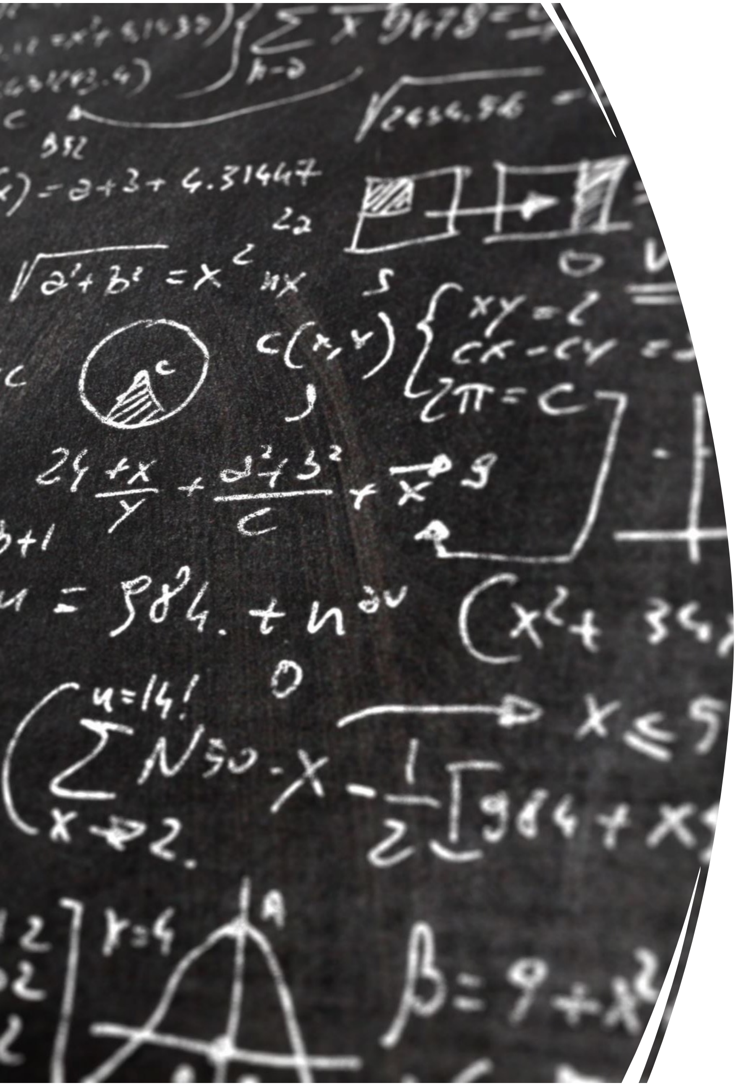


CS 383 – Computational Text Analysis

Lecture 23
Hypothesis Testing III – Quantifying Variability

Adam Poliak

04/17/2023

Announcements

Today's lecture:

- https://inferentialthinking.com/
- Chapter 9.4 – 14 (inclusive)

Final Project

- Powerpoint template is on webpage

Course evaluations

Outline

- Causality
- Estimation
- Boostrap
- Confidence Intervals
- Normal Distribution

Probability vs Statistics

Probability:

- Coming up with a view of the world then seeing if the data matches

Statistics:

- Creating a view of the world by looking at data

Probability vs Empirical Distribution

"Probability Distribution":

- All the possible values of a quantity
- The probability of each of the values

"Empirical" – based on observations

"Empirical Distribution":

- All observed values
  - The proportion of times each value appears

Steps in Assessing a Model

- Choose a statistic that will help you decide whether the data support the model or an alternative view of the world
- Simulate statistic under the assumptions of the model
- Draw a histogram of the simulated values
  - This is the model's prediction for how the statistic should come out
- Compute the statistic from the sample in the study
  - If the two are not consistent => evidence against the model
  - If the two are consistent => data supports the model ***so far***

Copyright © 2016 Barnard College

Null and Alternative

The method only works if we can simulate data under one of the hypotheses.

- **Null hypothesis**
  - A well defined chance model about how the data were generated
  - We can simulate data under the assumptions of this model
    - "Under the null hypothesis"
- **Alternative hypothesis:**
  - A different view about the origin of the data

Copyright © 2016 Barnard College

Prediction Under the Null Hypothesis

- Simulate the test statistic under the null hypothesis
  - Draw the histogram of simulated values
  - **The empirical distribution of the statistic under the null hypothesis**
- It is a prediction about the statistic, made by the null hypothesis
  - It shows all the likely values of the statistic
  - Also how likely they are (**if the null hypothesis is true)**
- The probabilities are approximate, because we can’t generate all the possible random samples

Copyright © 2016 Barnard College

Statistical Significance

Copyright © 2016 Barnard College

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

Copyright © 2016 Barnard College

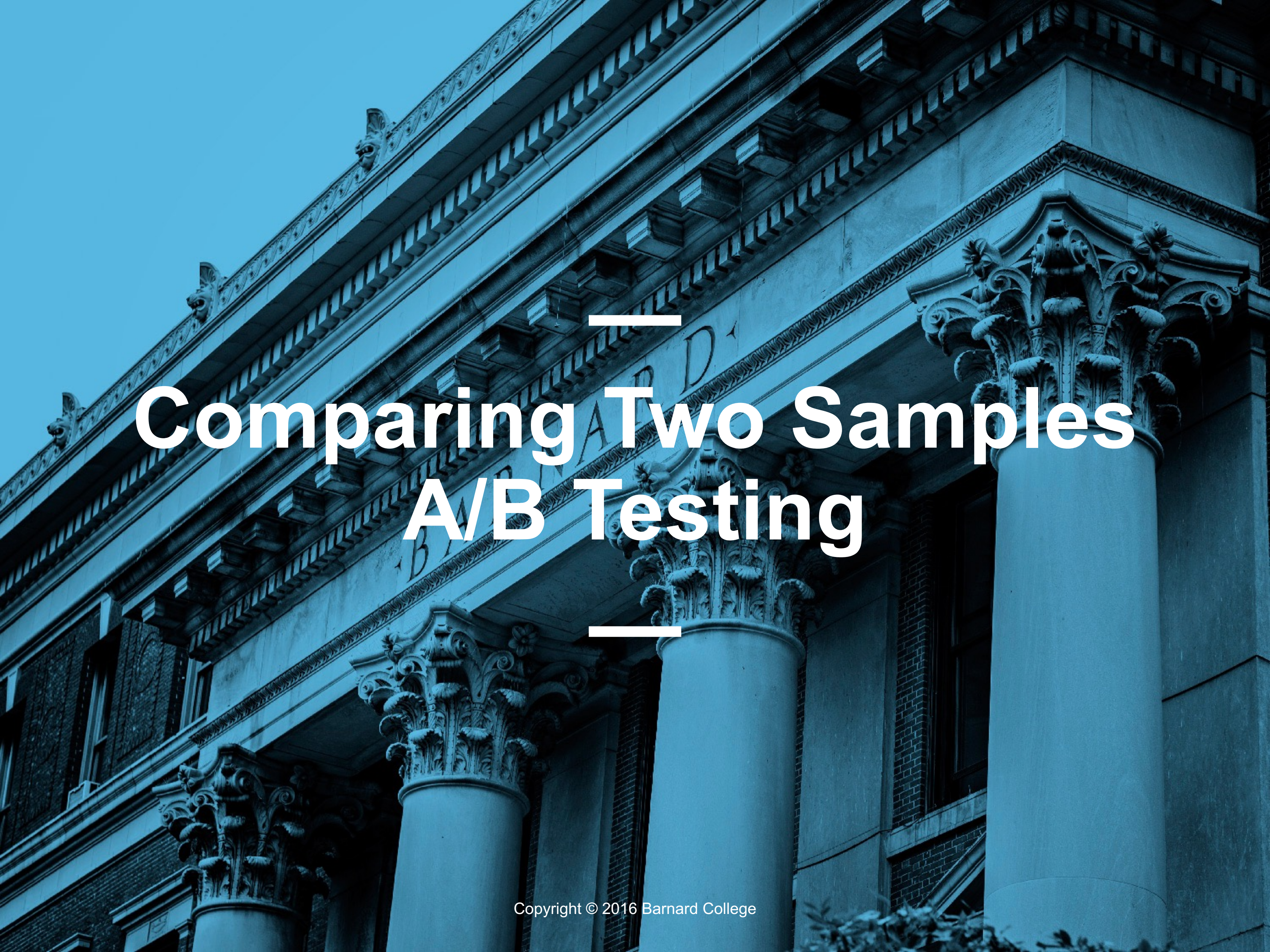
Conventions About Inconsistency

- **“Inconsistent with the null”**: The test statistic is in the tail of the empirical distribution under the null hypothesis

- **“In the tail,” first convention:**
  - The area in the tail is less than 5%
  - The result is “statistically significant”

- **”In the tail,” second convention:**
  - The area in the tail is less than 1%
  - The result is “highly statistically significant”

Copyright © 2016 Barnard College



Comparing Two Samples
A/B Testing

Copyright © 2016 Barnard College

Terminology

- Compare values of sampled *individuals* in **Group A** with values of sampled *individuals* in **Group B**.
- Question: Do the two sets of values come from the same underlying distribution?
- Answering this question by performing a statistical test is called **A/B testing**.

Copyright © 2016 Barnard College

The Groups and the Questions

- Random sample of mothers of newborns. Compare:
    A. Birth weights of babies of mothers who smoked during pregnancy
    B. Birth weights of babies of mothers who didn't smoke

- Question: Could the difference be due to chance alone?

Copyright © 2016 Barnard College

Hypotheses

Null Hypothesis:

- In the population, the distributions of the birth weights of the babies in the two groups are the same. (They are different in the sample just due to chance.)

