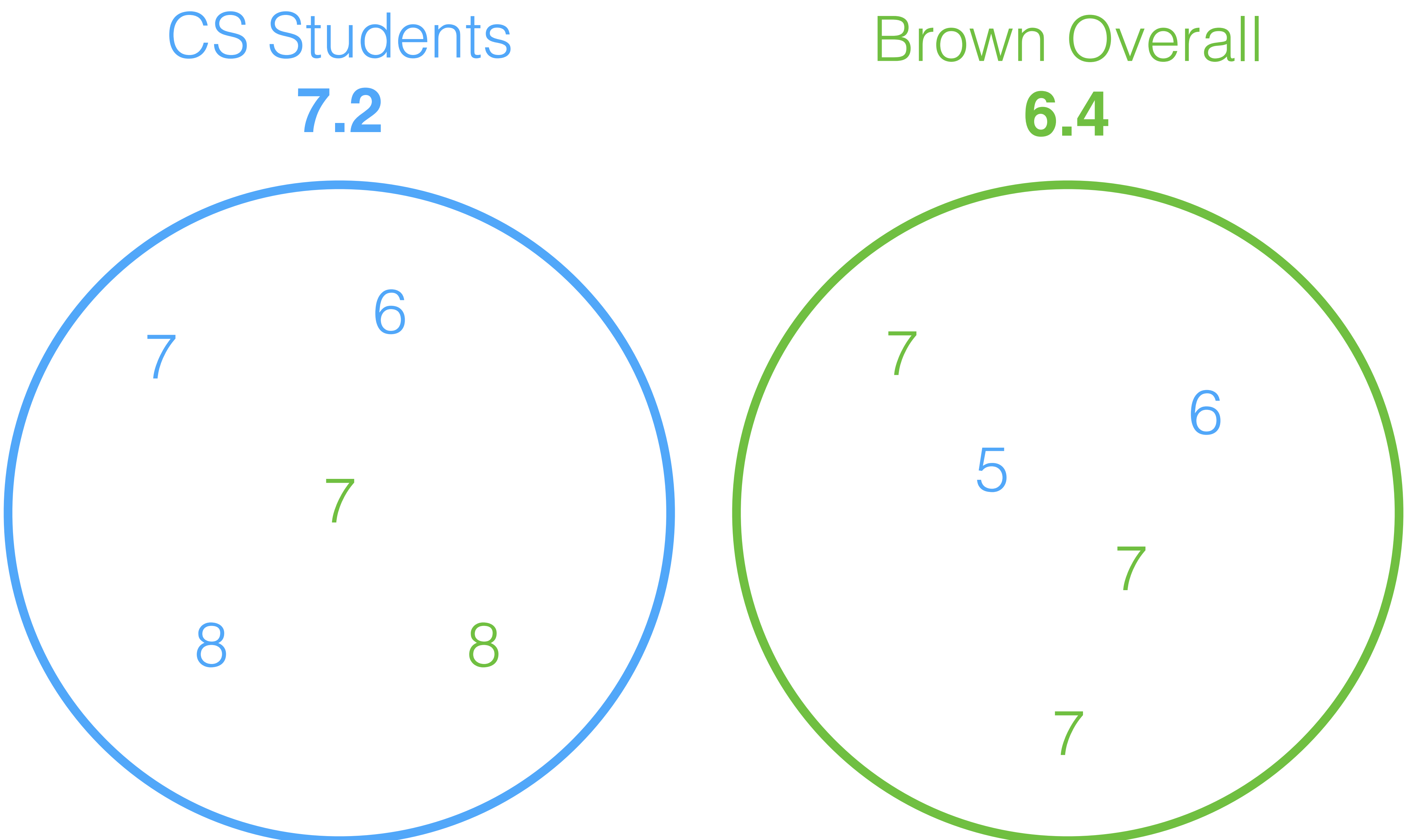
H_0: CS students sleep **the same** amount as everyone else

