BC COMS 1016: Intro to Comp Thinking & Data Science

Lecture 19 – Confidence Interval Standard Deviation Normal Distributions

Announcements

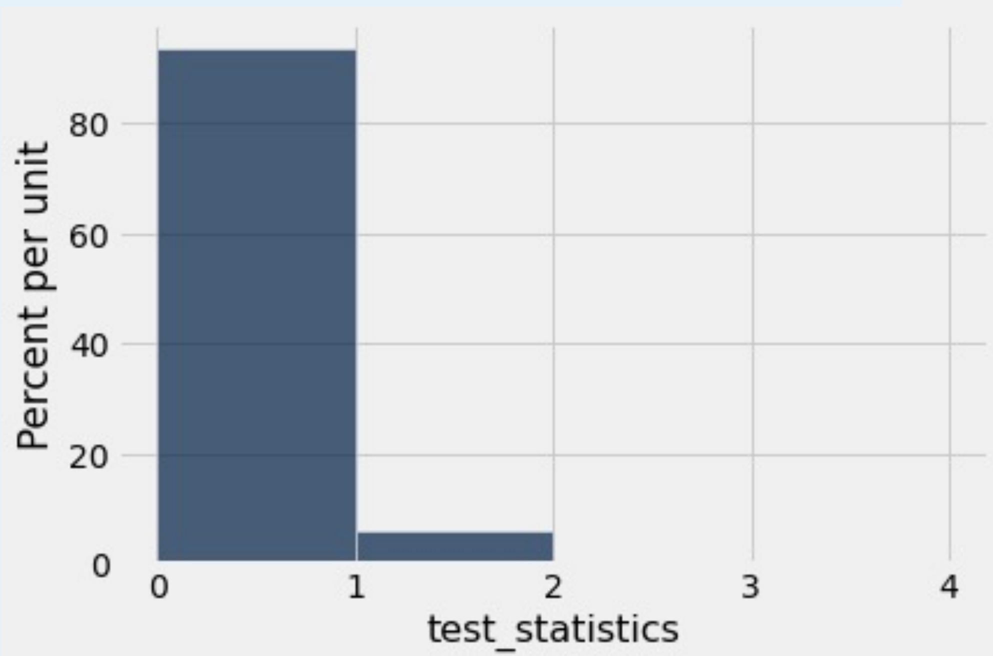
- Checkpoint/Project 2 (midterm):
  - due Monday 04/18
- No Lab this week
- Homework 7 - Confidence Intervals, Resampling, the Bootstrap, and the Central Limit Theorem
  - Due Thursday 04/07
- Dropping 1 homeworks and 1 lab
- Speak up!!
  - More posts on ed-stem – great job!

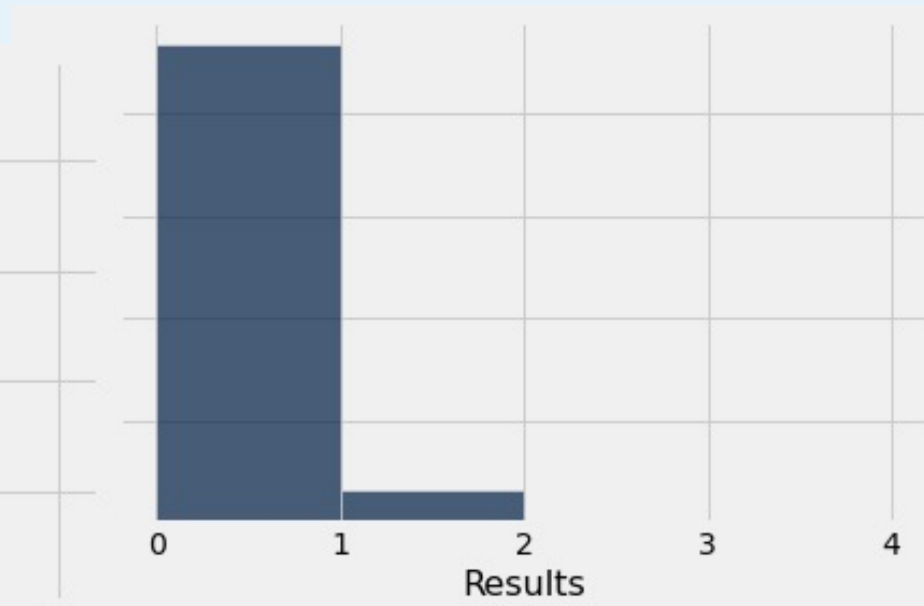
On naming histograms

- When running simulations, use label names to make it clear these are under simulation

On naming histograms

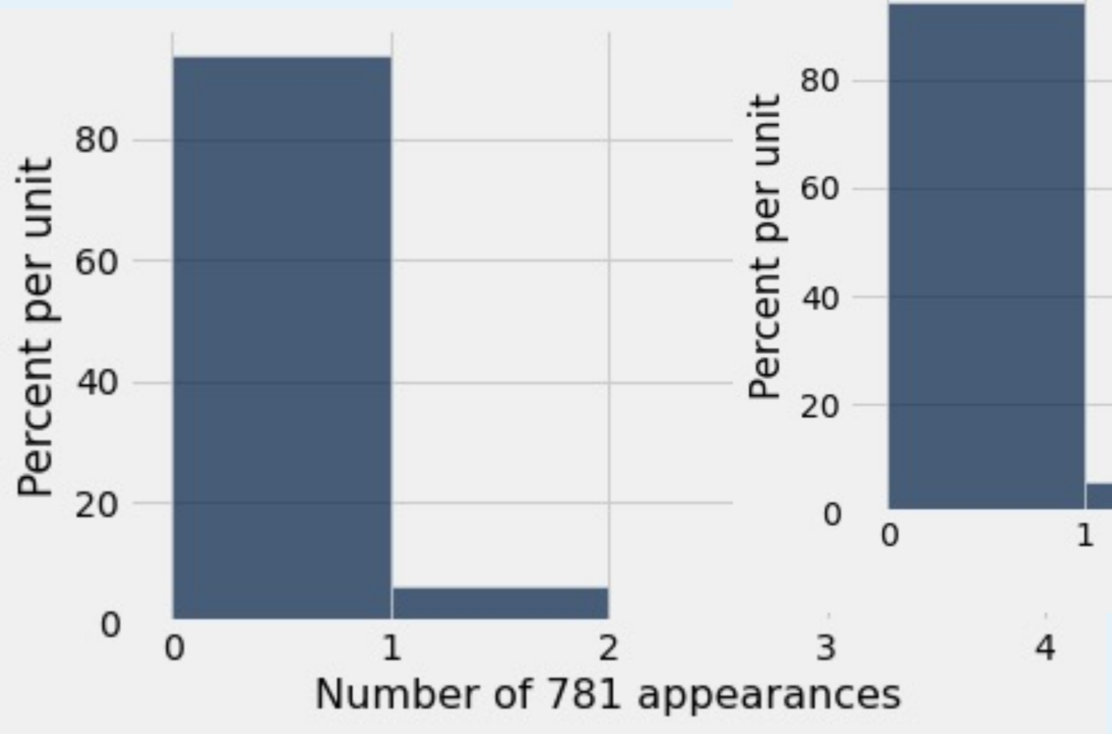
- When running simulations, use label names to make is clear these are under simulation