Alternative Hypothesis:

- In the population, the babies of the mothers who smoked weigh less, on average, than the babies of the non-smokers

Copyright © 2016 Barnard College

Test Statistic

Group A: non-smokers

Group B: smokers

Statistic:

- Difference between average weights:
  - Group B average - Group A average

Negative values of this statistic favor the alternative

Simulating Under the Null

If the null is true, all rearrangements of labels are equally likely

Permutation Test:

- Shuffle all birth weights
- Assign some to Group A and the rest to Group B
    - Key: keep the sizes of Group A and Group B that same from before
- Find the difference between the two shuffled groups
- Repeat

Copyright © 2016 Barnard College

Random Permutations

- Sample randomly with replacement

- With replacement:
  - Randomly choose a value from a set, then put it back into the set
  - Can result in duplicates

Copyright © 2016 Barnard College

A-B Testing for CTA

Difference in stress before vs during COVID

Observed Statistic:

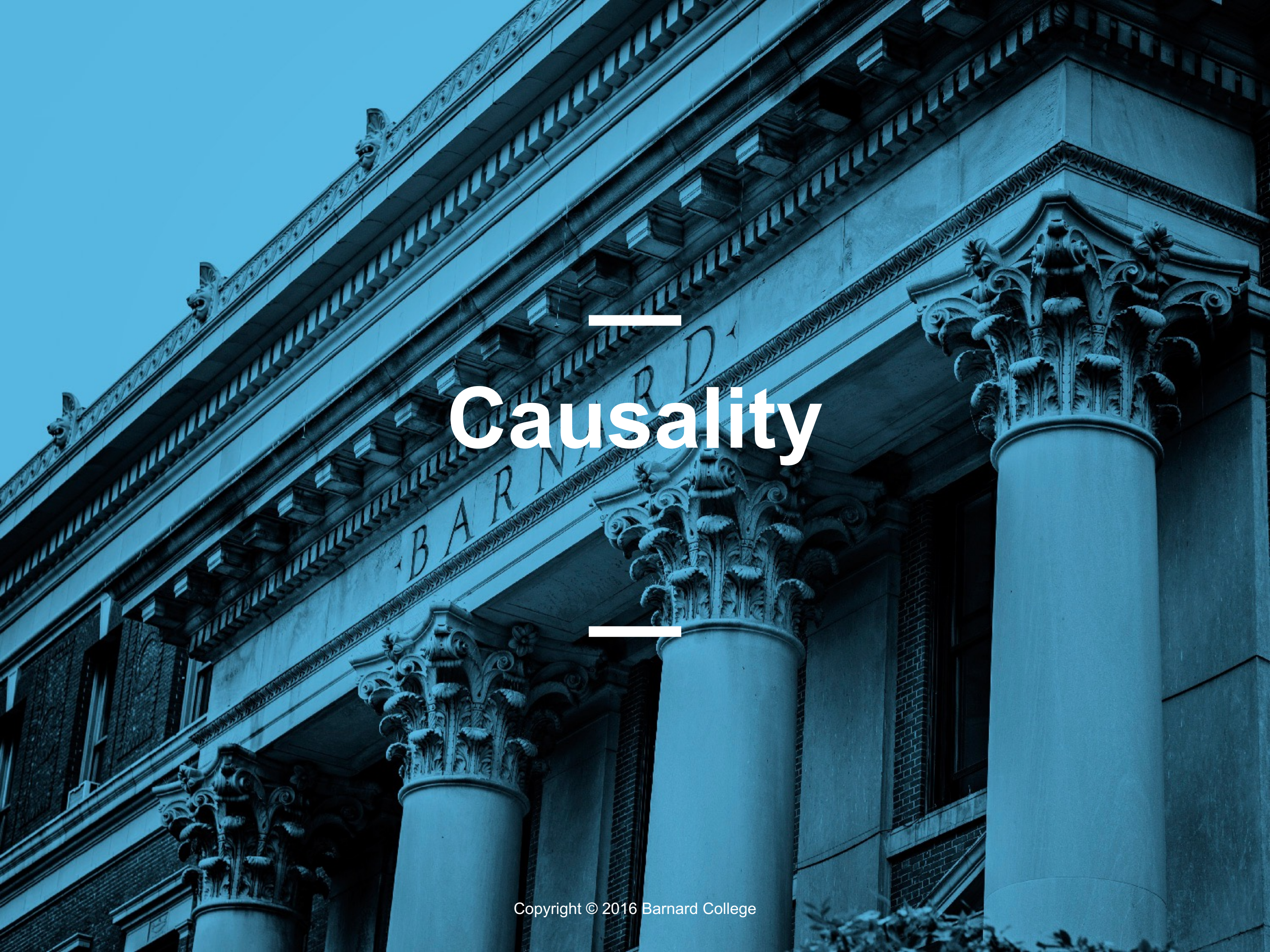
- Difference in avg LIWC score in *n* posts before COVID vs *m* posts during from a similar subreddit

Empirical distribution:

- Randomly assign *n* posts to before and *m* posts to during
- Compute difference between the two new groups

P-value

- Percent of simulated statistic that was like, or more extreme than observed statistic



Causality

Copyright © 2016 Barnard College

Randomized Controlled Experiment

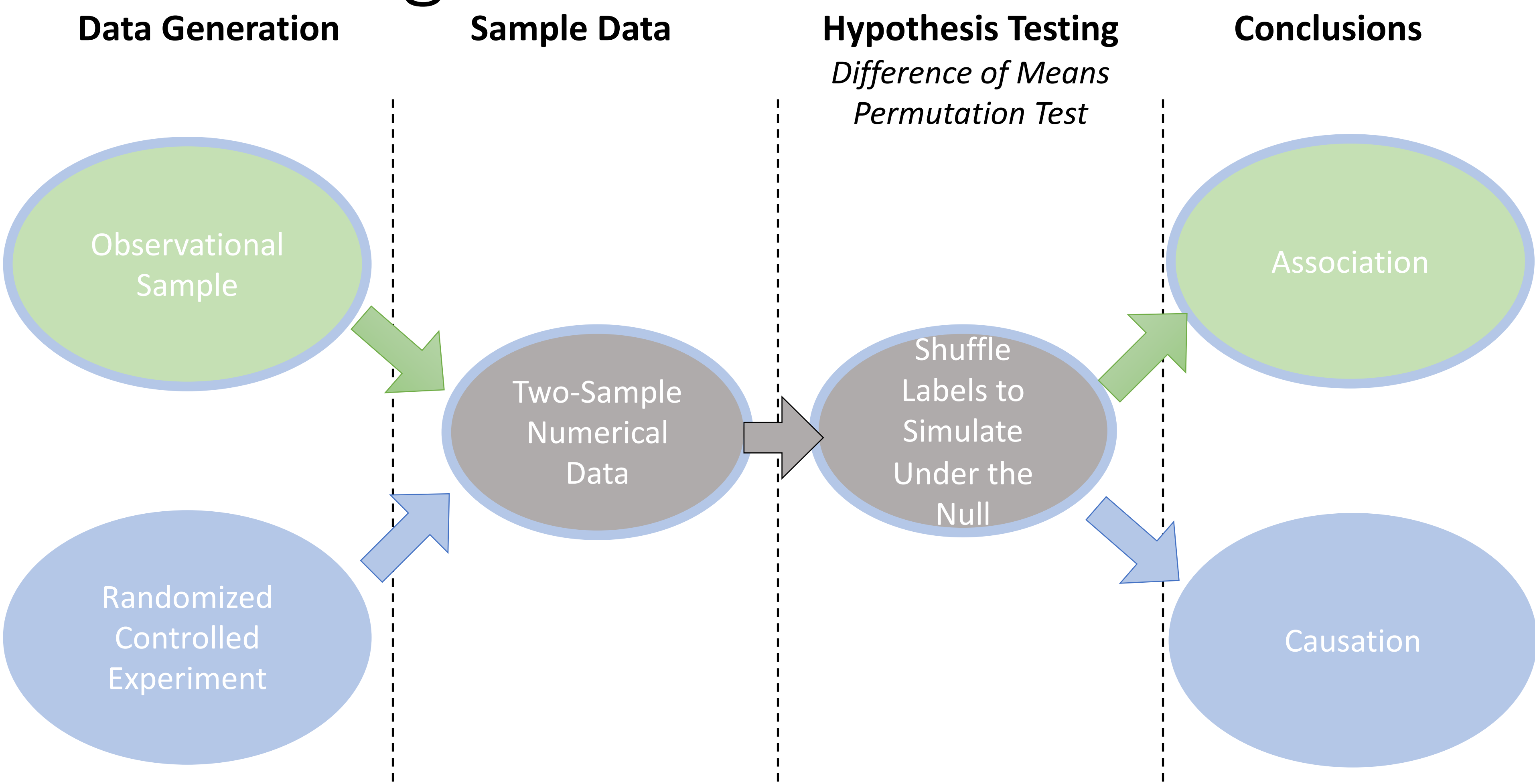
- Sample A: **control group**
- Sample B: **treatment group**

- **if the treatment and control groups are selected at random, then you can make causal conclusions.**

- Any difference in outcomes between the two groups could be due to
  - chance
  - the treatment

Copyright © 2016 Barnard College

Randomized Assignment & Shuffling



Copyright © 2016 Barnard College

Estimation

Copyright © 2016 Barnard College

Inference: Estimation

- How do we calculate the value of an unknown parameter?
- If you have a census (that is, the whole population):
  - Just calculate the parameter and you’re done
- If you don’t have a census:
  - Take a random sample from the population
  - Use a statistic as an **estimate** of the parameter

Copyright © 2016 Barnard College

Estimation Variability

Copyright © 2016 Barnard College

Variability of the Estimate

- One sample → One estimate

- But the random sample could have come out differently

- And so the estimate could have been different

- Big question:
  - How different would it be if we estimated again?

Copyright © 2016 Barnard College

Quantifying Uncertainty

- The estimate is usually not exactly right.
- Variability of the estimate tells us something about how accurate the estimate is:

Estimate = Parameter + Error

- How accurate is the estimate, usually?
- How big is a typical error?
- When we have a census, we can do this by simulation

Copyright © 2016 Barnard College

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?

