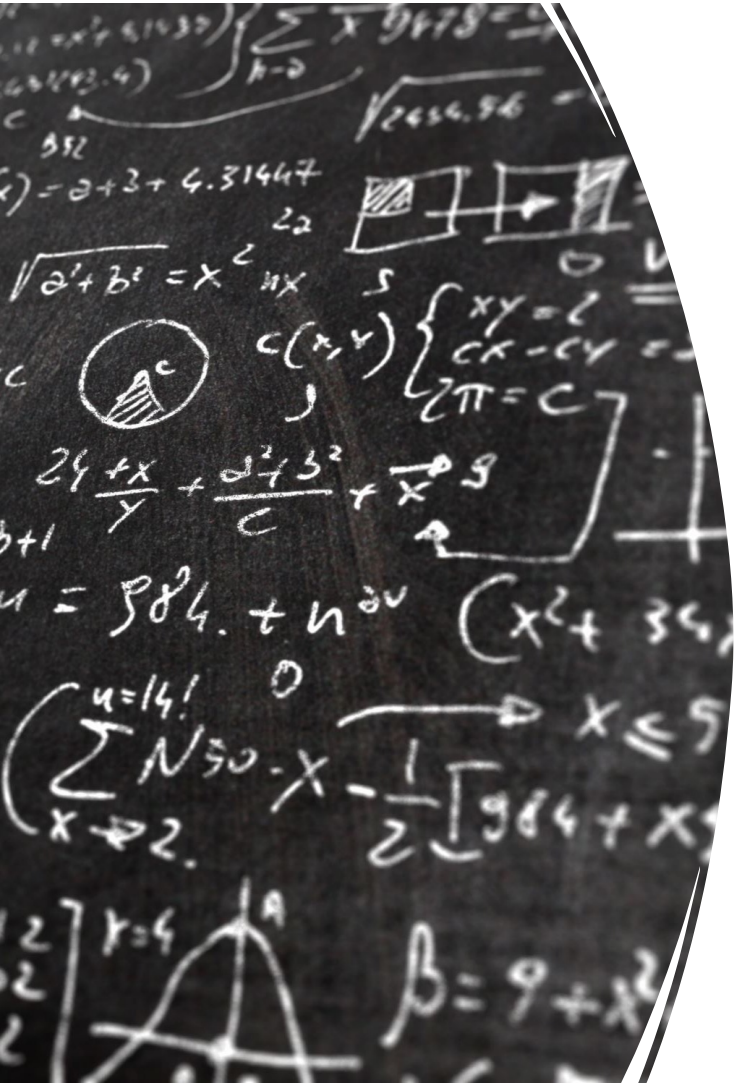


CS 383 – Computational Text Analysis

Lecture 24
Hypothesis Testing IV – Central Limit Theorem Bootstrapping for CTA

Adam Poliak

04/19/2023

Announcements

Today's lecture:

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

Final Project

- Powerpoint template is on webpage

Course evaluations

What do you see as the major strengths of Adam Poliak in this course? What areas do you see for improvement in instruction and/or in content?

How prepared were you to take this course? What courses, if any, would you have found useful to take before this course? Is this course listed at the appropriate level?

How did Adam Poliak effectively create an accessible and inclusive course experience? What areas do you see for commendation and/or improvement in the instructor's attention to accessibility and inclusivity?

Would you recommend this course, as taught by Adam Poliak, to other students? Why or why not?