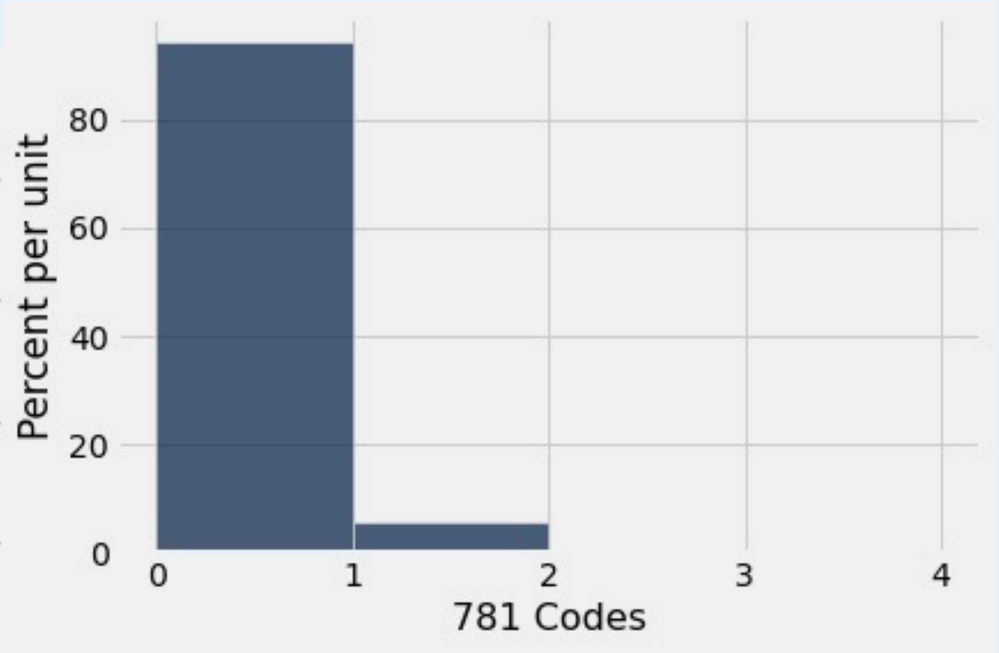




On naming histograms

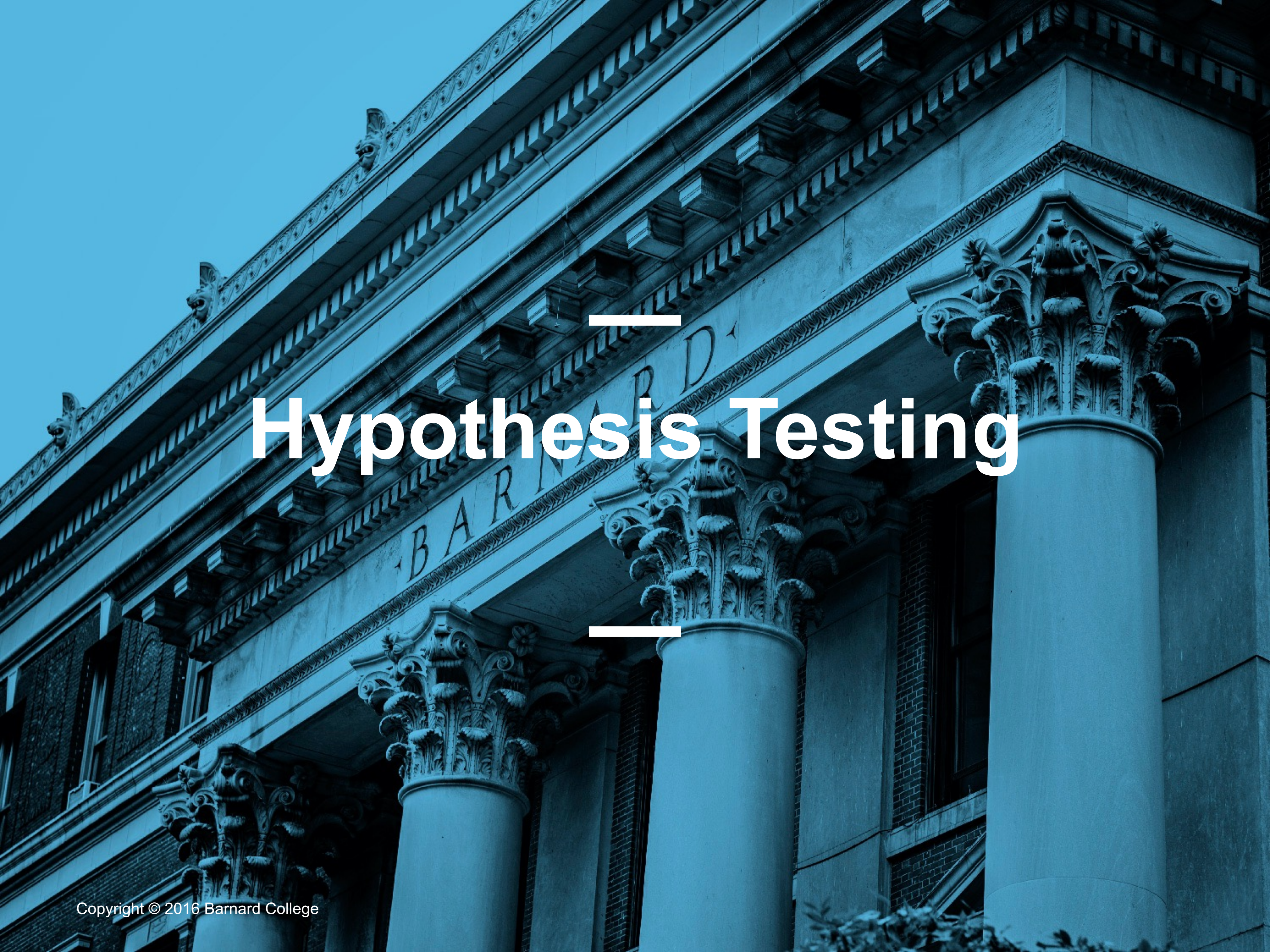
- When running simulations, use label names to make is clear these are under simulation



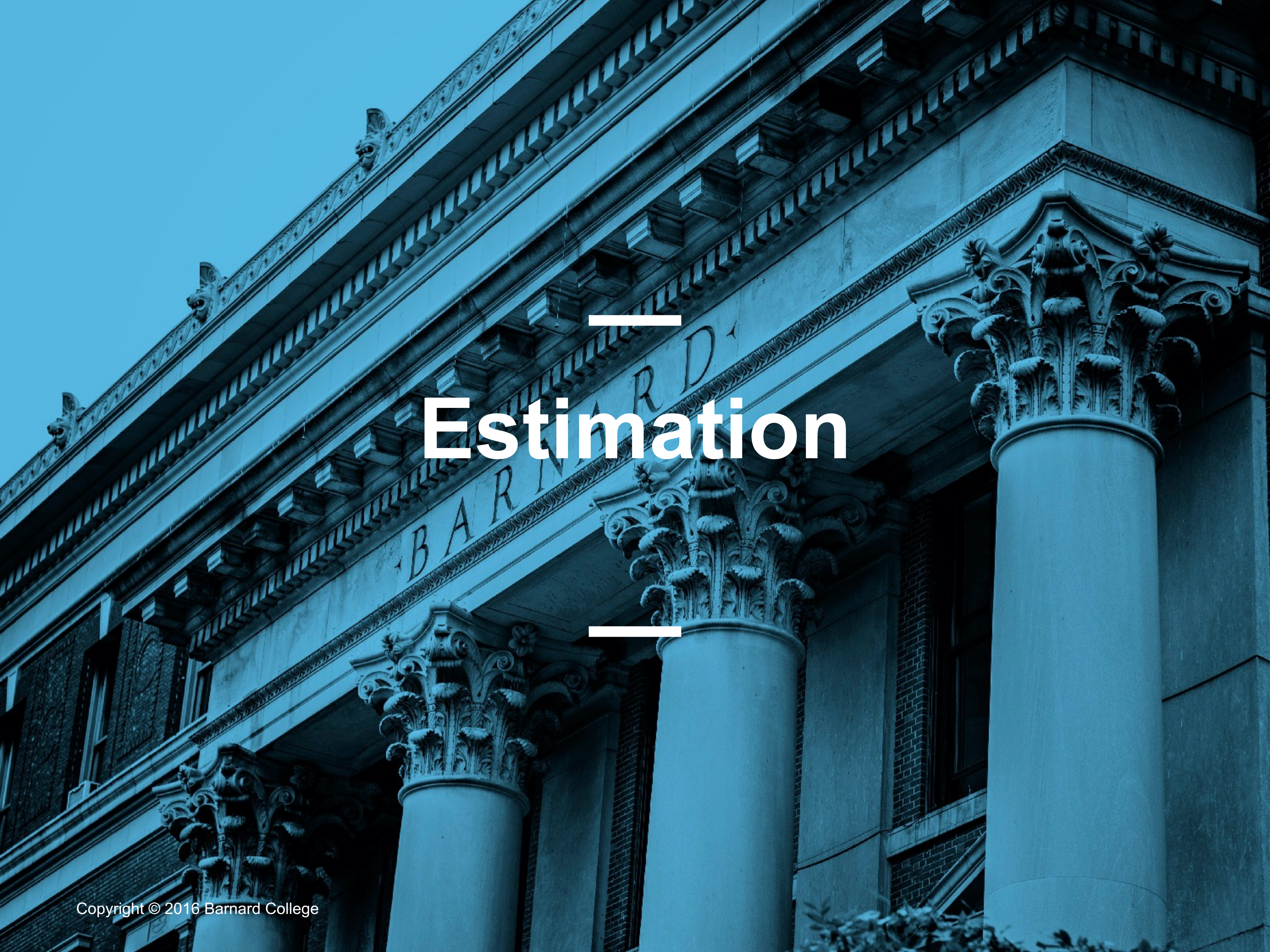


Data Science in this course

- Exploration
  - Discover patterns in data
  - Articulate insights (visualizations)
- Inference
  - Make reliable conclusions about the world
  - Statistics is useful
- Prediction
  - Informed guesses about unseen data