Copyright © 2016 Barnard College

The Bootstrap

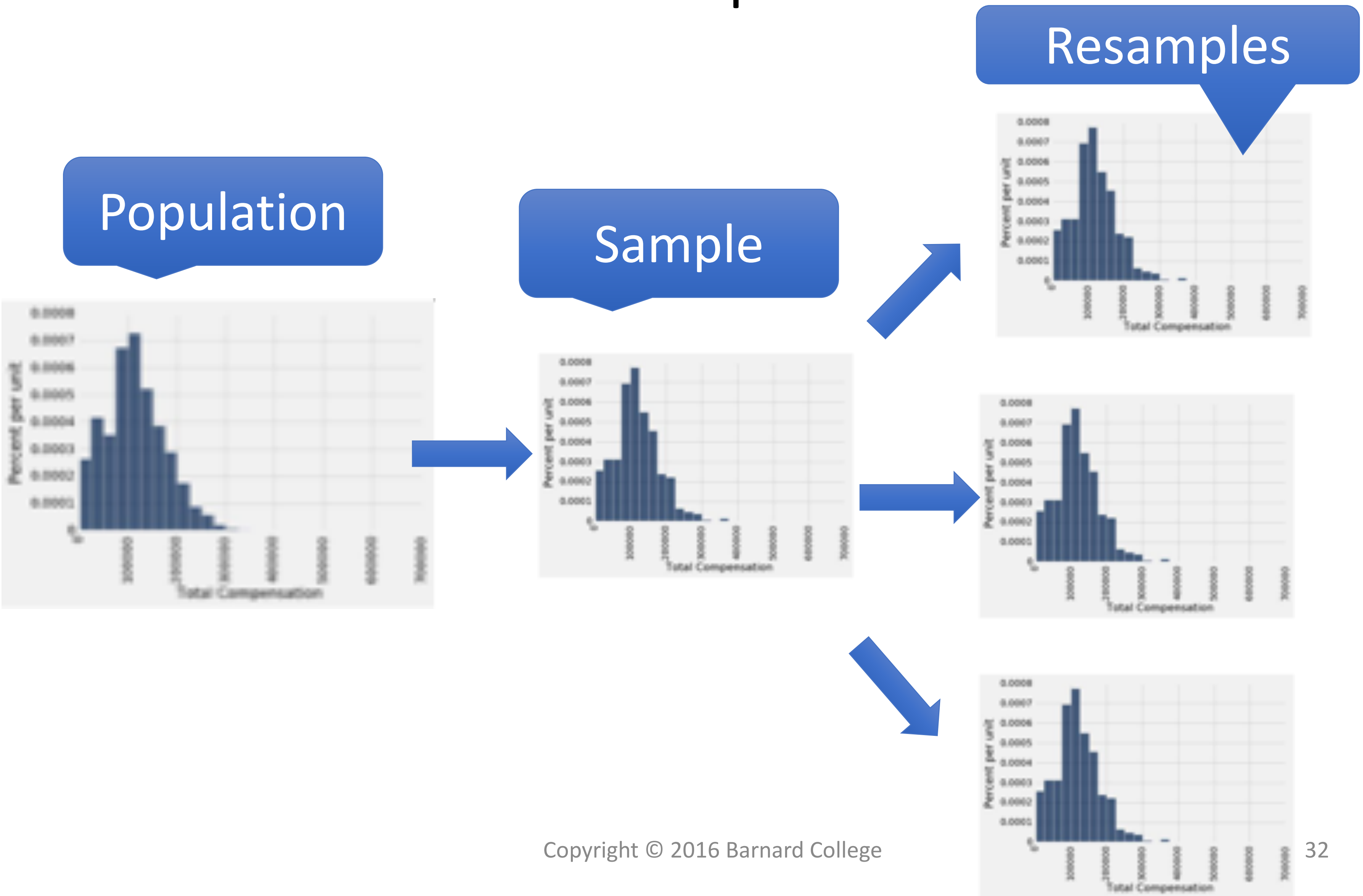
Copyright © 2016 Barnard College

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!

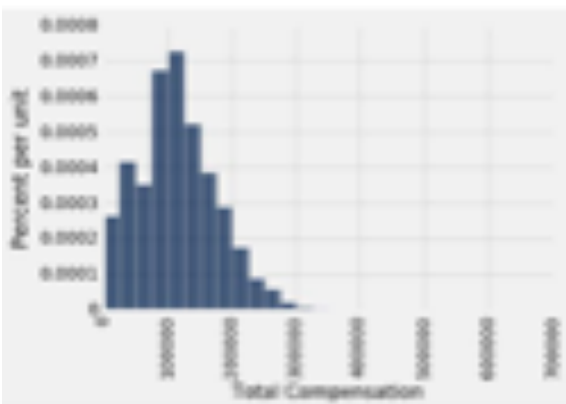
Copyright © 2016 Barnard College

How the Bootstrap works



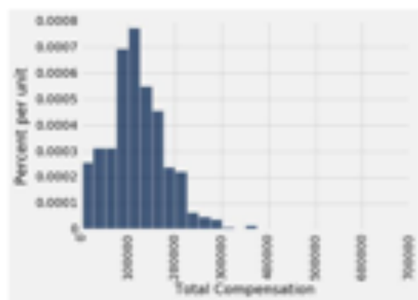
Why the Bootstrap works

Population



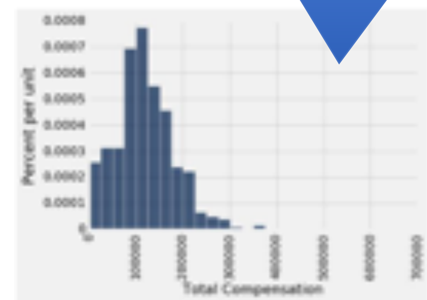
What we wish we could get

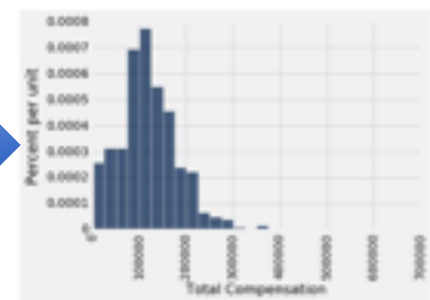
Sample

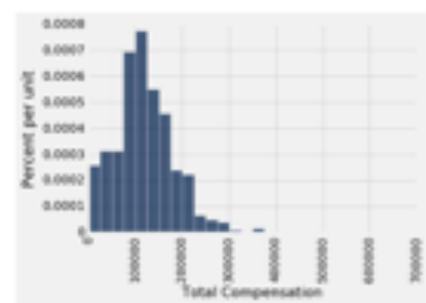


What we actually can get

Resamples







Copyright © 2016 Barnard College

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

Copyright © 2016 Barnard College

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

Copyright © 2016 Barnard College

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

Copyright © 2016 Barnard College

Variability

Our results might be different based on the original sample

How can we quantify this variability?

Confidence Intervals

Copyright © 2016 Barnard College

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

Copyright © 2016 Barnard College

Use Methods Appropriately

Copyright © 2016 Barnard College

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.

Copyright © 2016 Barnard College

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

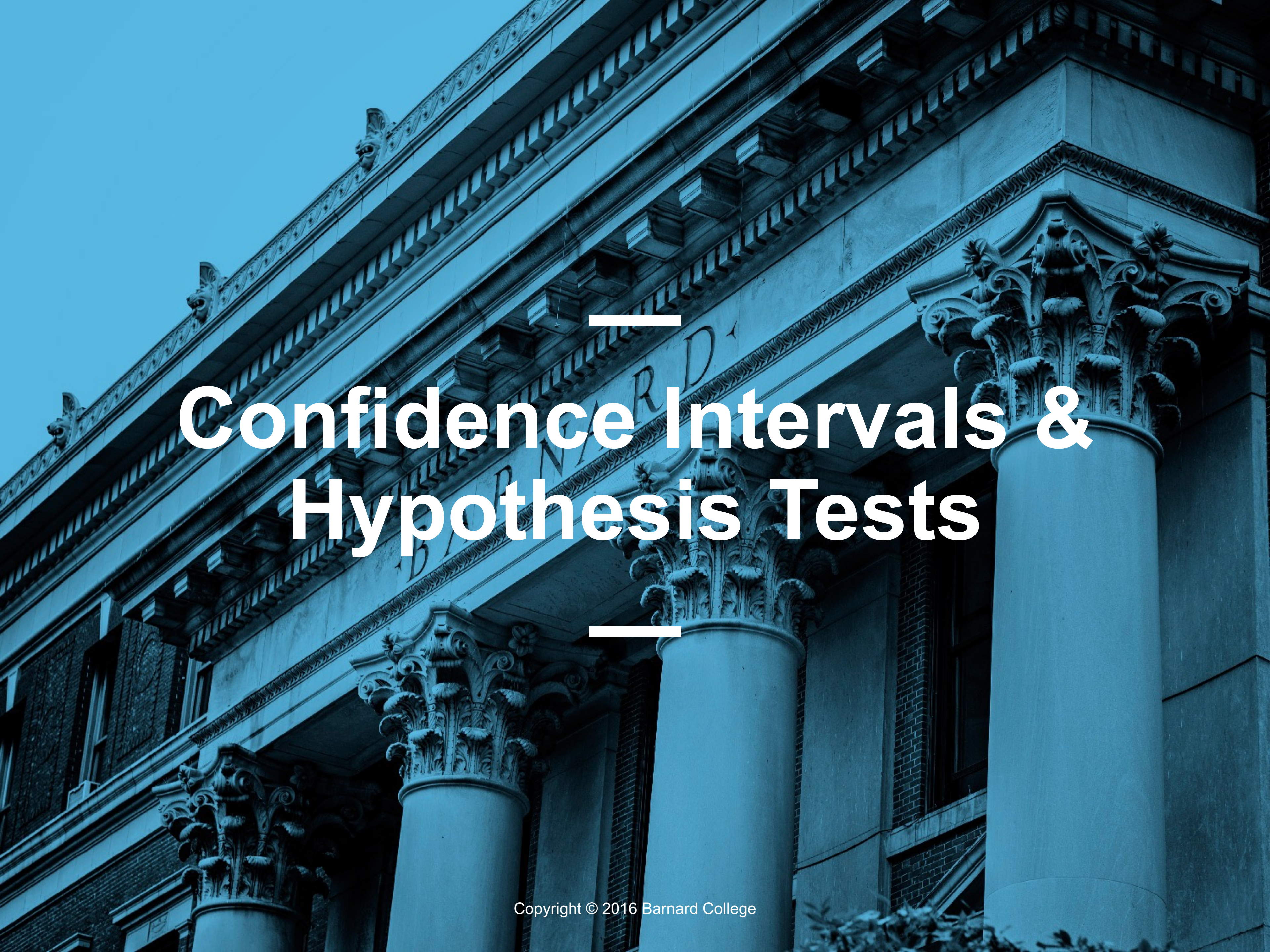
- if you're trying to estimate very high or very low percentiles, or min and max
- If you're trying to estimate any parameter that's greatly affected by rare elements of the population
- If the probability distribution of your statistic is not roughly bell shaped (the shape of the empirical distribution will be a clue)
- If the original sample is very small