H_a: CS students sleep less than the rest of Brown students



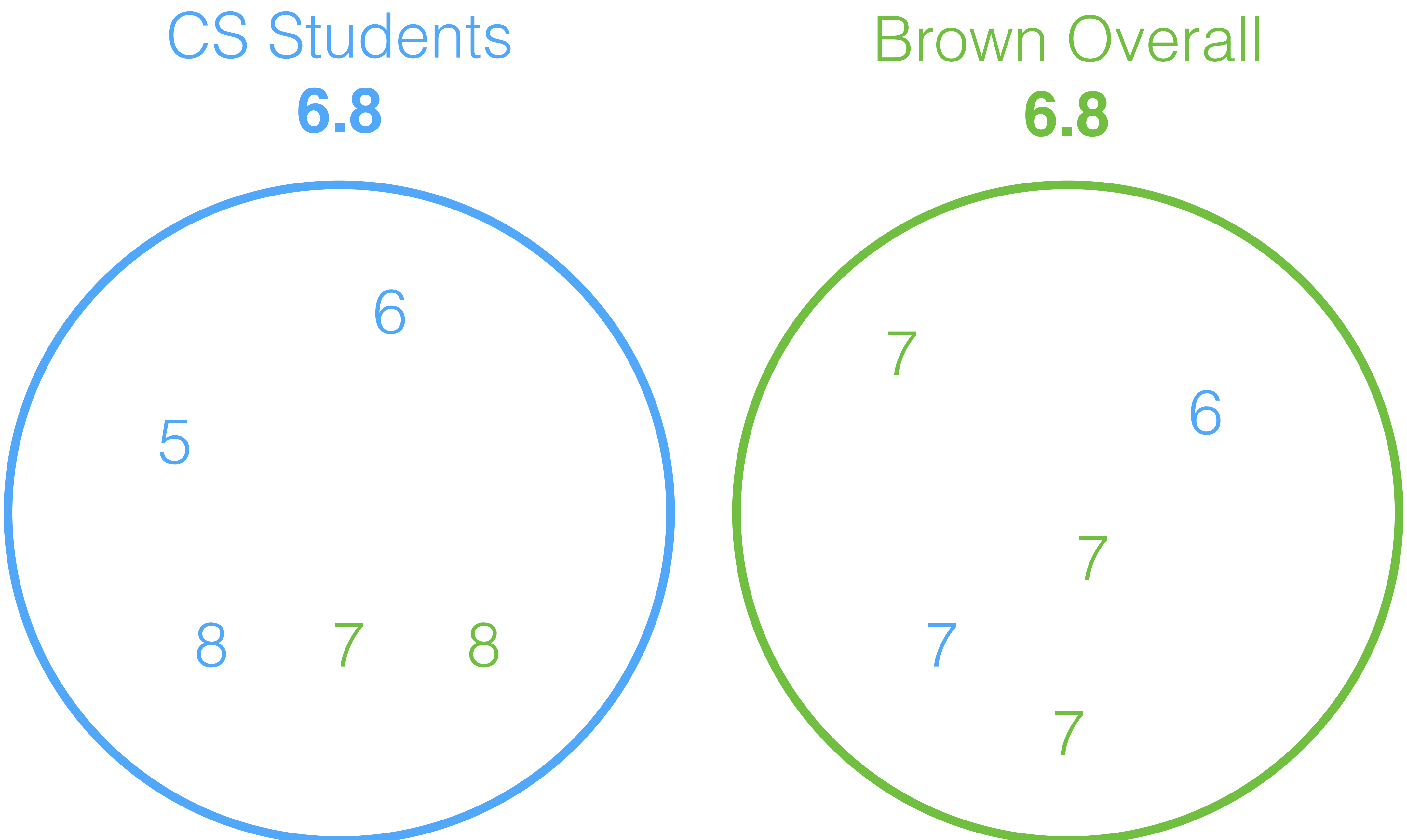
Permutation Test

H_0: CS students sleep the same amount as everyone else
H_a: CS students sleep less than the rest of Brown students