Hypothesis Testing



Estimation

Estimation Variability

The Bootstrap

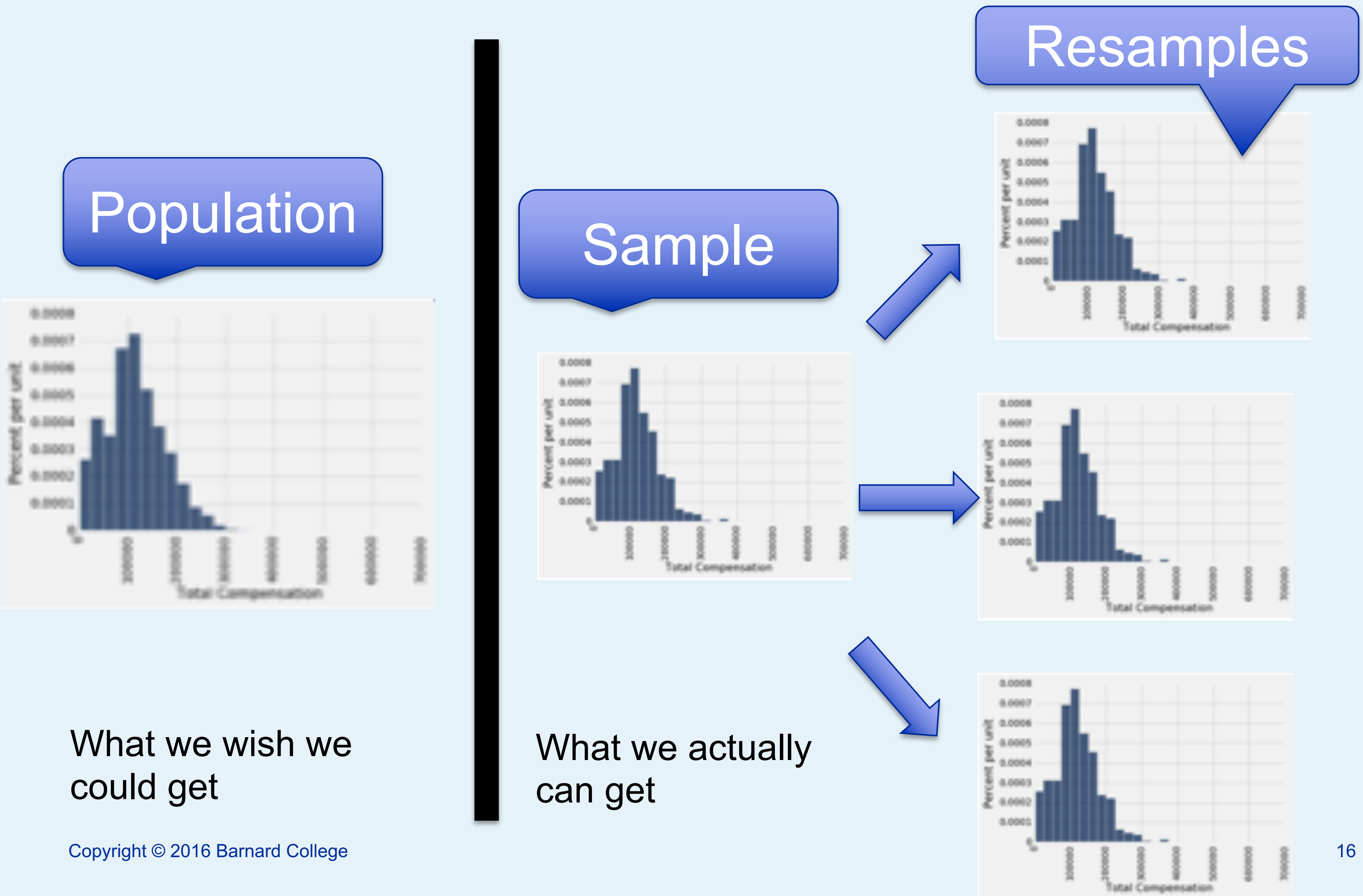
The Bootstrap

- 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!

Why the Bootstrap works



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

Use Methods Appropriately

When to find a Confidence Interval

- You have to guess a parameter for a population
- You have a random sample from the population
  - But not access to the population
- You want to quantify uncertainty
- A statistic is a reasonable estimate of the parameter

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

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?

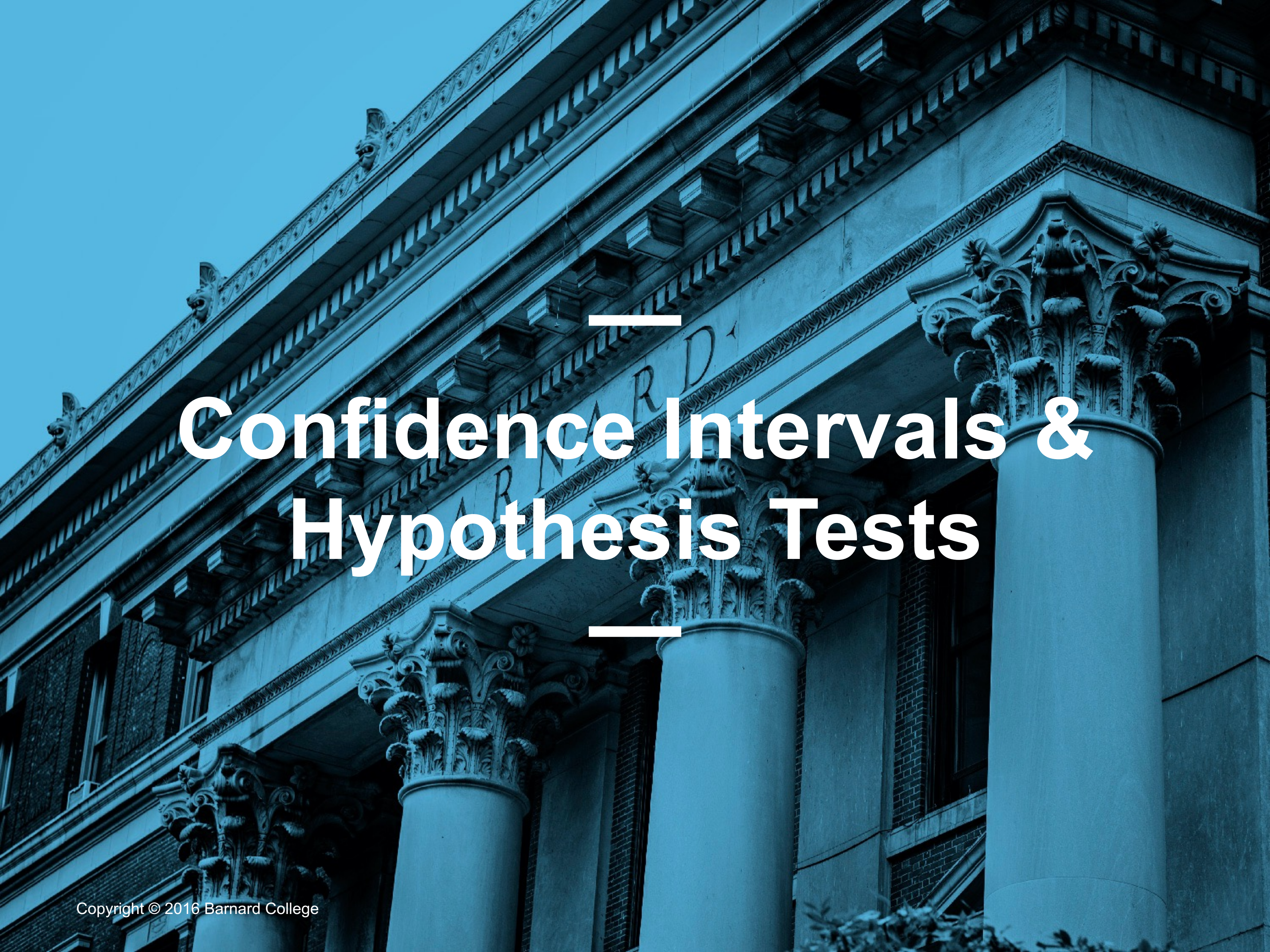
An approximate 95% confidence interval for the average age of the mothers in the population is (26.9, 27.6) years.

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



Confidence Intervals & Hypothesis Tests

Using a CI for Testing

- Null hypothesis: **Population average = *x***
- Alternative hypothesis: **Population average ≠ *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

Data Science in this course

- Exploration
  - Discover patterns in data
  - Articulate insights (visualizations)

- Inference
  - Make reliable conclusions about the world
  - Statistics is useful

- **Prediction**
  - **Informed guesses about unseen data**

Center & Spread

Questions/Goals

- How can we quantify natural concepts like "center" and "variability"?
- Why do many of the empirical distributions that we generate come out bell shaped?
- How is sample size related to the accuracy of an estimate?

Average and the Histogram

The average (mean)

Data: 2, 3, 3, 9

Average = (2+3+3+9)/4 = 4.25

- Need not be a value in the collection
- Need not be an integer even if the data are integers
- Somewhere between min and max, but not necessarily halfway in between
- Same units as the data
- Smoothing operator: collect all the contributions in one big pot, then split evenly