Copyright © 2016 Barnard College

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

Copyright © 2016 Barnard College



Confidence Intervals & Hypothesis Tests

Copyright © 2016 Barnard College

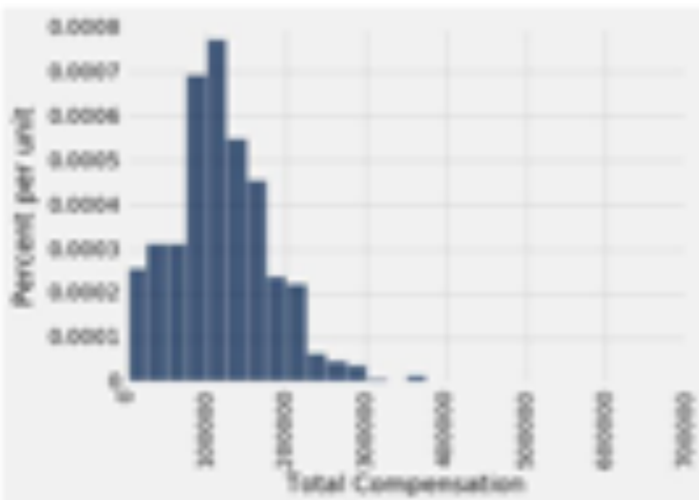
Using a CI for Testing

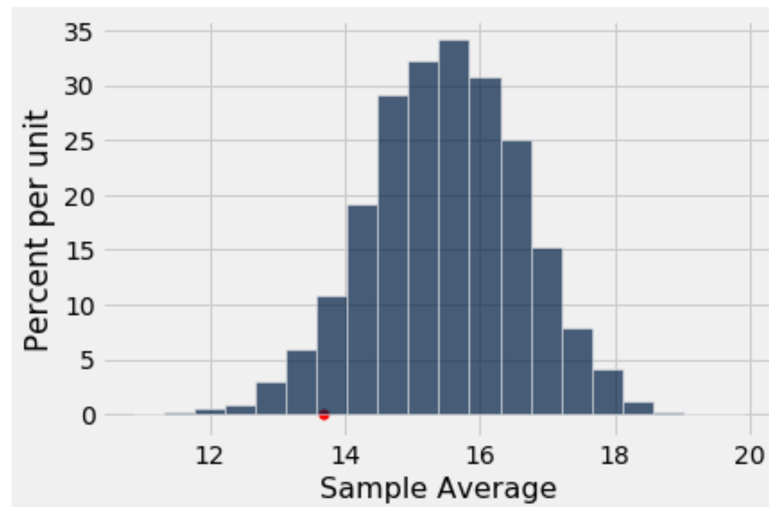
- Null hypothesis: **Population average = *x***
- Alternative hypothesis: **Population average ≠ *x***
- Cutoff for P-value: *p*%
- Method:
  - Construct a (100-*p*)% confidence interval for the population average
  - If *x* is not in the interval, reject the null
  - If *x* is in the interval, can't reject the null

Copyright © 2016 Barnard College

Empirical Distribution

When we simulate the statistic under the null hypothesis, we often see a distribution like:





Why?

Center Limit Theorem

Center & Spread

Copyright © 2016 Barnard College

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?

Copyright © 2016 Barnard College

Average and the Histogram

Copyright © 2016 Barnard College

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

Copyright © 2016 Barnard College

Relation to the histogram

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

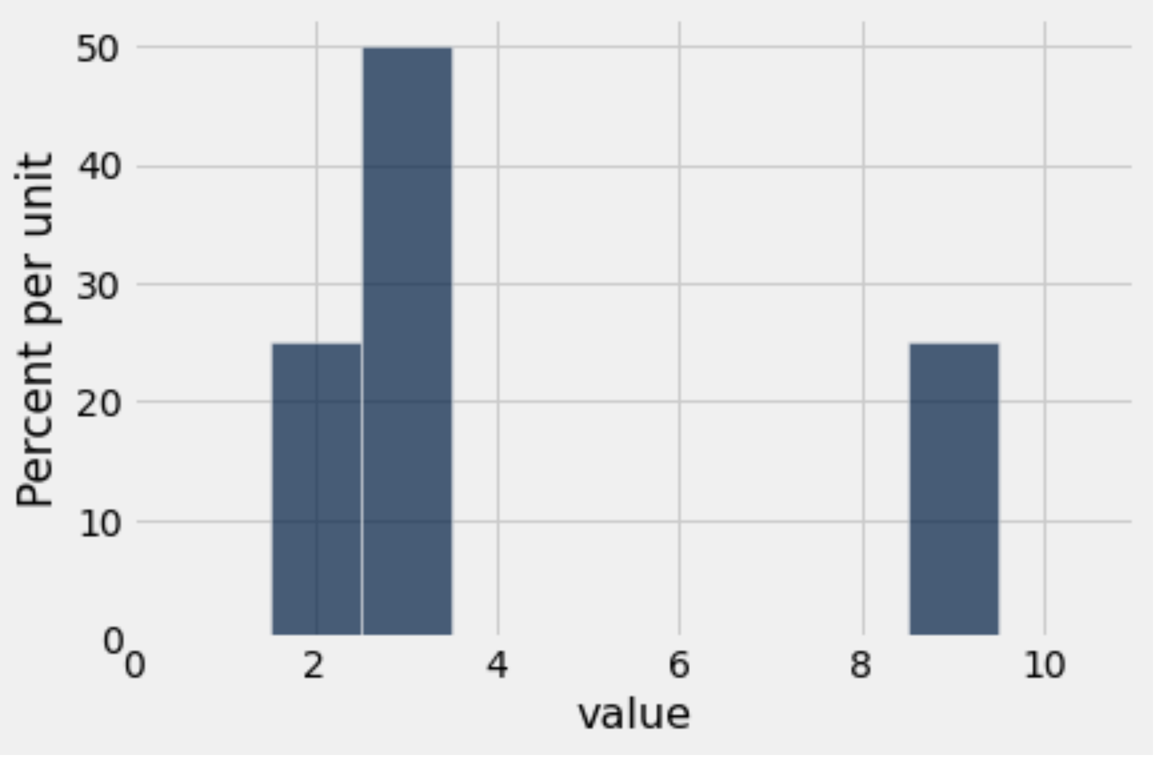
- The average is the **center of gravity** of the histogram

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

Copyright © 2016 Barnard College

Average as balance point

- Average is 4.25

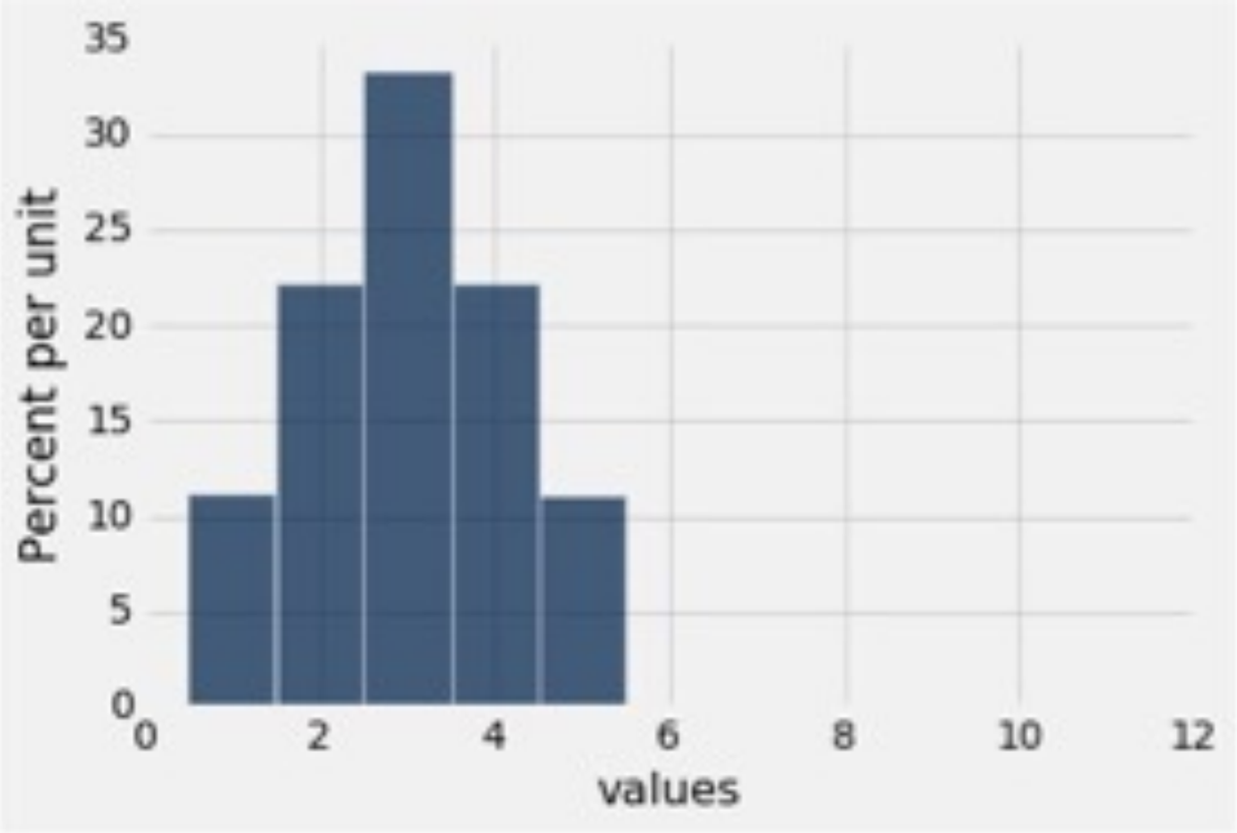


Average and Median

Copyright © 2016 Barnard College

Question

- What list produces this histogram?