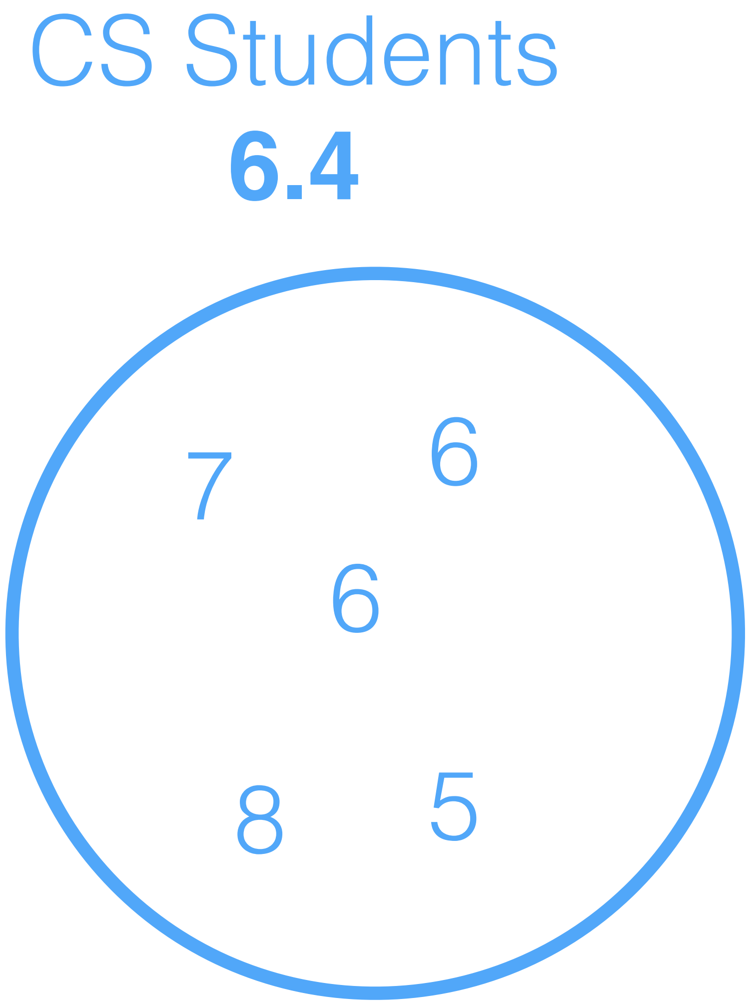
Permutation Test

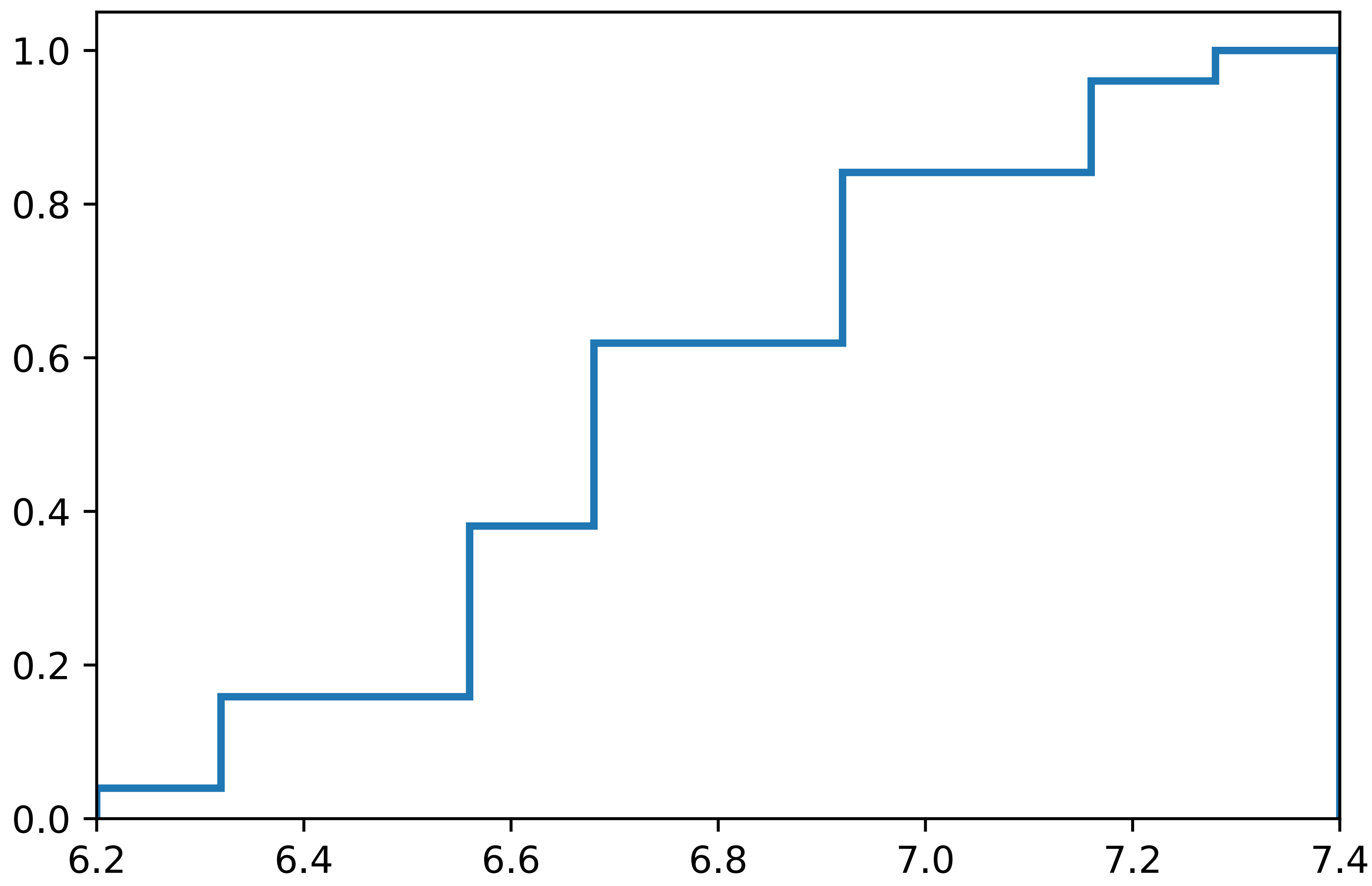
H_0: CS students sleep the same amount as everyone else
H_a: CS students sleep less than the rest of Brown students



Permutation Test

H_0: CS students sleep the same amount as everyone else
H_a: CS students sleep less than the rest of Brown students