Relation to the histogram

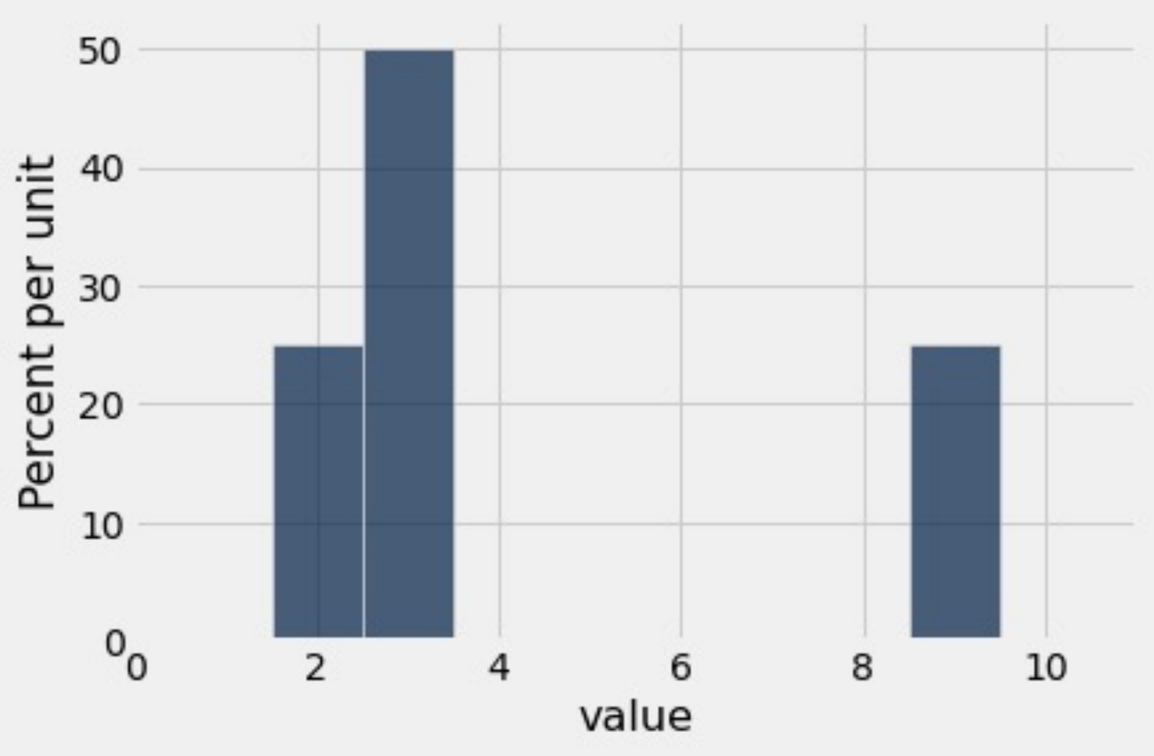
- The average depends only on the **proportions** in which the distinct values appears

- The average is the **center of gravity** of the histogram

- It is the point on the horizontal axis where the histogram balances

Average as balance point

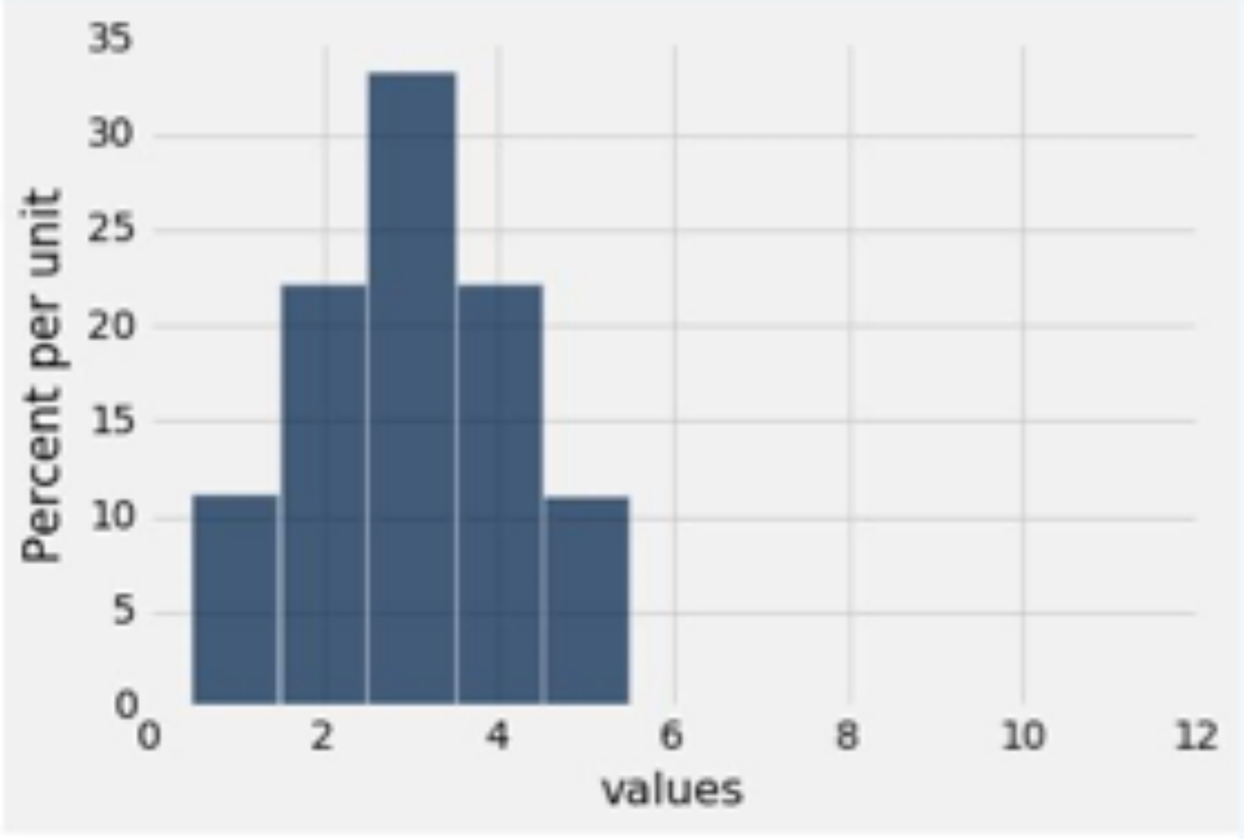
- Average is 4.25



Average and Median

Question

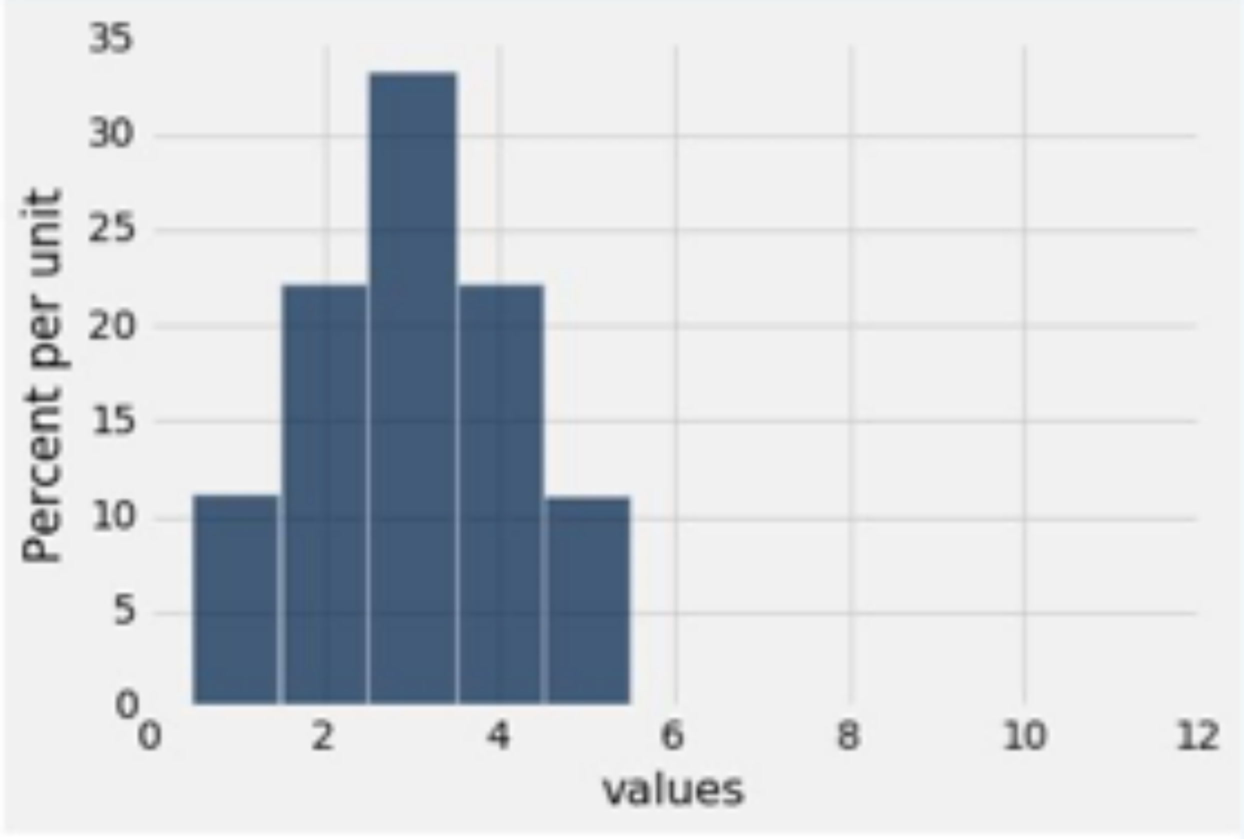
- What list produces this histogram?



Question

- What list produces this histogram?