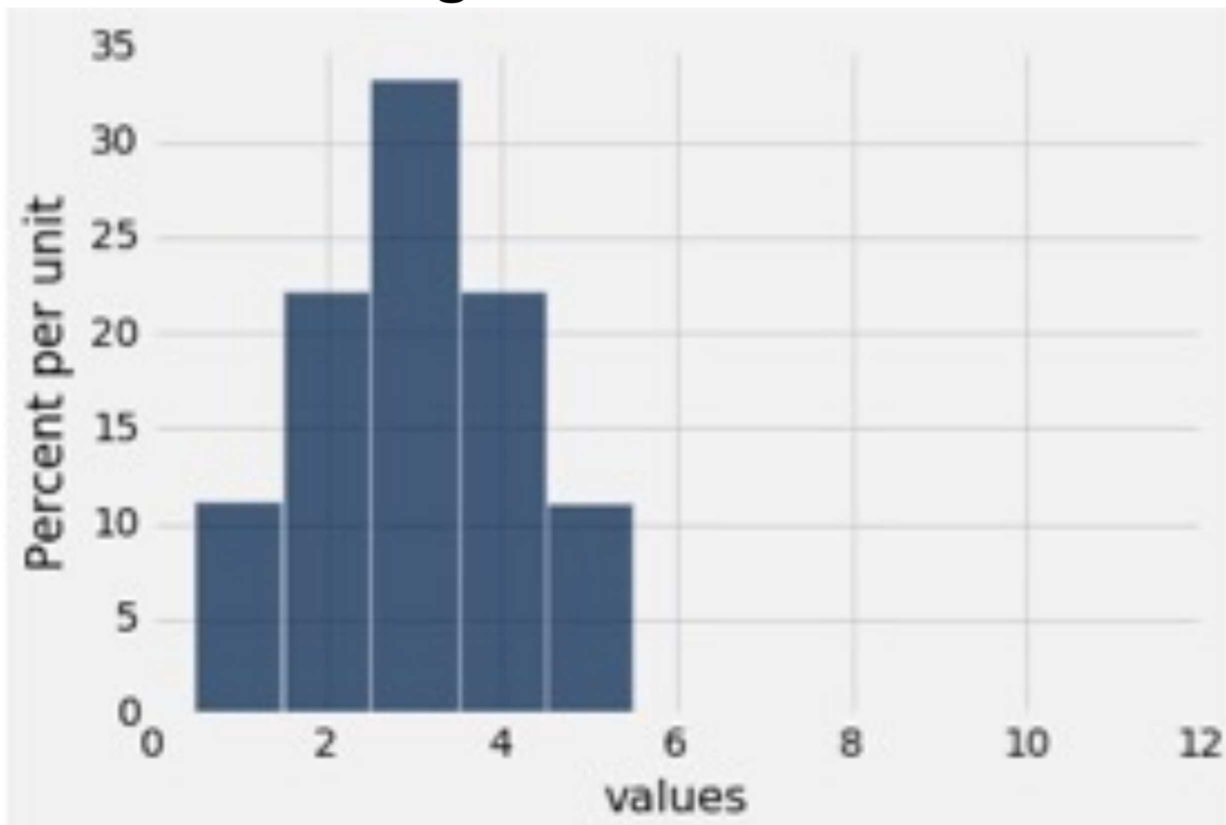
Copyright © 2016 Barnard College

Question

- What list produces this histogram?

1, 2, 2, 3, 3

3, 4, 4, 5



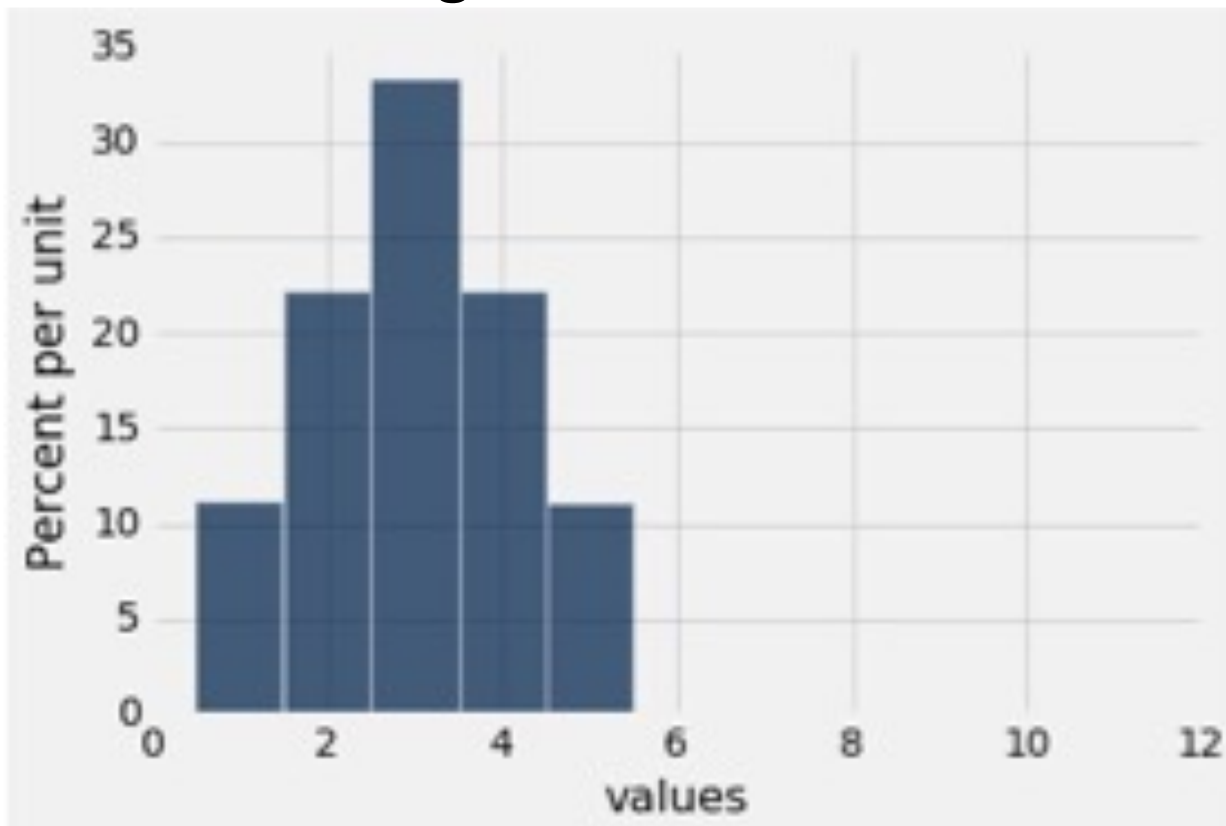
Copyright © 2016 Barnard College

Question

- What list produces this histogram?

1, 2, 2, 3, 3

3, 4, 4, 5



- Average?

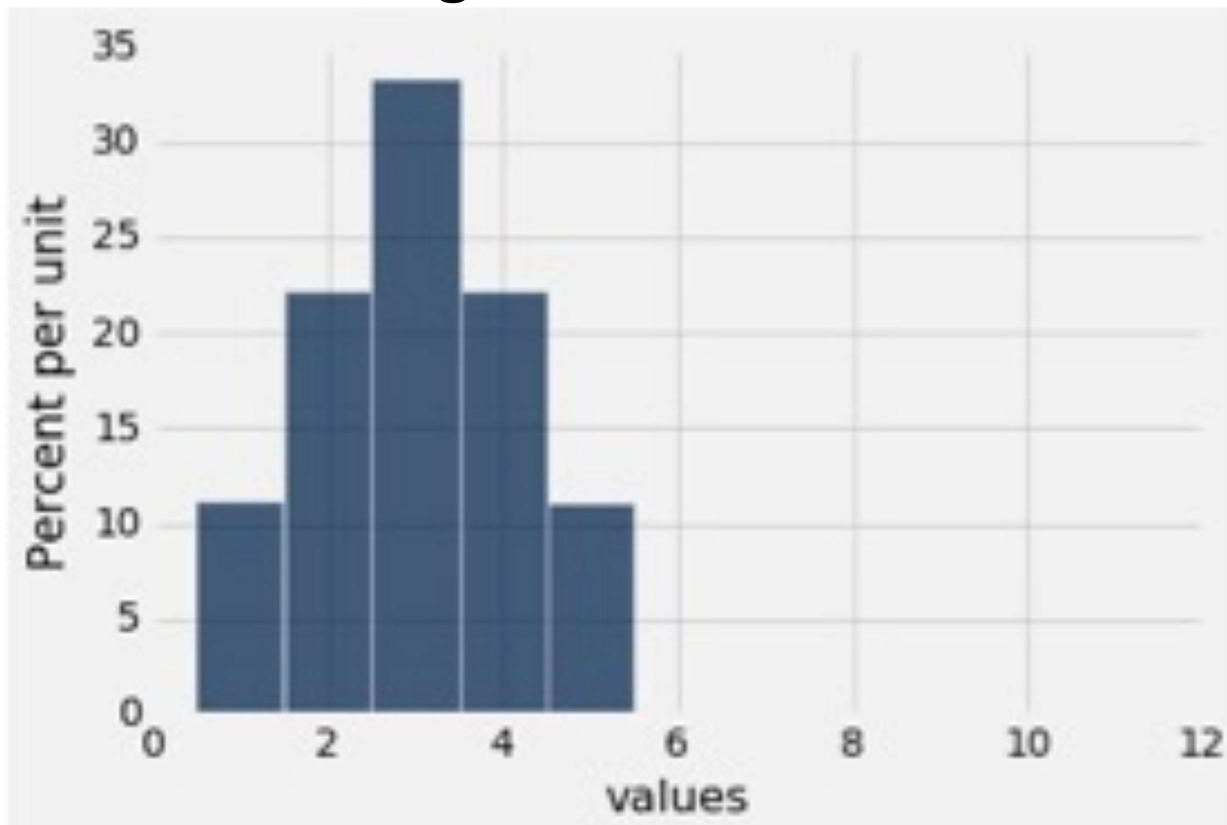
Copyright © 2016 Barnard College

Question

- What list produces this histogram?