Enrollments	Responded	Response Rate
16	0	0%
16	1	6.25%

gaut
@0xgaut

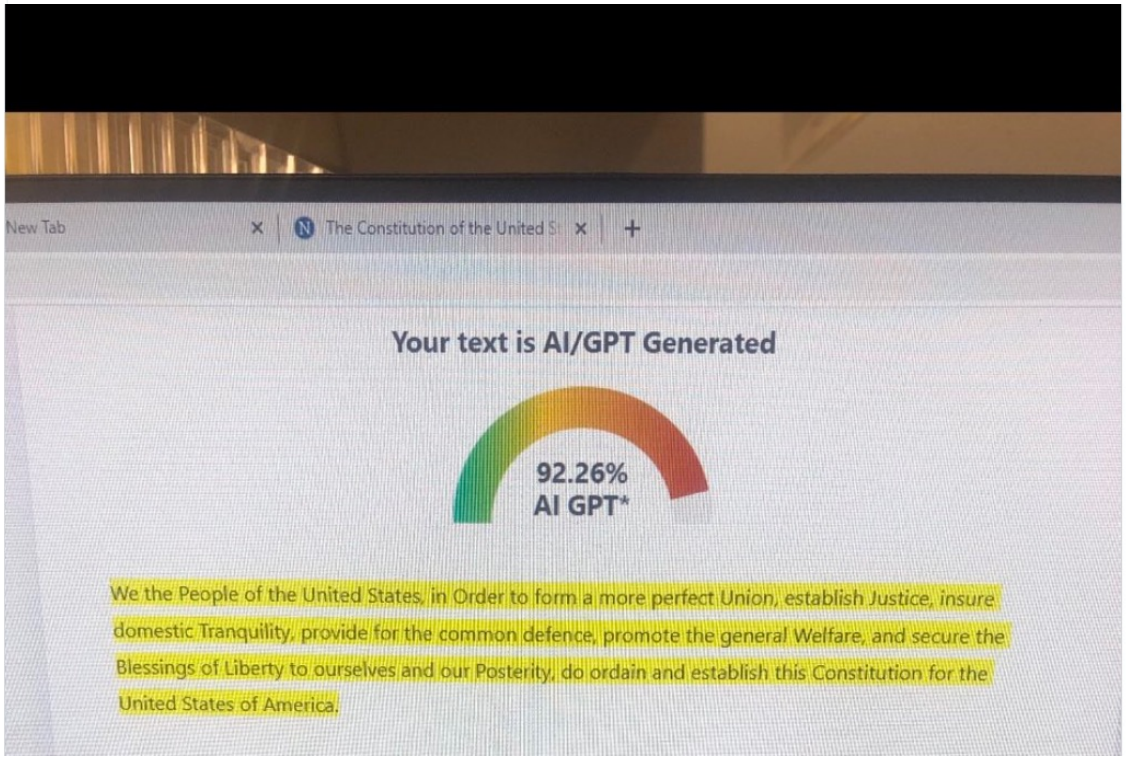
someone used an AI detector on the US Constitution and the results are concerning.

Explain this, OpenAI!

[redacted]

1:52 PM · Apr 18, 2023 · **2.9M** Views

2,054 Retweets **387** Quotes **28.9K** Likes **2,107** Bookmarks



Why might this be the case?

gaut
@0xgaut

someone used an AI detector on the US Constitution and the results are concerning.

Explain this, OpenAI!

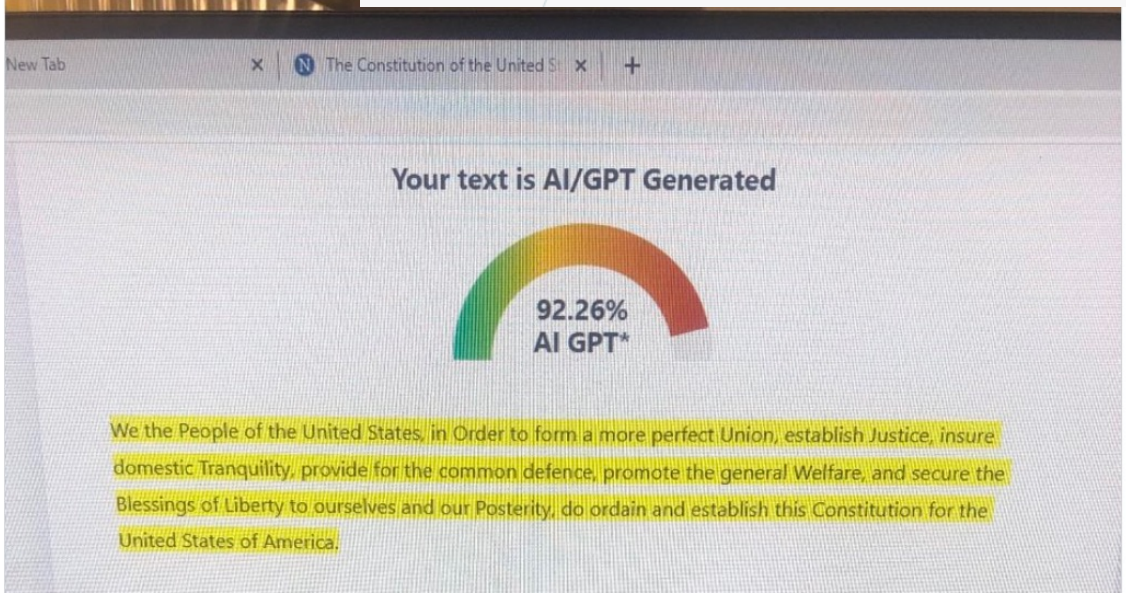
[redacted]