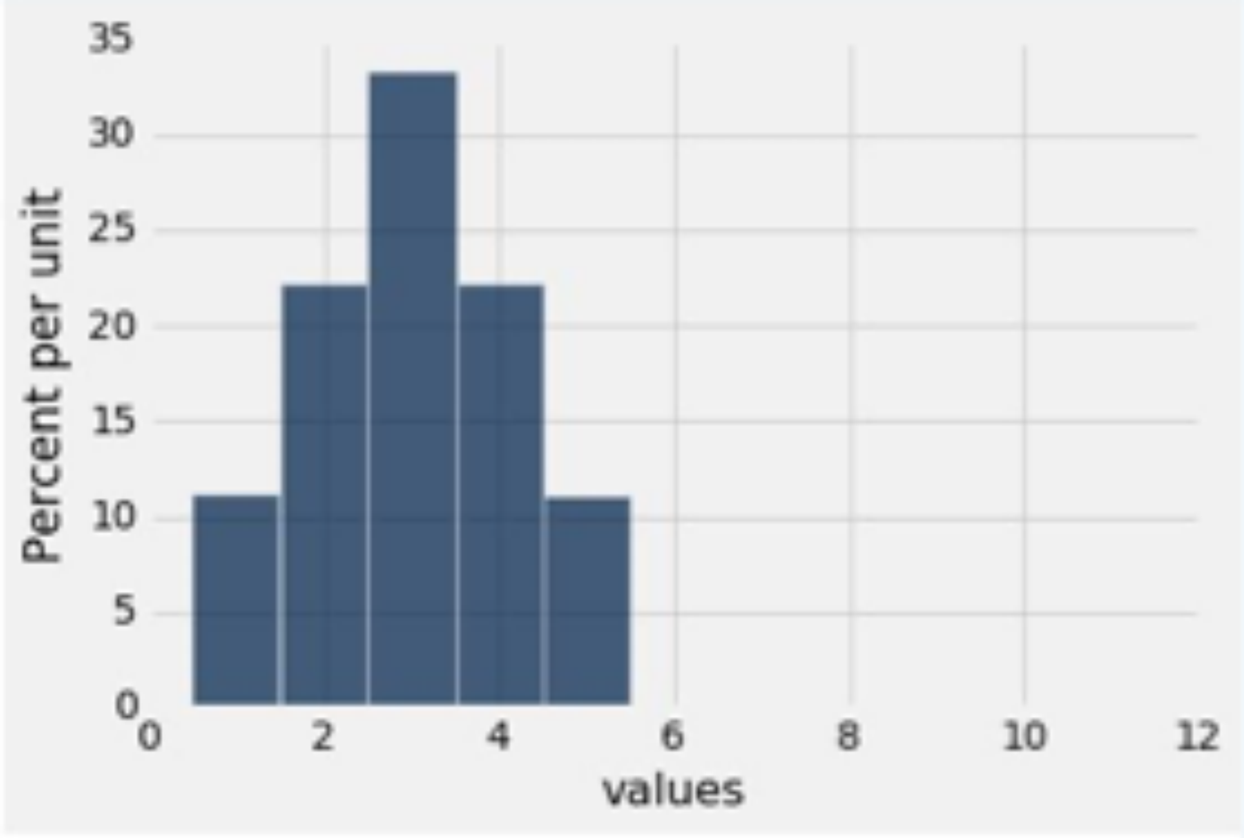
1, 2, 2, 3, 3
3, 4, 4, 5



Question

- What list produces this histogram?

1, 2, 2, 3, 3
3, 4, 4, 5

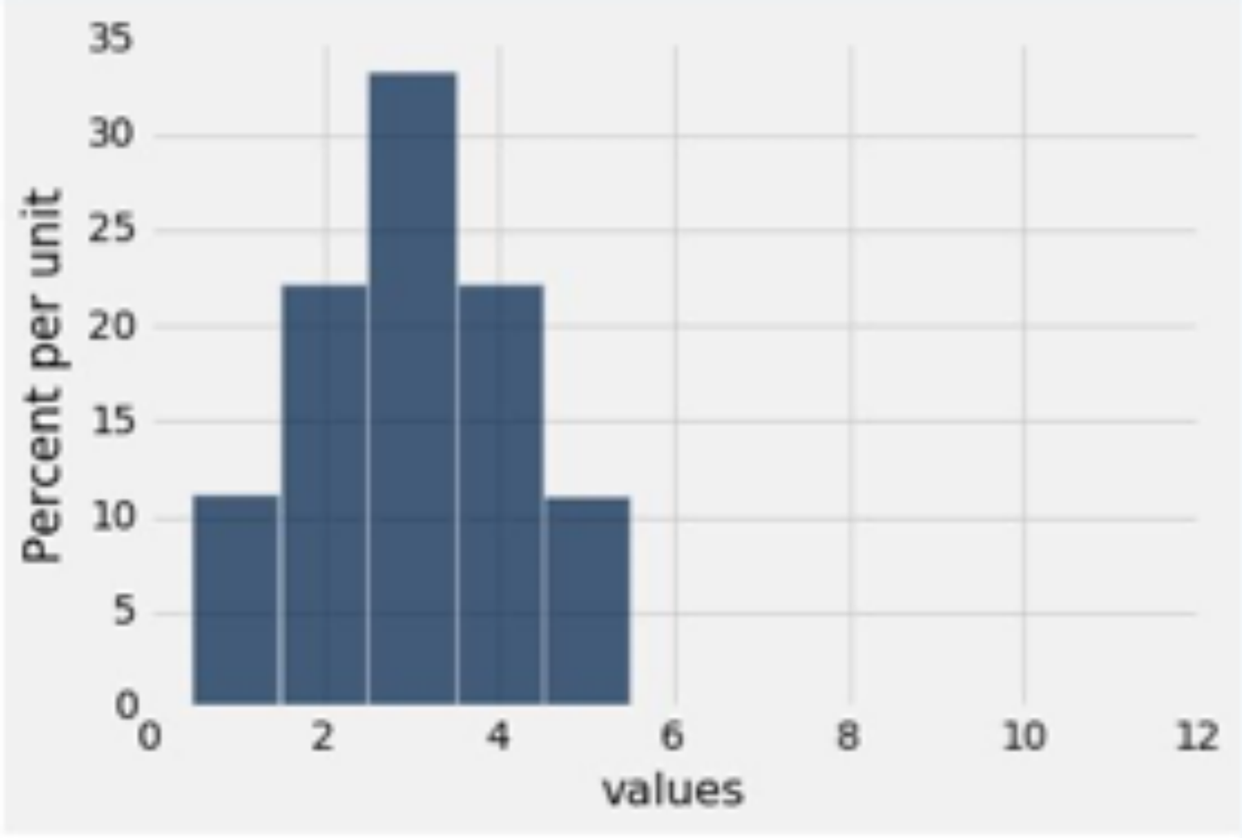


- Average?

Question

- What list produces this histogram?

1, 2, 2, 3, 3
3, 4, 4, 5

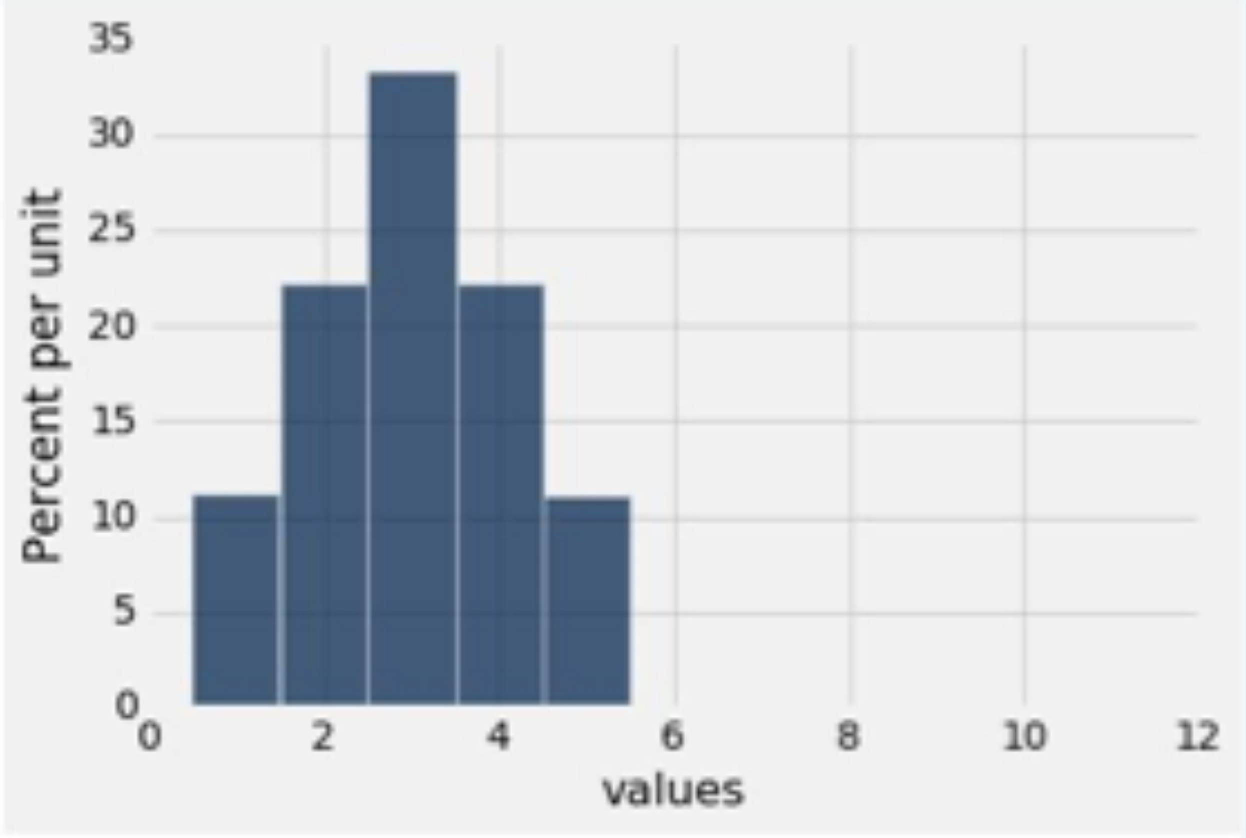


- Average?
  - 3

Question

- What list produces this histogram?