1, 2, 2, 3, 3

3, 4, 4, 5



- Average?
  - 3

Copyright © 2016 Barnard College

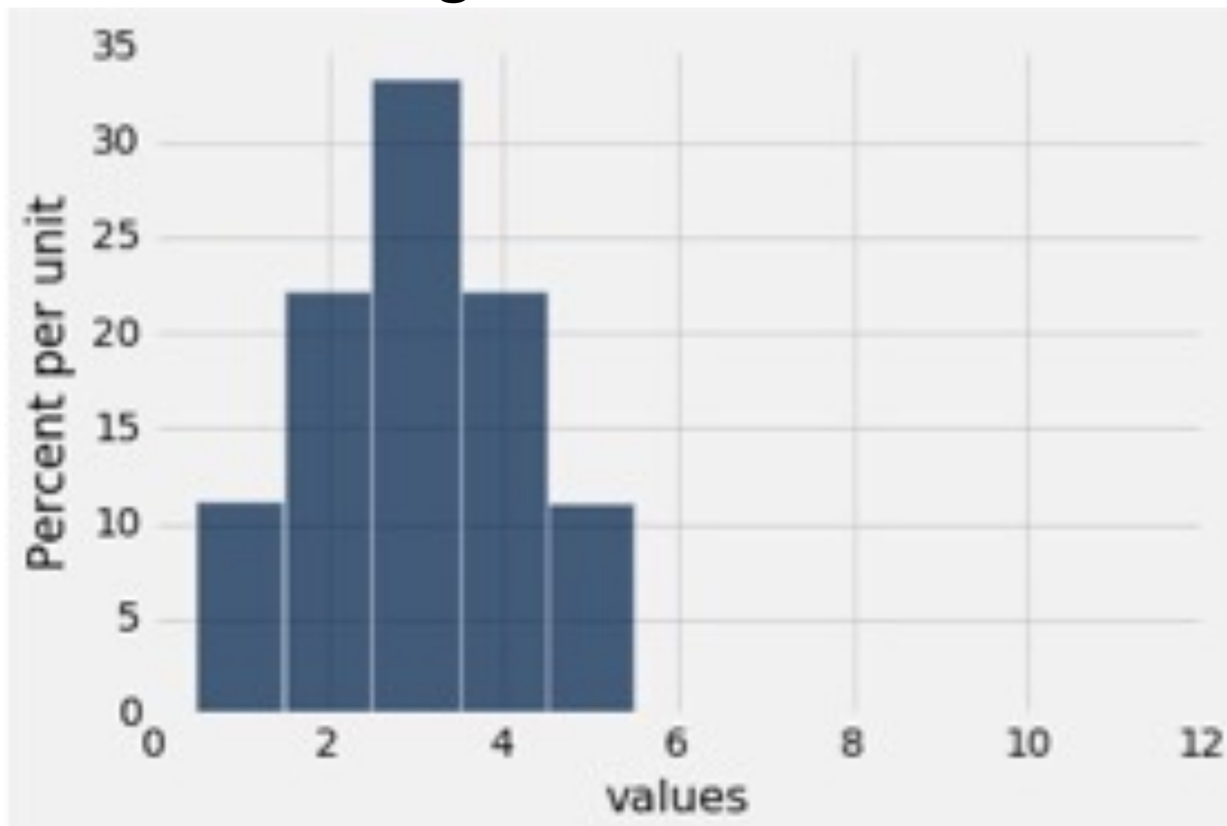
Question

- What list produces this histogram?

1, 2, 2, 3, 3

3, 4, 4, 5

- Average?
  - 3
- Median?



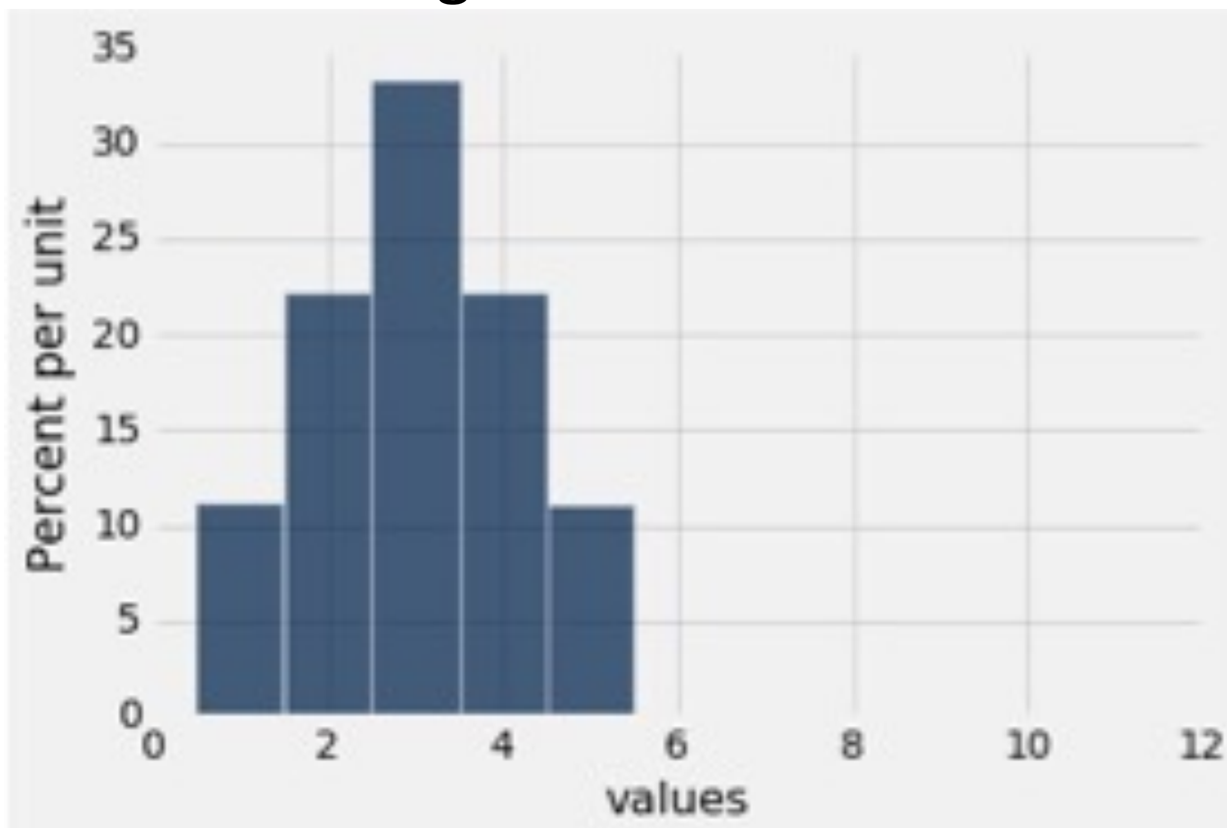
Copyright © 2016 Barnard College

Question

- What list produces this histogram?