1:52 PM · Apr 18, 2023 · **2.9M** Views

2,054 Retweets **3**

David Mimno @dmimno · 48m

LLM detection works by looking for text that has unexpectedly high probability according to the language model. Text that occurs frequently in training data=high probability. I'd guess same for Gutenberg license text and lorem ipsum.



Outline

- Review
- Bootstrap:
  - My model is better than your model?
  - Is a covariate (feature) actually a strong indicator? (next lecture)
- Normal Distribution
- Central Limit Theorem

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

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

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
- Compute the statistic from your sample
- Simulate statistic under the assumptions of the model
  - **Empirical distribution**
- Compare observed test statistic to empirical distribution
  - If the two are not consistent => evidence against the model
  - If the two are consistent => data supports the model ***so far***

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

A-B Testing

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

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

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?

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

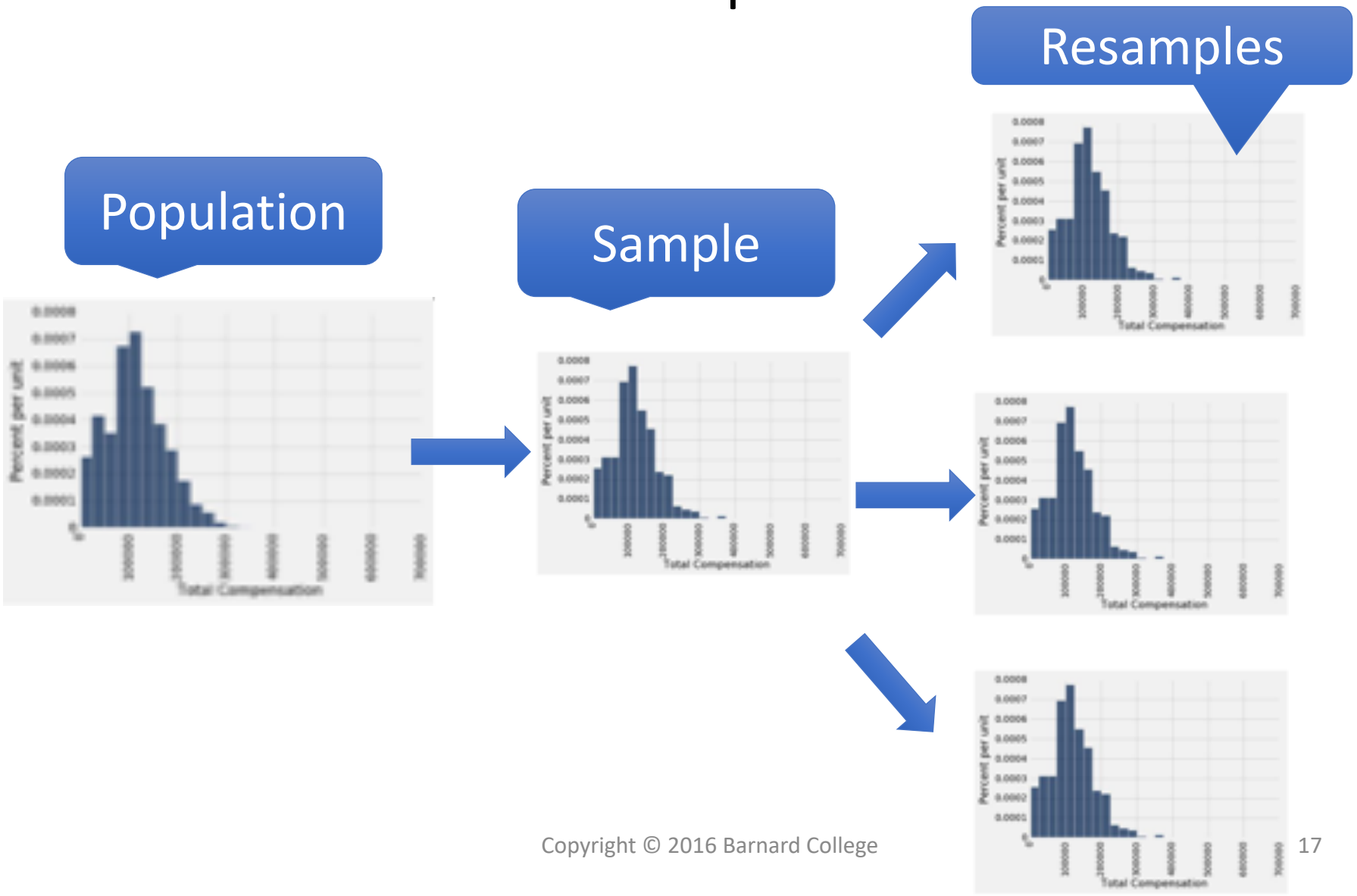
Where to Get Another Sample?