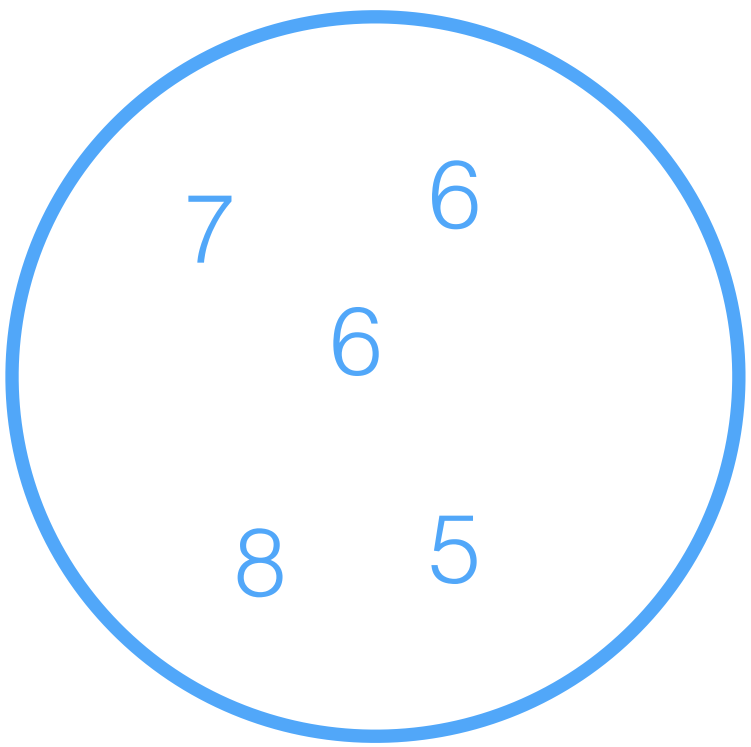


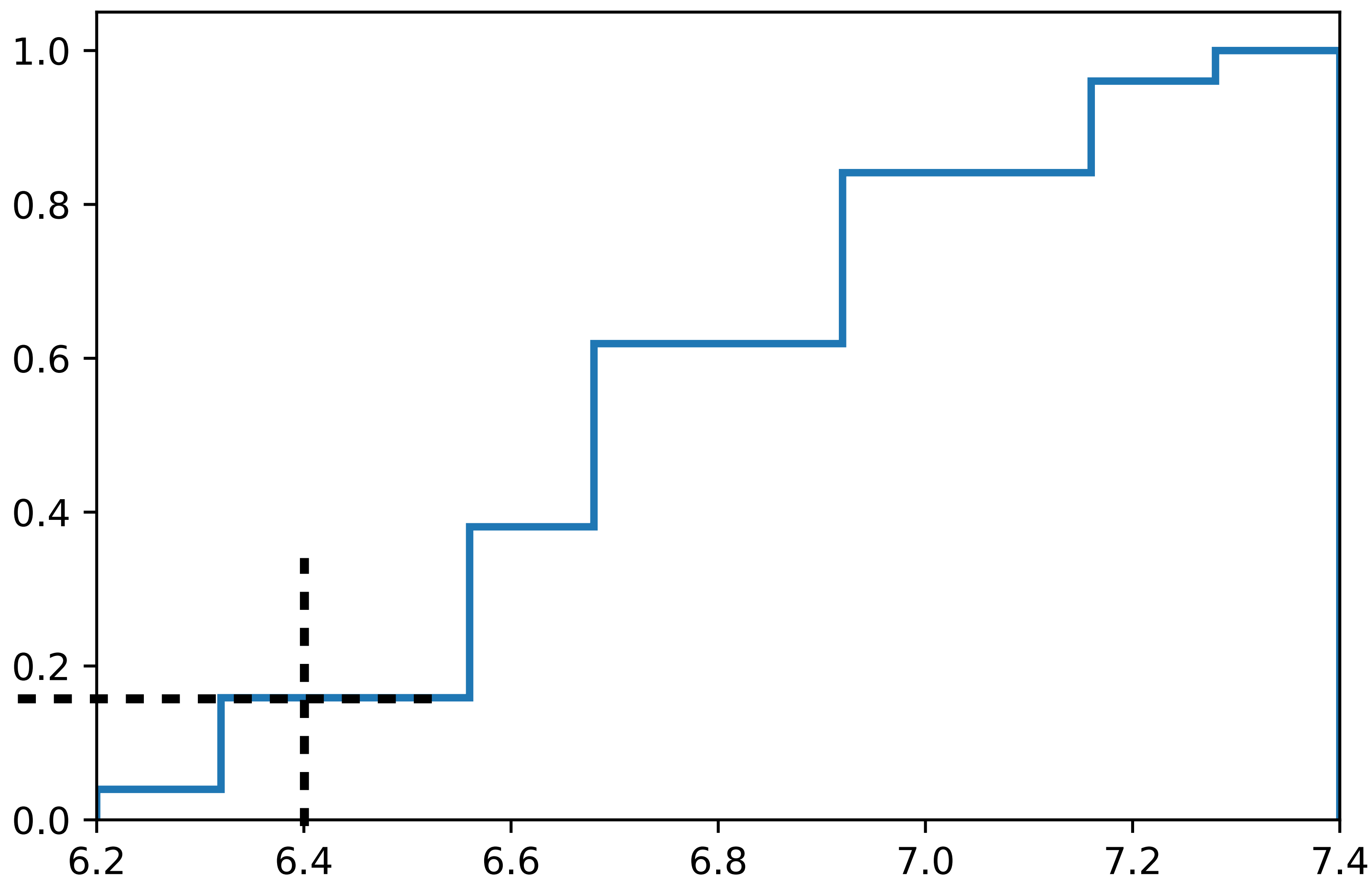


Permutation Test

H_0: CS students sleep the same amount as everyone else
H_a: CS students sleep less than the rest of Brown students

CS Students
6.4





Today

- Non-Parametric Methods
- **Simulations (example using Gaussian Mixture Models)**

Simulations

H_0: I swear there are two types of TAs: nice ones and mean ones. If you get a mean one, you fail, otherwise you pass. Your work doesn't really factor in at all.

Simulations

H_0: I swear there are two types of TAs: nice ones and mean ones. If you get a mean one, you fail, otherwise you pass. Your work doesn't really factor in at all.

???

Simulations

H_0: I swear there are two types of TAs: nice ones and mean ones. If you get a mean one, you fail, otherwise you pass. Your work doesn't really factor in at all.

```
if (TA is nice):
    student passes (grade of 90)
else:
    student fails (grade of 60)
```