1, 2, 2, 3, 3

3, 4, 4, 5

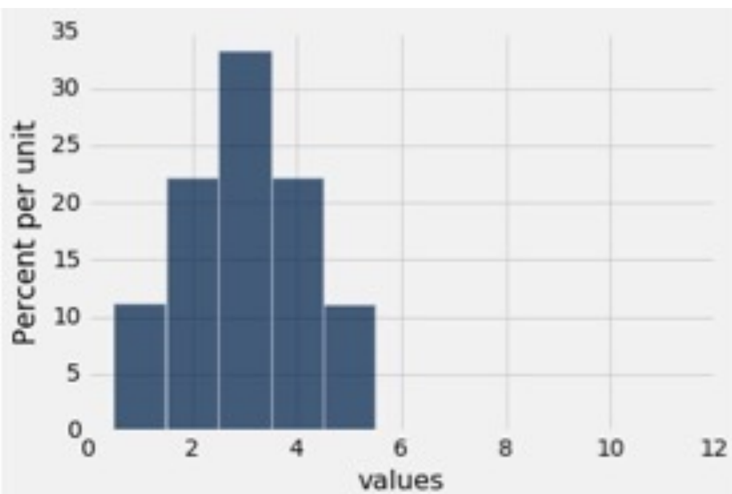


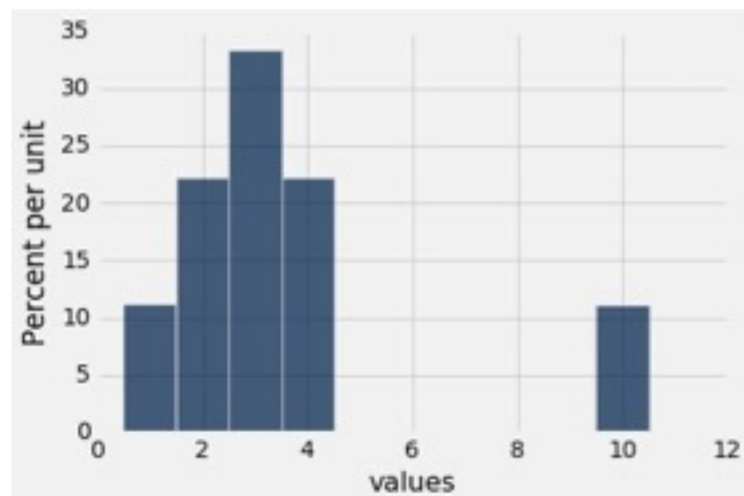
- Average?
  - 3
- Median?
  - 3

Copyright © 2016 Barnard College

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





Copyright © 2016 Barnard College

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

Copyright © 2016 Barnard College

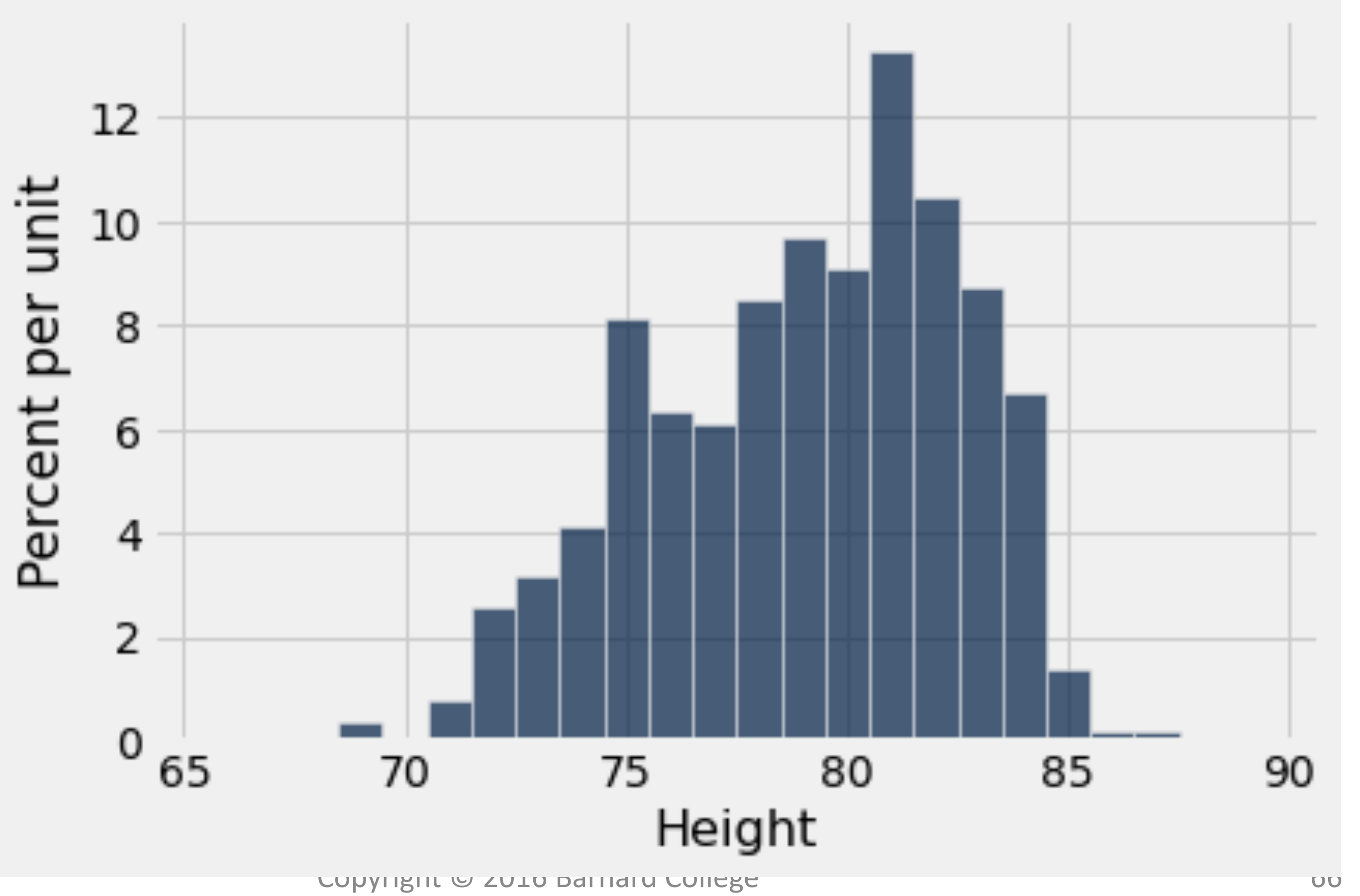
Comparing Mean and Median

- **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.

Copyright © 2016 Barnard College

Question

- Which is bigger, median or mean?



Standard Deviation

Copyright © 2016 Barnard College

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

Copyright © 2016 Barnard College

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

Copyright © 2016 Barnard College

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 in the lecture

Copyright © 2016 Barnard College

Chebyshev’s Inequality

Copyright © 2016 Barnard 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/z2

Copyright © 2016 Barnard College

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%)

Copyright © 2016 Barnard College

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...%)

Copyright © 2016 Barnard College

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%)

Copyright © 2016 Barnard College

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%)
average ± 5 SDs	at least 1 - 1/25 (96%)