- 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?

The Bootstrap

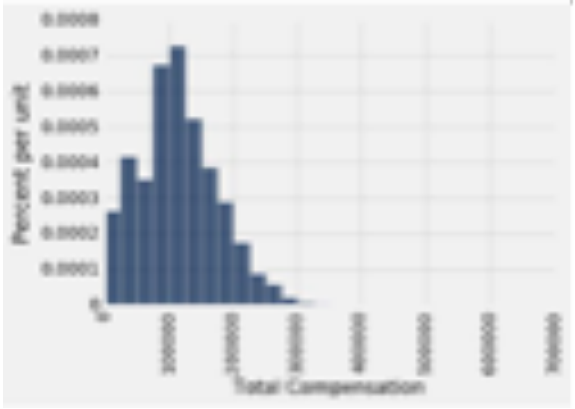
- A technique for simulating repeated random sampling
- All that we have is the original sample
  - ... which is large and random
  - Therefore, it probably resembles the population
- So we sample at random from the original sample!

How the Bootstrap works



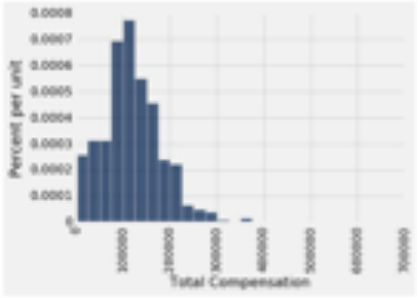
Why the Bootstrap works

Population



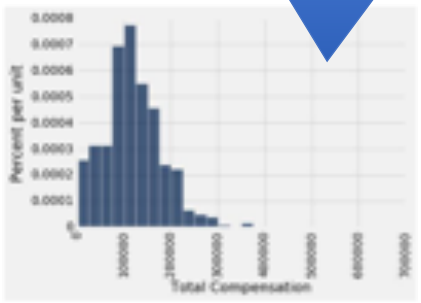
What we wish we could get

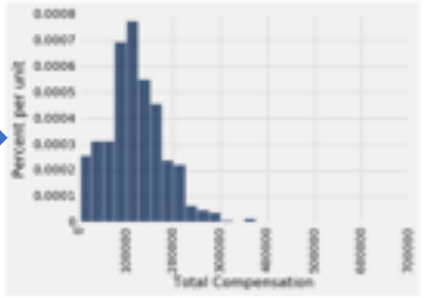
Sample

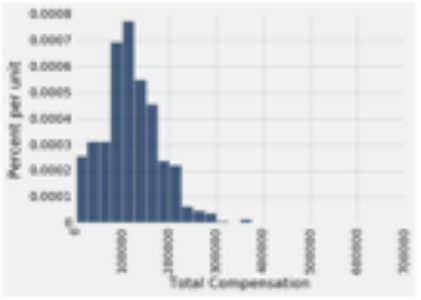


What we actually can get

Resamples







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

Variability

Our results might be different based on the original sample

How can we quantify this variability?