1, 2, 2, 3, 3
3, 4, 4, 5

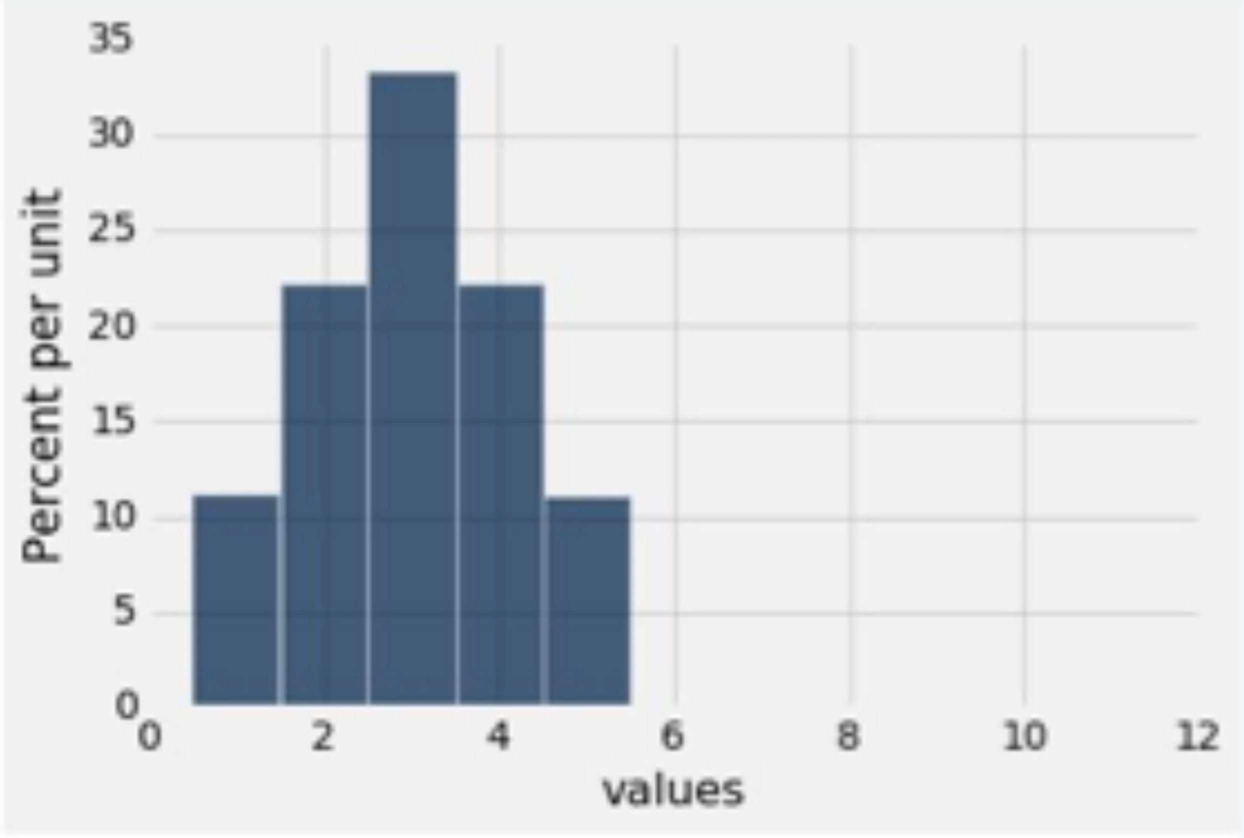


- Average?
  - 3
- Median?

Question

- What list produces this histogram?

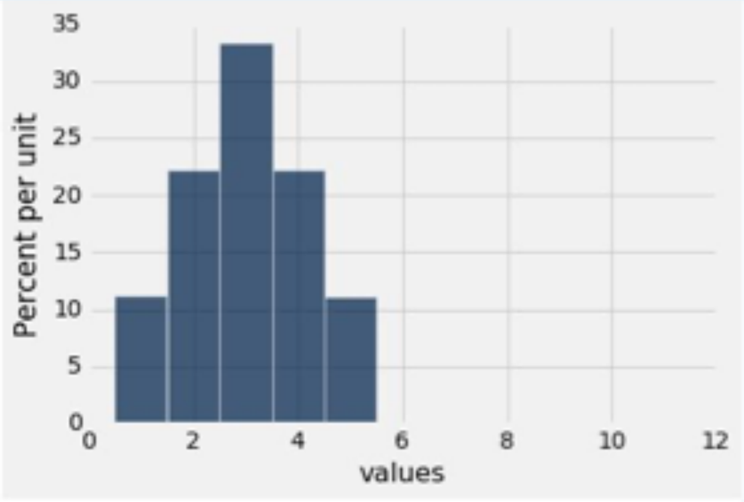
1, 2, 2, 3, 3
3, 4, 4, 5

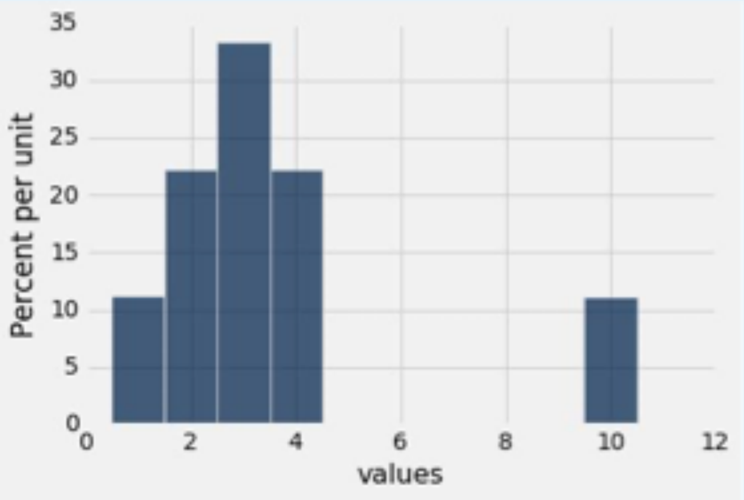


- Average?
  - 3
- Median?
  - 3

Question 2

- Are the medians of these two distributions the same or different?
- Are the means the same or different?
  - If you say "different," then say which one is bigger





Answer 2

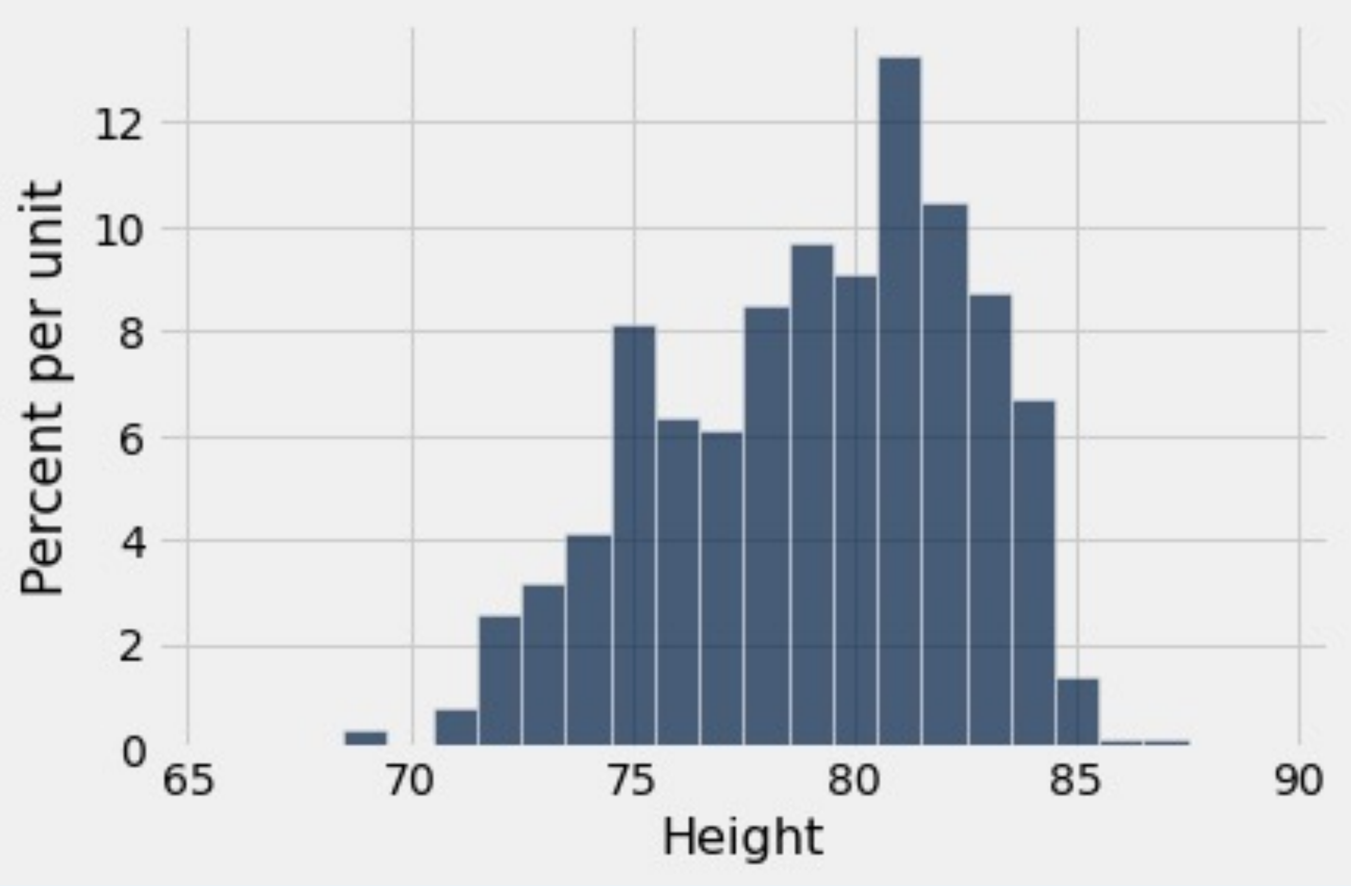
- List 1
  - 1, 2, 2, 3, 3, 3, 4, 4, 5
- List 2
  - 1, 2, 2, 3, 3, 3, 4, 4, 10
- Medians = 3
- Mean(List1) = 3
- Mean (List 2) = 3.55556

Comparing Mean and Median

- **Mean:** Balance point of the histogram
- **Median:** Half-way point of data; half the area of histogram is on either side of median
- If the distribution is symmetric about a value, then that value is both the average and the median.
- If the histogram is skewed, then the mean is pulled away from the median in the direction of the tail.