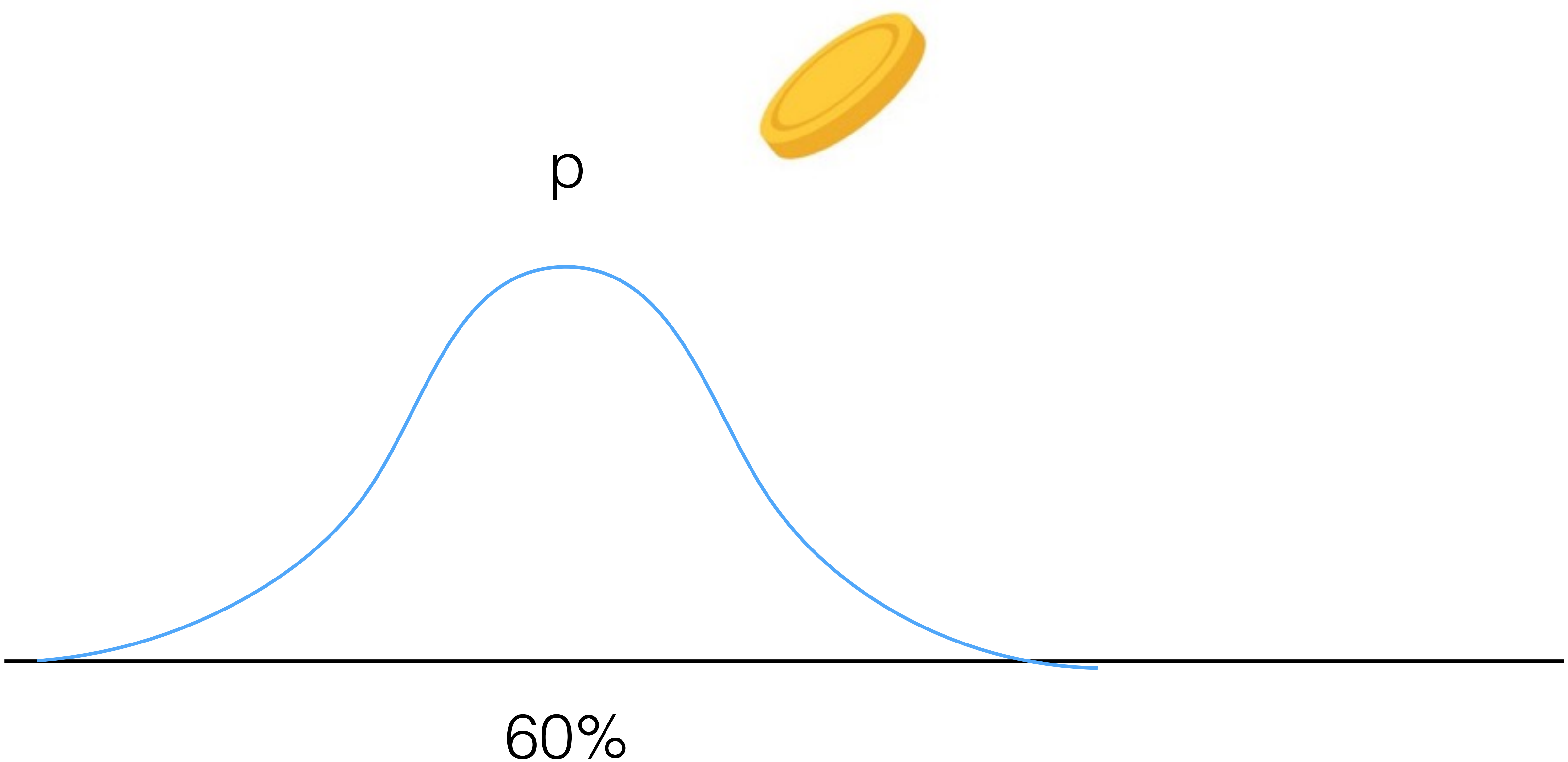
Clicker Question!

Simulations

H_0: I swear there are two types of TAs: nice ones and mean ones. If you get a mean one, you fail, otherwise you pass. Your work doesn't really factor in at all.