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

How to generate a CI

In each bootstrap sample, compute the statistic under the null

Create an empirical distribution of the simulated statistic across all bootstrapped samples

95% CI:

- lower bound: 2.5 percentile on empirical distribution
- upper bound: 97.5 percentile on empirical distribution

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 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

Outline

- Review
- **Bootstrap**:
  - My model is better than your model?
  - Is a covariate (feature) actually a strong indicator? (next lecture)
- Normal Distribution
- Central Limit Theorem

How do we know if one classifier is better than another?

Given:

- Classifier A and B
- Metric M: M(A,x) is the performance of *A* on testset *x*
- δ(x): the performance difference between A, B on x:
  - δ(x) = M(A,x) – M(B,x)

- We want to know if δ(x)>0, meaning A is better than B
- δ(x) is called the **effect size**
- Suppose we look and see that δ(x) is positive. Are we done?
- No! This might be just an accident of this one test set, or circumstance of the experiment. Instead:

Statistical Hypothesis Testing

- Consider two hypotheses:
  - Null hypothesis: A isn't better than B
  - A is better than B

$$H_0 : \delta(x) \leq 0$$
$$H_1 : \delta(x) > 0$$

- How can we rule out H_0 ?
- We create a random variable X ranging over test sets
- And ask, how likely, if H_0 is true, is it that among these test sets we would see the δ(x) that we did see?
  - Formalized as the p-value:

$$P(\delta(X) \geq \delta(x) | H_0 \text{ is true})$$

Statistical Hypothesis Testing

$$P(\delta(X) \geq \delta(x) | H_0 \text{ is true})$$

- In our example, this p-value is the probability that we would see δ(*x*) assuming H_0 (null is that *A* is not better than *B)*.
  - If H_0 is true but δ(*x*) is huge, that is surprising! Very low probability!
- A very small p-value means that the difference we observed is very unlikely under the null hypothesis, and we can reject the null hypothesis
- Very small: .05 or .01
- A result(e.g., "*A* is better than *B*") is **statistically significant** if the δ we saw has a probability that is below the threshold and we therefore reject this null hypothesis.

Statistical Hypothesis Testing

How do we compute this probability?

For example, suppose we had created zillions of testsets x'.

- Now we measure the value of δ(x') on each test set
- That gives us a distribution
- Now set a threshold (say .01).
- So if we see that in 99% of the test sets δ(x) > δ(x')
  - We conclude that our original test set delta was a real delta and not an artifact.

Statistical Hypothesis Testing

- Two common approaches:
  - approximate randomization
  - bootstrap test
- Paired tests:
  - Comparing two sets of observations in which each observation in one set can be paired with an observation in another.
  - For example, when looking at systems A and B **on the same test set**, we can compare the performance of system *A* and *B* on each same observation x_i

Bootstrap example

Consider a text classification example with a test set x of 10 documents, using accuracy as metric.

Suppose these are the results of systems A and B on x, with 4 outcomes (A & B both right, A & B both wrong, A right/B wrong, A wrong/B right):

either A+B both correct, or

	1	2	3	4	5	6	7	8	9	10	A%	B%	$\delta()$
x	AB	AB	AB	AB	AB	AB	AB	AB	AB	AB	.70	.50	.20

Bootstrap example

Now we create, many, say, *b*=10,000 bootstrapped test sets *x*(*i*), each of size *n* = 10.

To make each *x*(*i*), we randomly select a cell from row *x*, with replacement, 10 times:

	1	2	3	4	5	6	7	8	9	10	A%	B%	$\delta()$
x	AB	AB	AB	AB	AB	AB	AB	AB	AB	AB	.70	.50	.20
$x^{(1)}$	AB	AB	AB	AB	AB	AB	AB	AB	AB	AB	.60	.60	.00
$x^{(2)}$	AB	AB	AB	AB	AB	AB	AB	AB	AB	AB	.60	.70	-.10
...													
$x^{(b)}$													

Bootstrap example

- Now we have a distribution! We can check how often A has an **accidental** advantage, to see if the original δ(x) we saw was very common.
- Now assuming H_0, that means normally we expect δ(x')=0
- So we just count how many times the δ(x') we found exceeds the expected 0 value by δ(x) or more:

$$\text{p-value}(x) = \sum_{i=1}^{b} \mathbb{1}\left(\delta(x^{(i)}) - \delta(x) \geq 0\right)$$

Bootstrap example

Alas, it's slightly more complicated.