True no matter what the distribution looks like

Copyright © 2016 Barnard College

Understanding HW 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

Copyright © 2016 Barnard College

Standard Units

Copyright © 2016 Barnard 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

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

Copyright © 2016 Barnard College

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

Copyright © 2016 Barnard College

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

Copyright © 2016 Barnard College

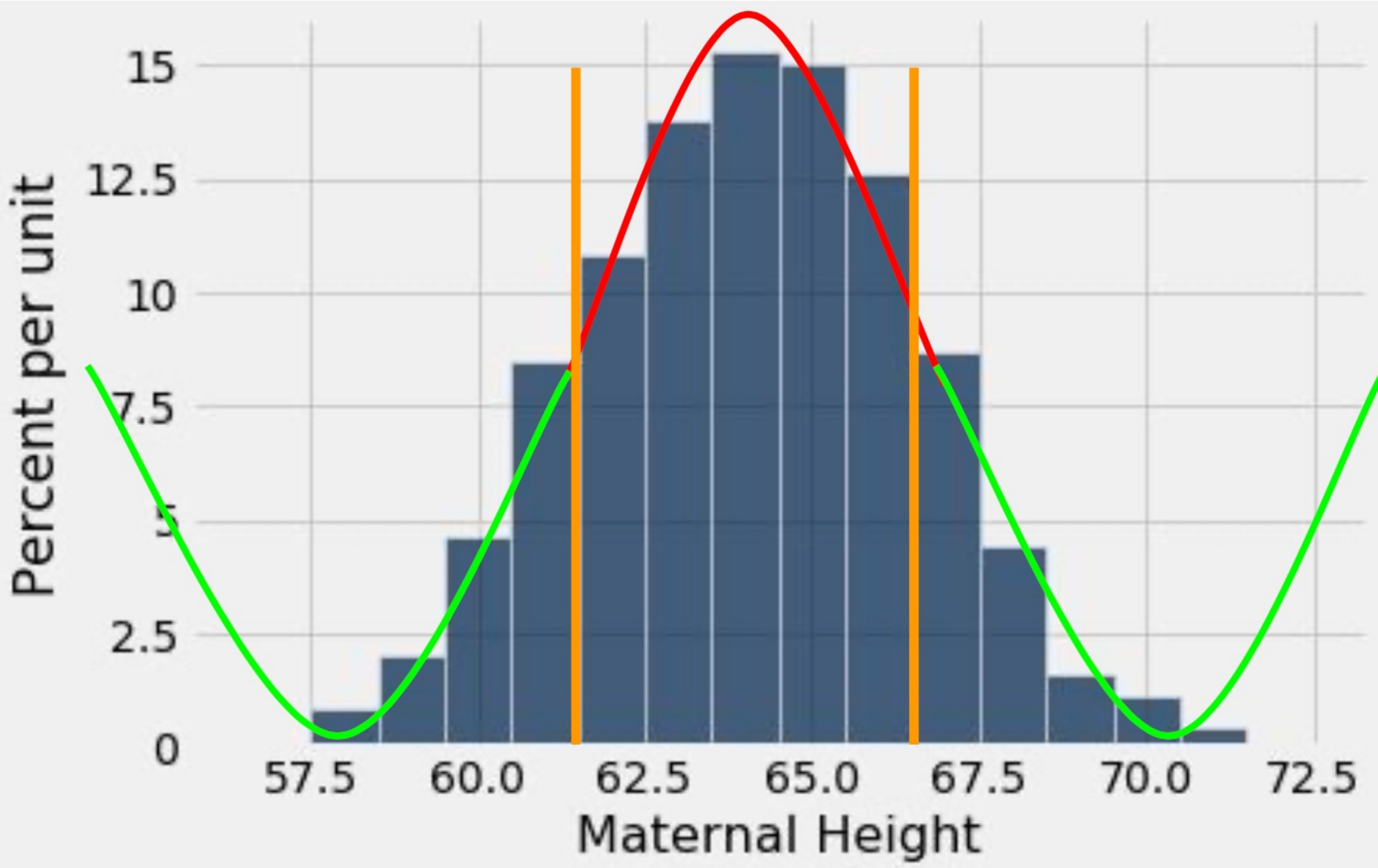
The SD and Bell Shaped Curves

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

Copyright © 2016 Barnard College

Points of Inflection



Normal Distribution

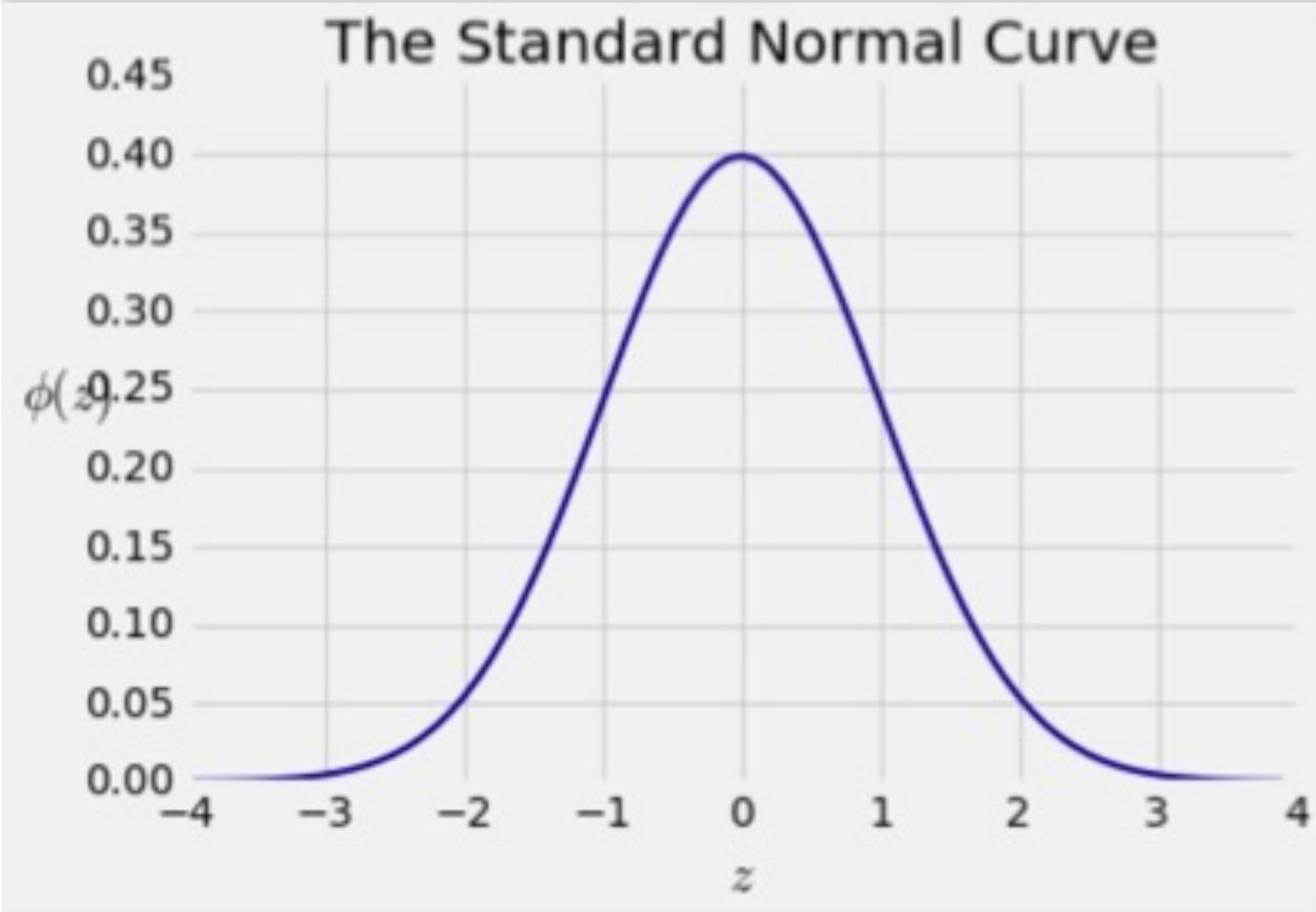
Copyright © 2016 Barnard College

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

Equation for the normal curve

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

Bell Curve



Copyright © 2016 Barnard College

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

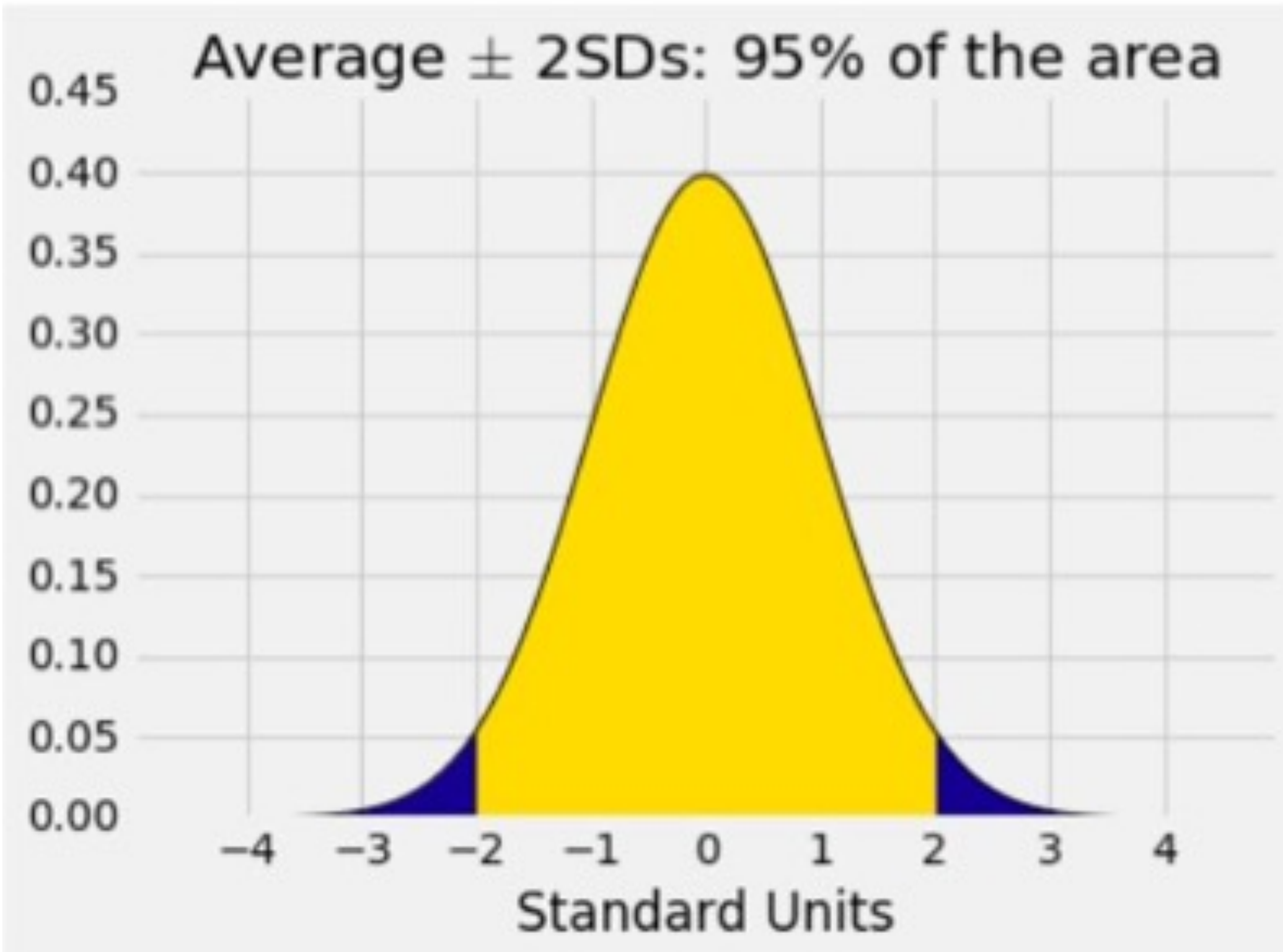
- Almost all of the data are in the range “average ± 3 SDs

Bounds and Approximations

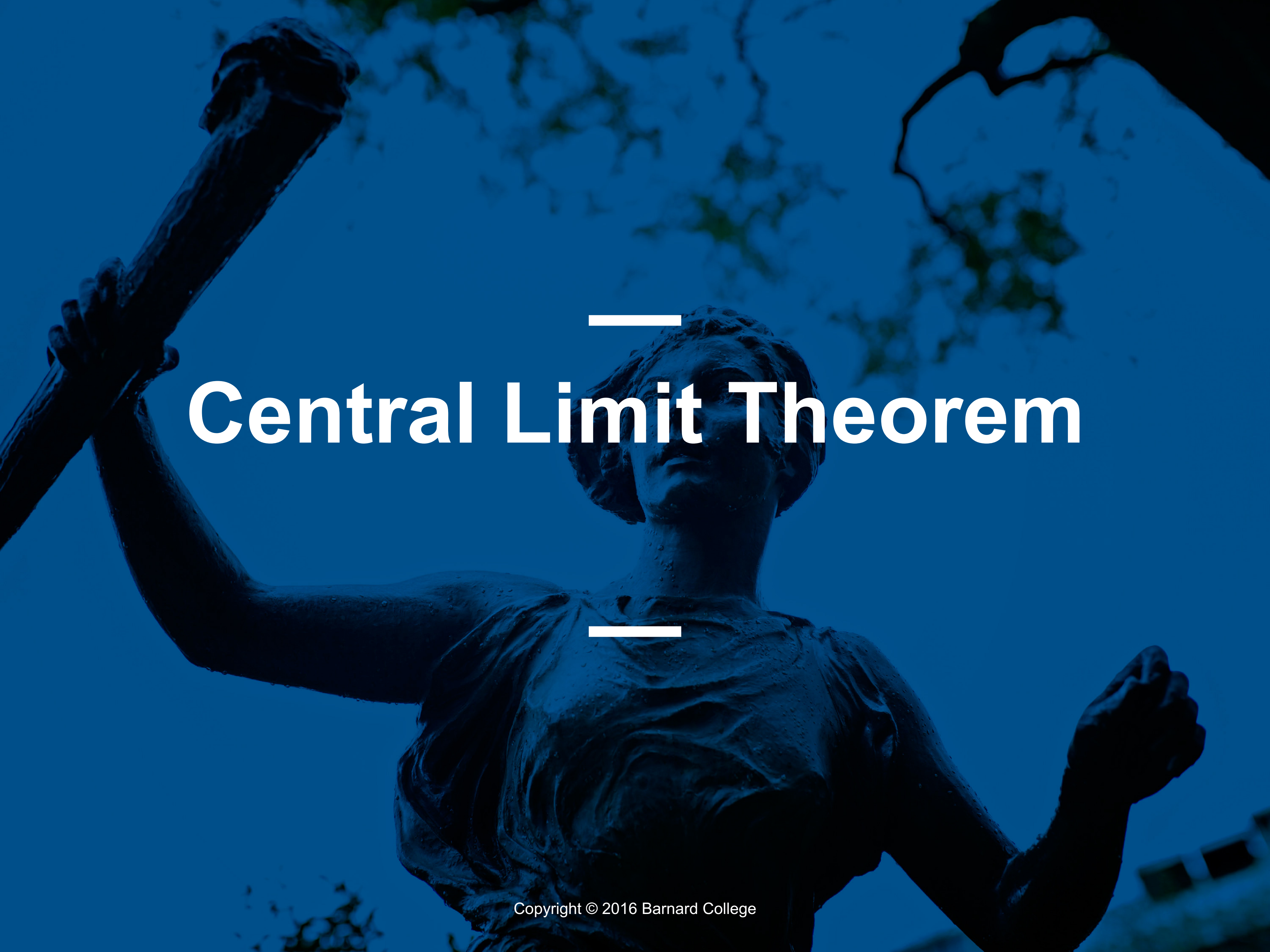
Percent in Range	All Distributions	Normal Distributions
Average +- 1 SD	At least 0%	About 68%
Average +- 2 SDs	At least 75%	About 95%
Average +- 3 SDs	At least 88.888…%	About 99.73%

Copyright © 2016 Barnard College

A "Central" Area



Copyright © 2016 Barnard College



Central Limit Theorem

Copyright © 2016 Barnard College

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

Copyright © 2016 Barnard College

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

Copyright © 2016 Barnard College