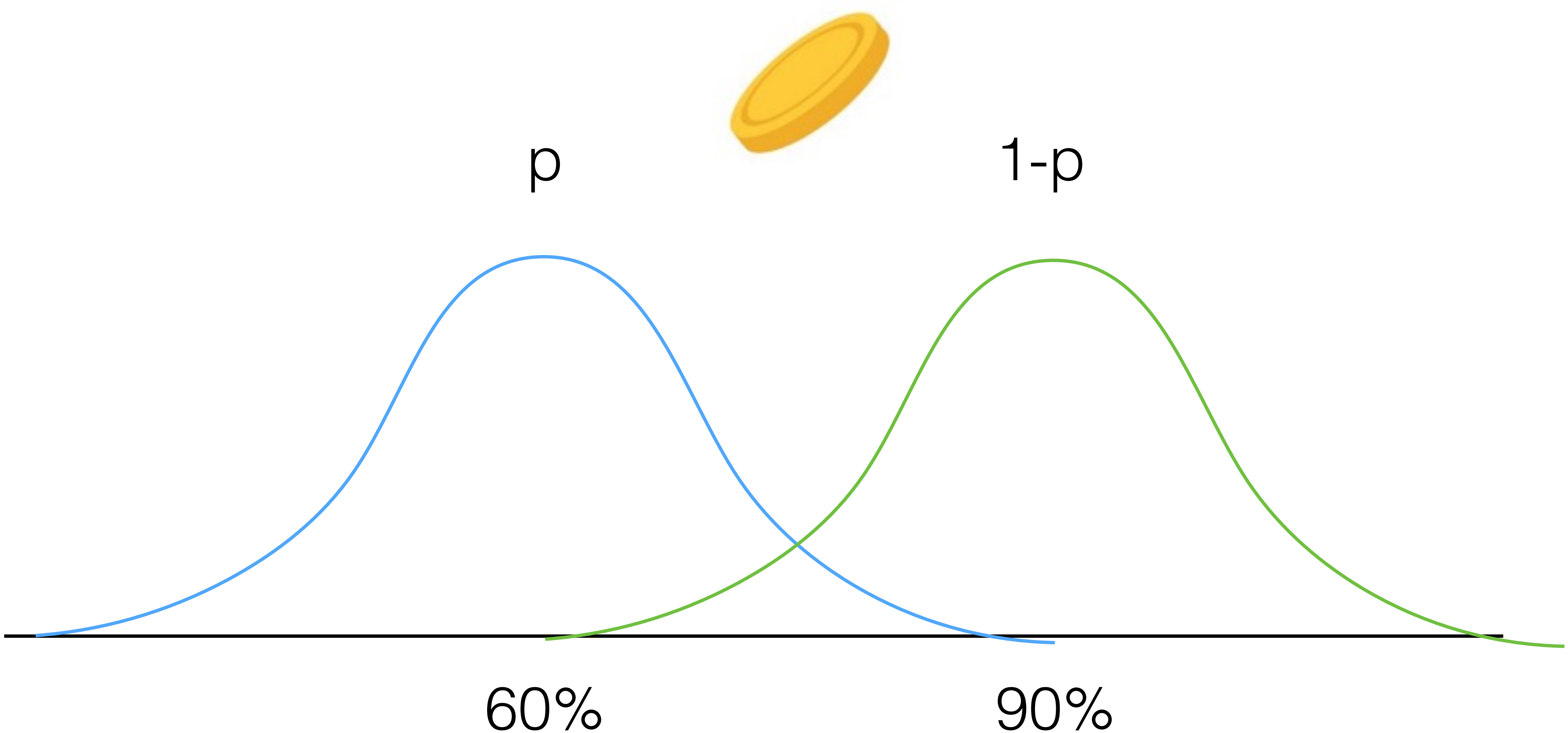
Simulations

H_0: I swear there are two types of TAs: nice ones and mean ones. If you get a mean one, you fail, otherwise you pass. Your work doesn't really factor in at all.