We didn't draw these samples from a distribution with 0 mean; we created them from the original test set *x*, which happens to be biased (by .20) in favor of *A*.

So to measure how surprising is our observed δ(*x*), we actually compute the p-value by counting how often δ(*x'*) exceeds the expected value of δ(*x*) by δ(*x*) or more:

$$\begin{aligned}\text{p-value}(x) &= \sum_{i=1}^{b} \mathbb{1}\left(\delta(x^{(i)}) - \delta(x) \geq \delta(x)\right) \\ &= \sum_{i=1}^{b} \mathbb{1}\left(\delta(x^{(i)}) \geq 2\delta(x)\right)\end{aligned}$$

Bootstrap example

We have 10,000 test sets $x(i)$ and a threshold of .01

And in only 47 of the test sets do we find that $\delta(x(i)) \geq 2\delta(x)$

The resulting p-value is .0047

This is smaller than .01, indicating $\delta(x)$ is indeed sufficiently surprising

And we reject the null hypothesis and conclude *A* is better than *B*.

Paired bootstrap example

After Berg-Kirkpatrick et al (2012)

function BOOTSTRAP(test set x, num of samples b) **returns** p-*value*(x)

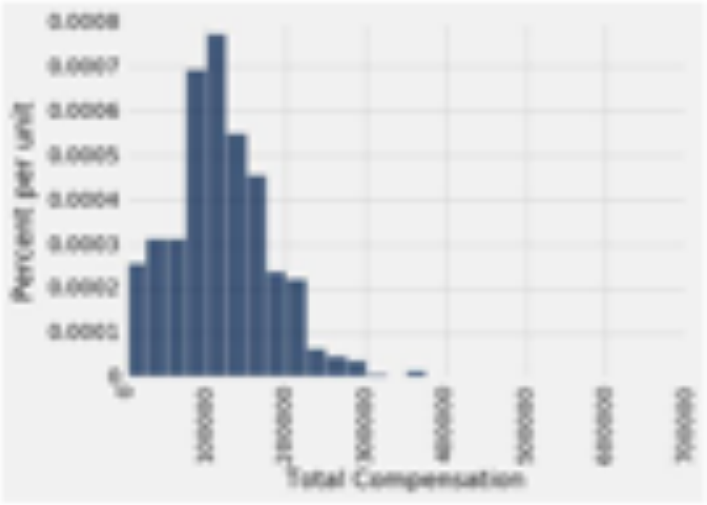
Calculate $\delta(x)$ # how much better does algorithm A do than B on x
$s = 0$
for $i = 1$ **to** b **do**
    for $j = 1$ **to** n **do** # Draw a bootstrap sample $x^{(i)}$ of size n
        Select a member of x at random and add it to $x^{(i)}$
    Calculate $\delta(x^{(i)})$ # how much better does algorithm A do than B on $x^{(i)}$
    $s \leftarrow s + 1$ **if** $\delta(x^{(i)}) > 2\delta(x)$
p-value(x) $\approx \frac{s}{b}$ # on what % of the b samples did algorithm A beat expectations?
return p-value(x) # if very few did, our observed δ is probably not accidental