Question

- Which is bigger, median or mean?



Standard Deviation

Defining Variability

- **Plan A:** “biggest value - smallest value”
  - Doesn’t tell us much about the shape of the distribution
- **Plan B**:
  - Measure variability around the mean
  - Need to figure out a way to quantify this

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
  - Discuss this later

Chebyshev's Inequality

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/z2

Chebyshev's Bounds

Range	Proportion

Chebyshev's Bounds

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

Chebyshev's Bounds

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

Chebyshev's Bounds

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

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

Understanding HW05 Results

Statistics:

Minimum: 7.5

Maximum: 29.0

Mean: 24.55

Median: 25.0

Standard Deviation: 3.96

- At least 50% of the class had scores between 20.59 and 28.51
- At least 75% of the class had scores between 16.62 and 32.47

Standard Units

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

Question

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

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

(2) The SD is about 6 years (standard unit at 33 is close to 1. 33 – 27 = 6)

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

The SD and the Histogram

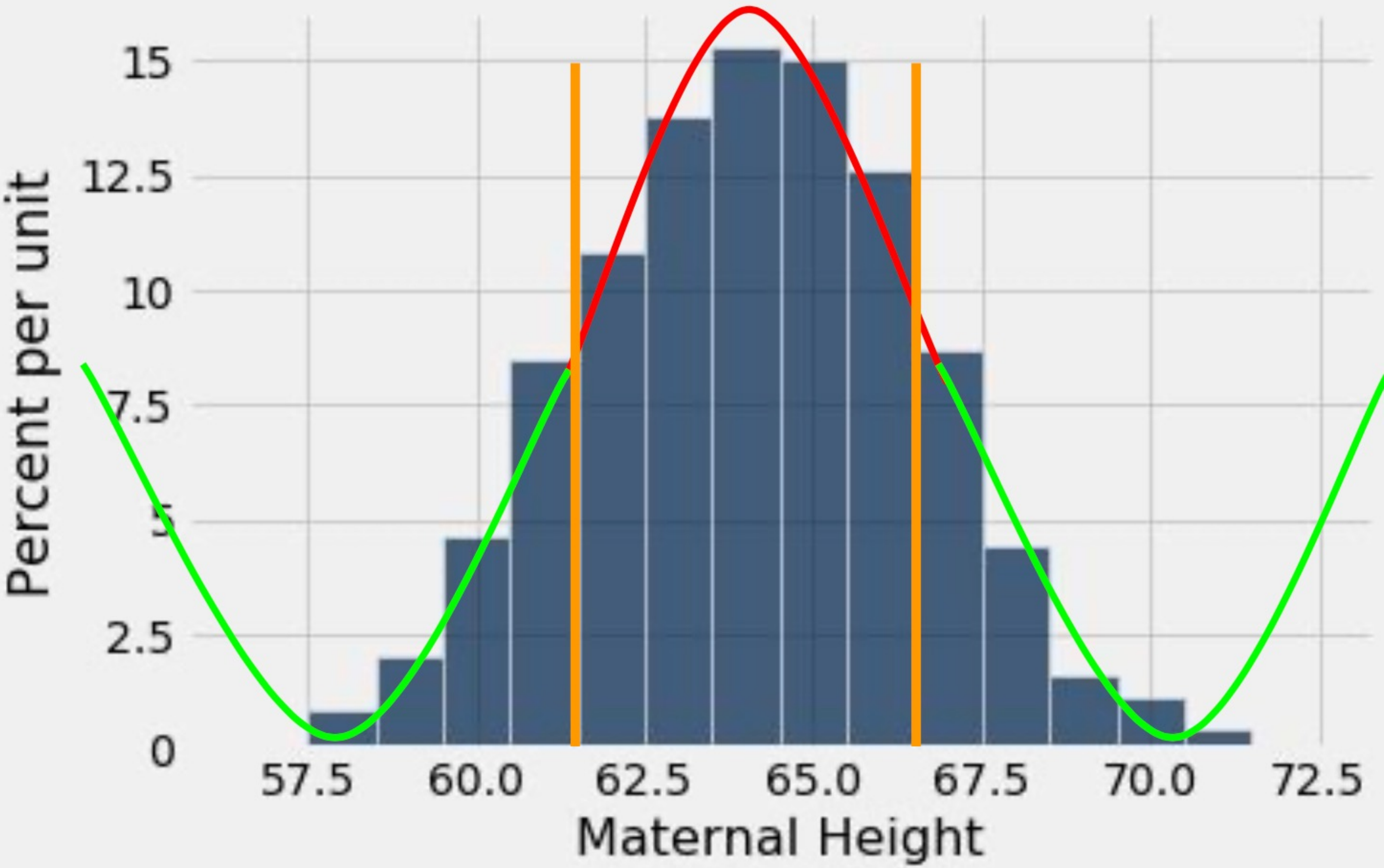
- Usually, it's not easy to estimate the SD by looking at a histogram.
- But if the histogram has a bell shape, then you can

The SD and Bell Shaped Curves

If a histogram is bell-shaped, then

- the average is at the center
- the SD is the distance between the average and the points of inflection on either side

Points of Inflection



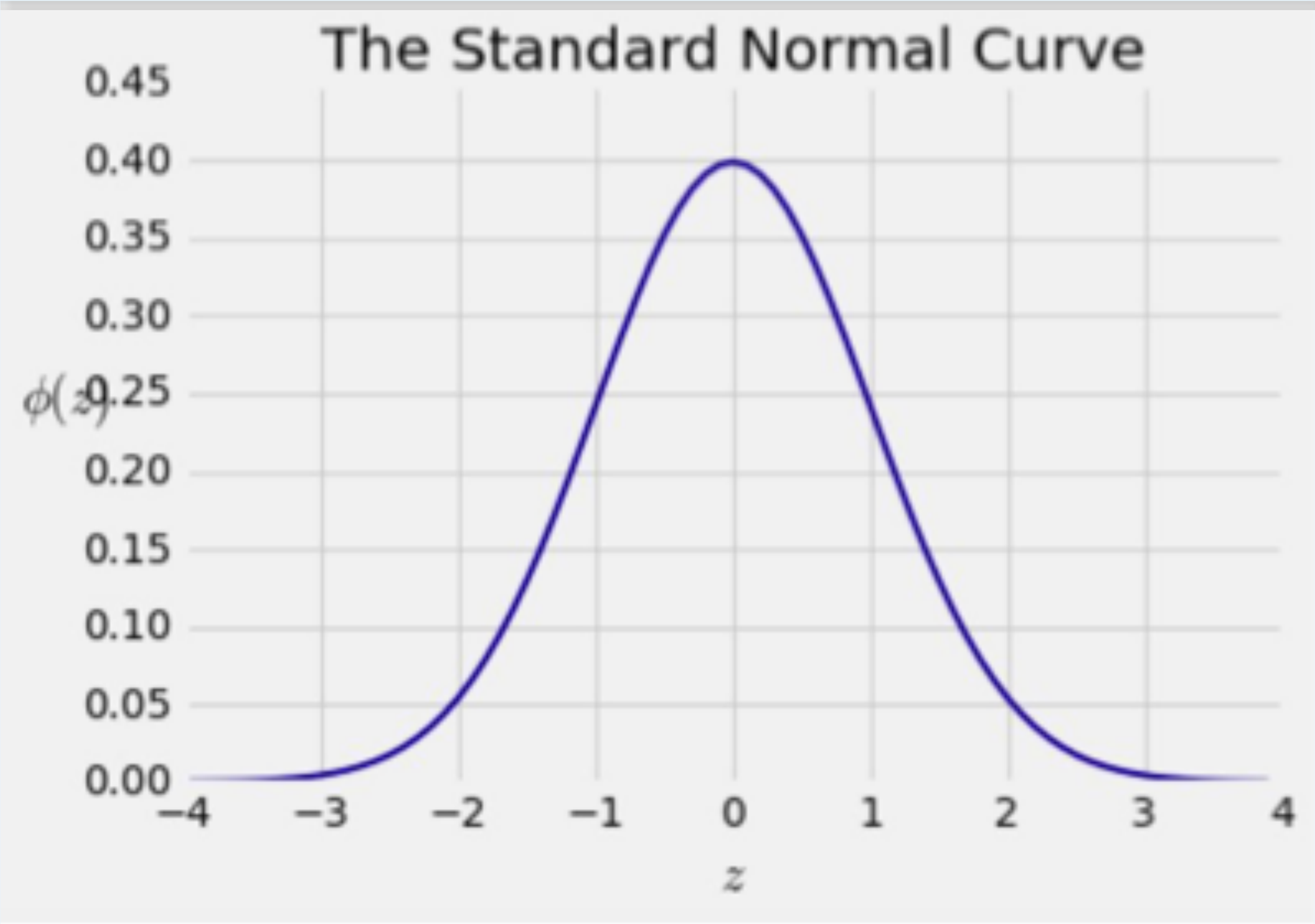
Normal Distribution

Standard Normal Curve

Equation for the normal curve

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}, \qquad -\infty < z < \infty$$