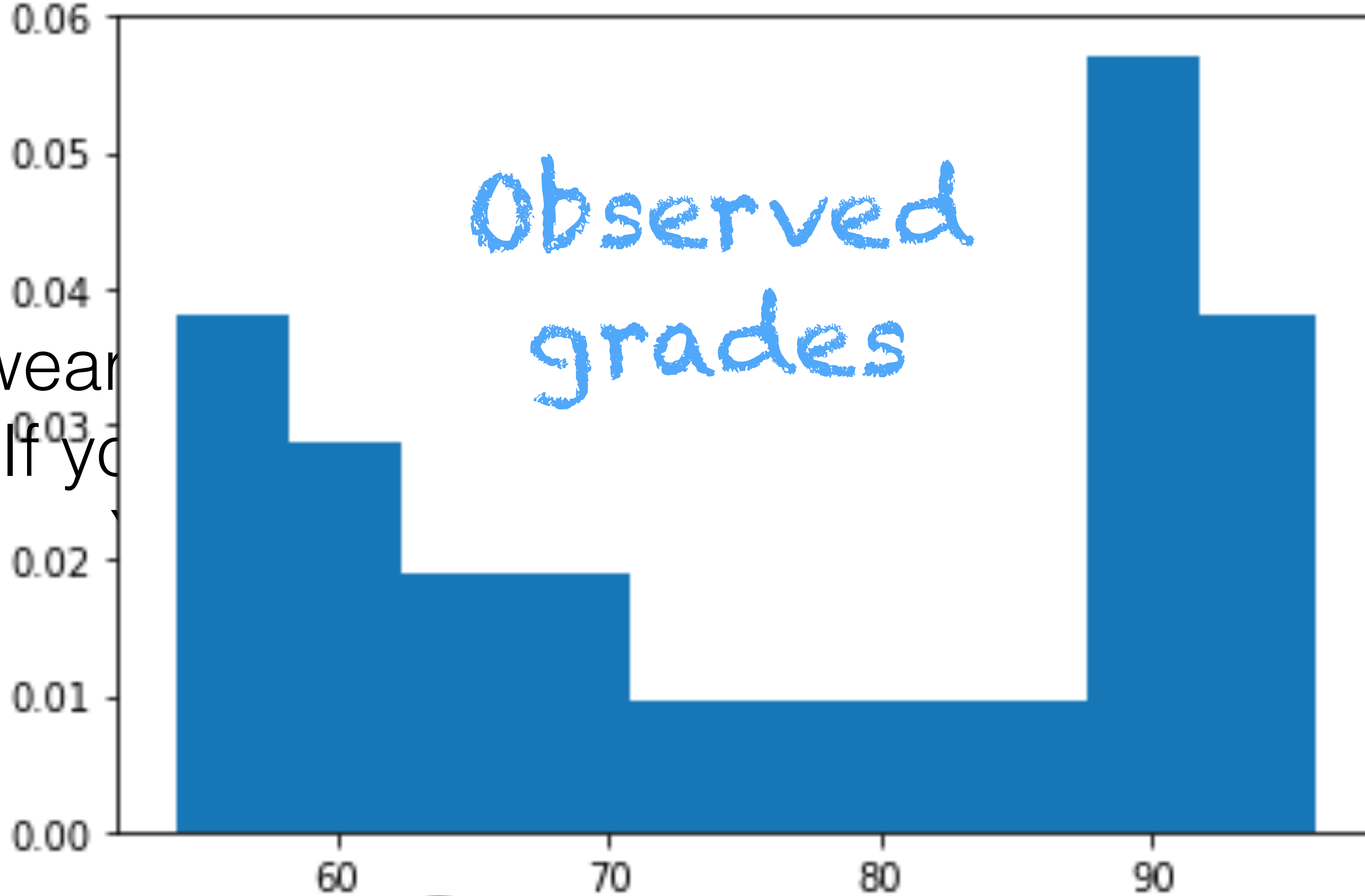
Simulations

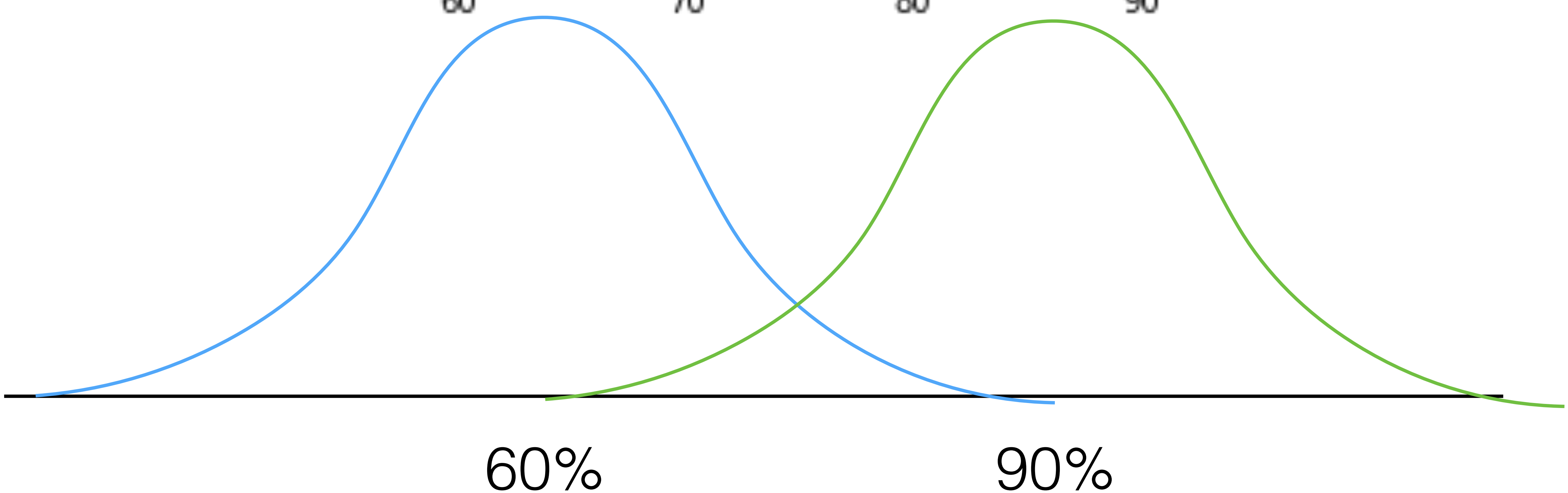
H_0: I swear there are two types of TAs: nice ones and mean ones. If you get a mean one, you fail, otherwise you pass. Your work doesn't really factor in at all.