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"

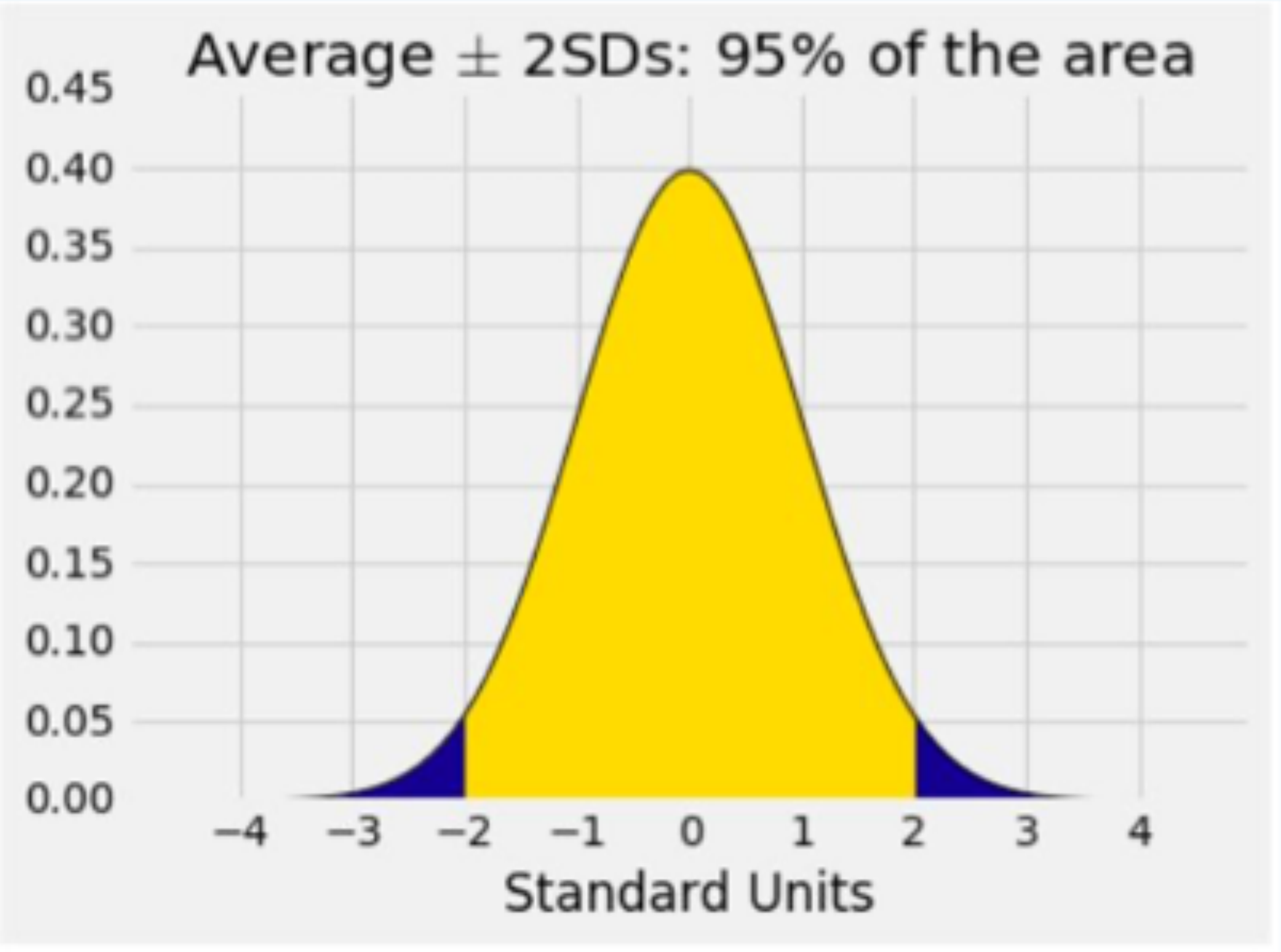
If a histogram is bell-shaped, then

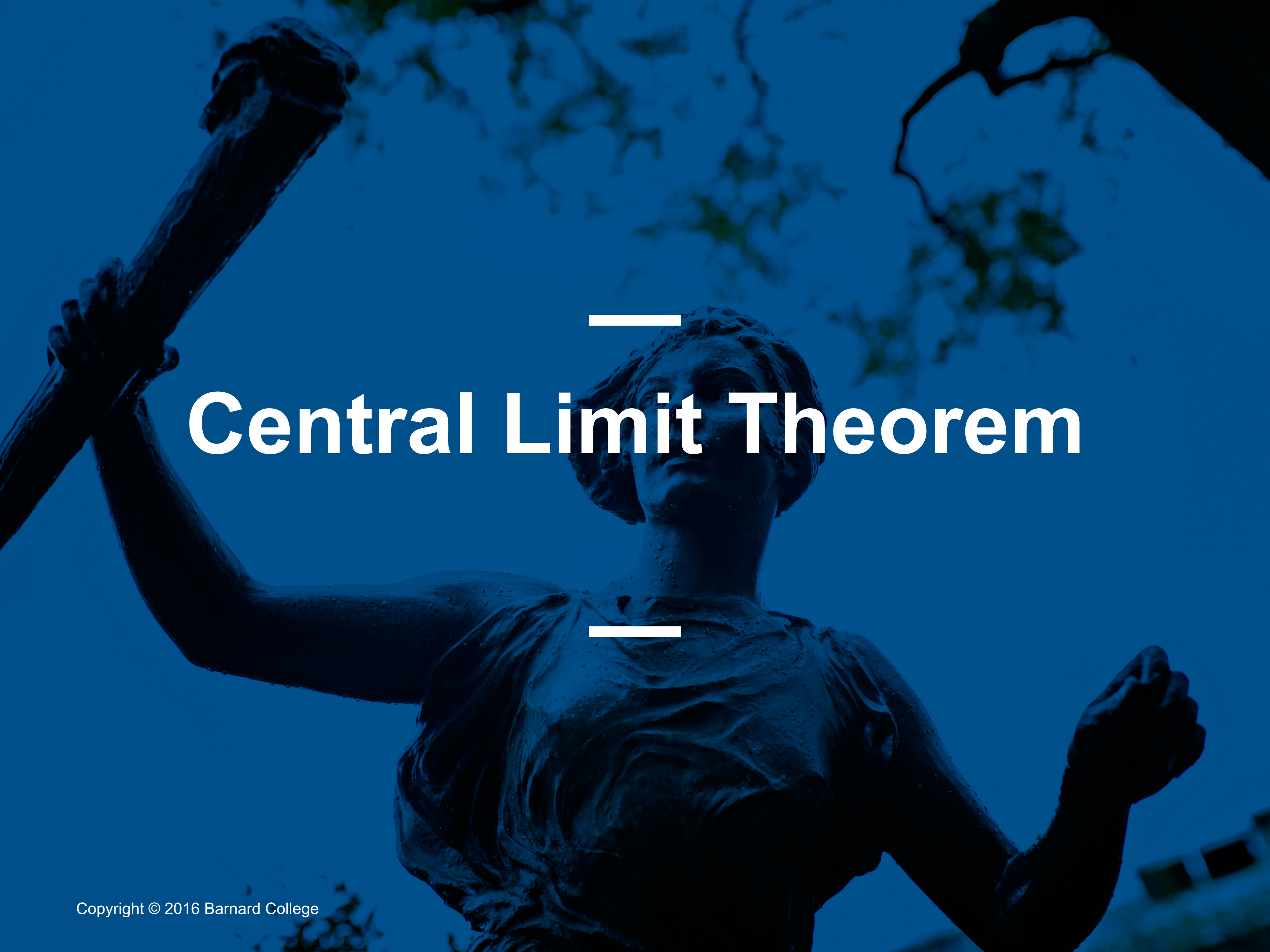
- Almost all of the data are in the range "average ± 3 SDs

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%

A “Central” Area





Central Limit Theorem

Central Limit Theorem

If the sample is

- large, and
- drawn at random with replacement,

Then, *regardless of the distribution of the population,*

the probability distribution of the sample sum (or the sample average) is roughly normal

Sample Average

- We often only have a sample
- We care about sample averages because they estimate population averages.
- The Central Limit Theorem describes how the normal distribution (a bell-shaped curve) is connected to random sample averages.
- CLT allows us to make inferences based on averages of random samples

Correlation