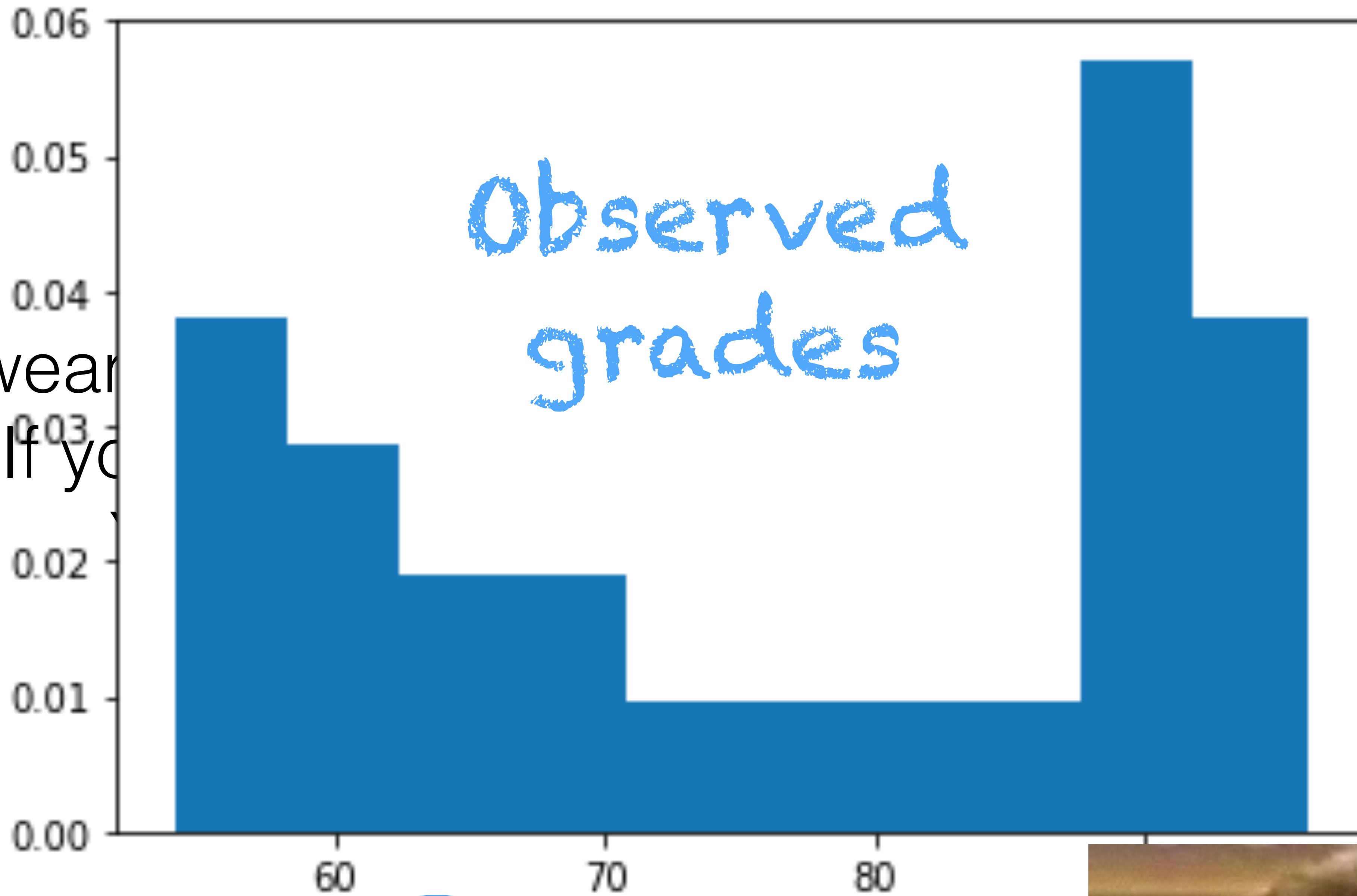
H_0: I swear ... and mean

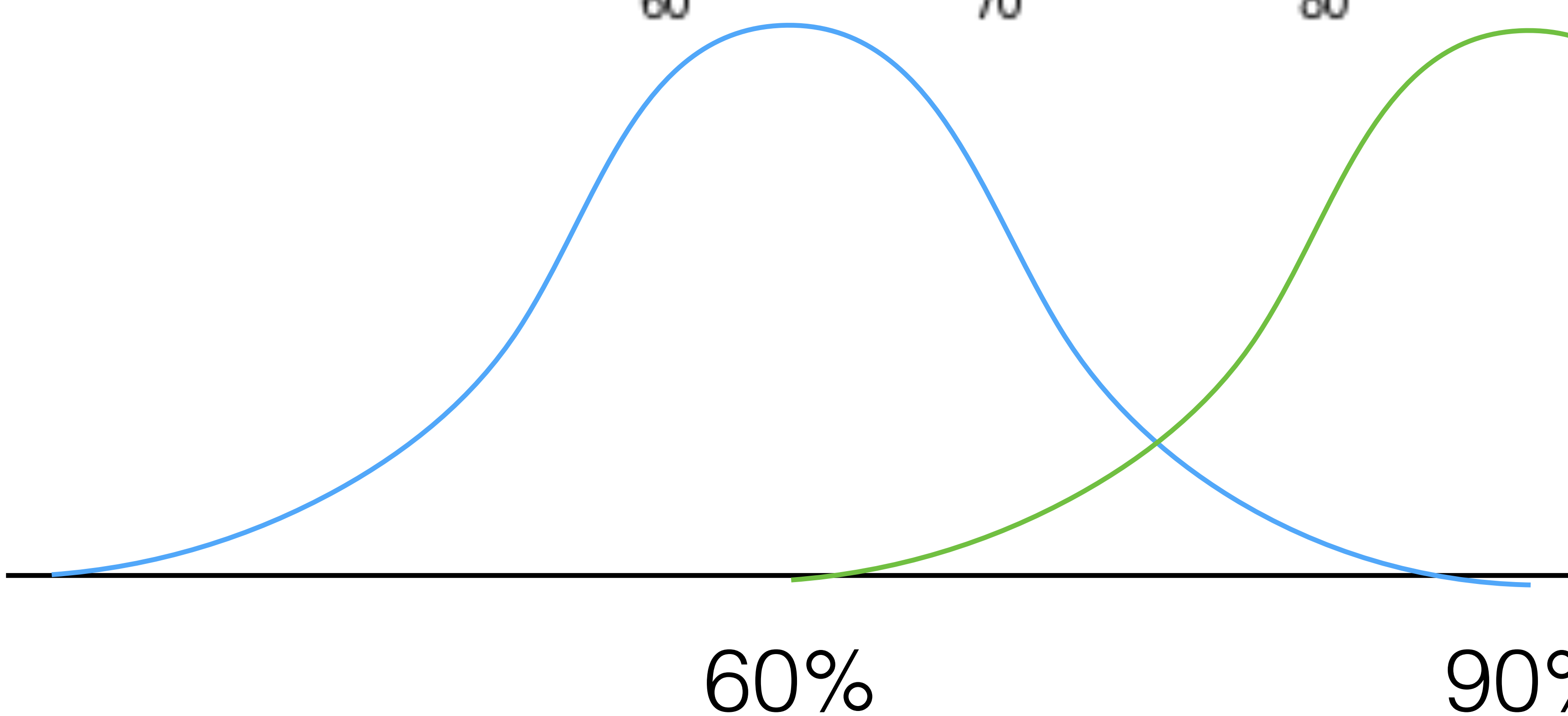
ones. If yo ... ou pass.





H_0: I swear … and mean ones. If yo… ou pass.





Clicker Question!