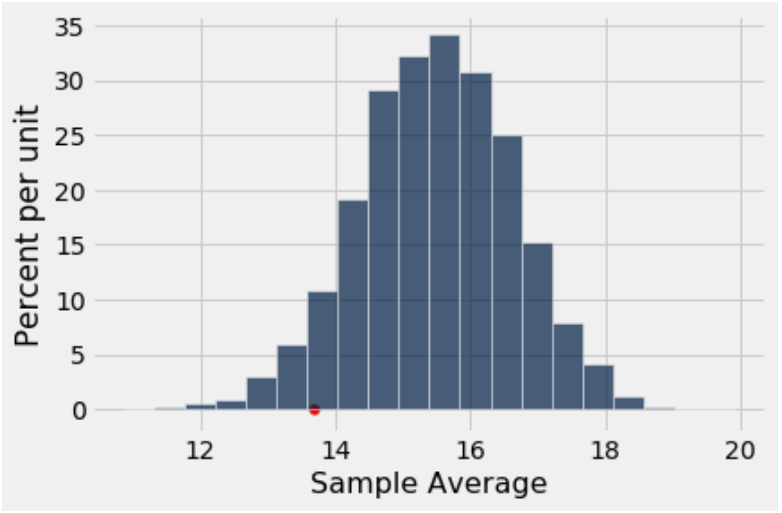
Outline

- Review
- Bootstrap:
  - My model is better than your model?
  - Is a covariate (feature) actually a strong indicator? (next lecture)
- **Normal Distribution**
- Central Limit Theorem

Empirical Distribution

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





Why?

Center Limit Theorem

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

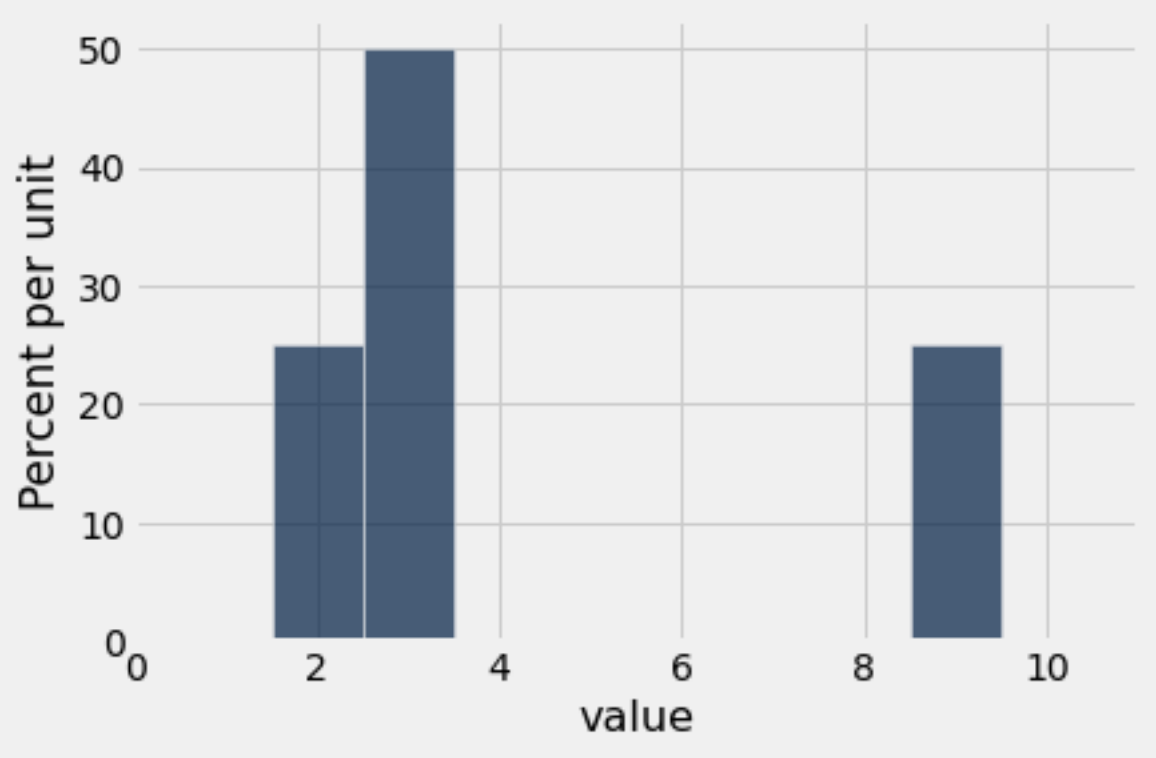
- Need not be a value in the collection
- Need not be an integer even if the data are integers
- Somewhere between min and max, but not necessarily halfway in between
- Same units as the data
- Smoothing operator: collect all the contributions in one big pot, then split evenly

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

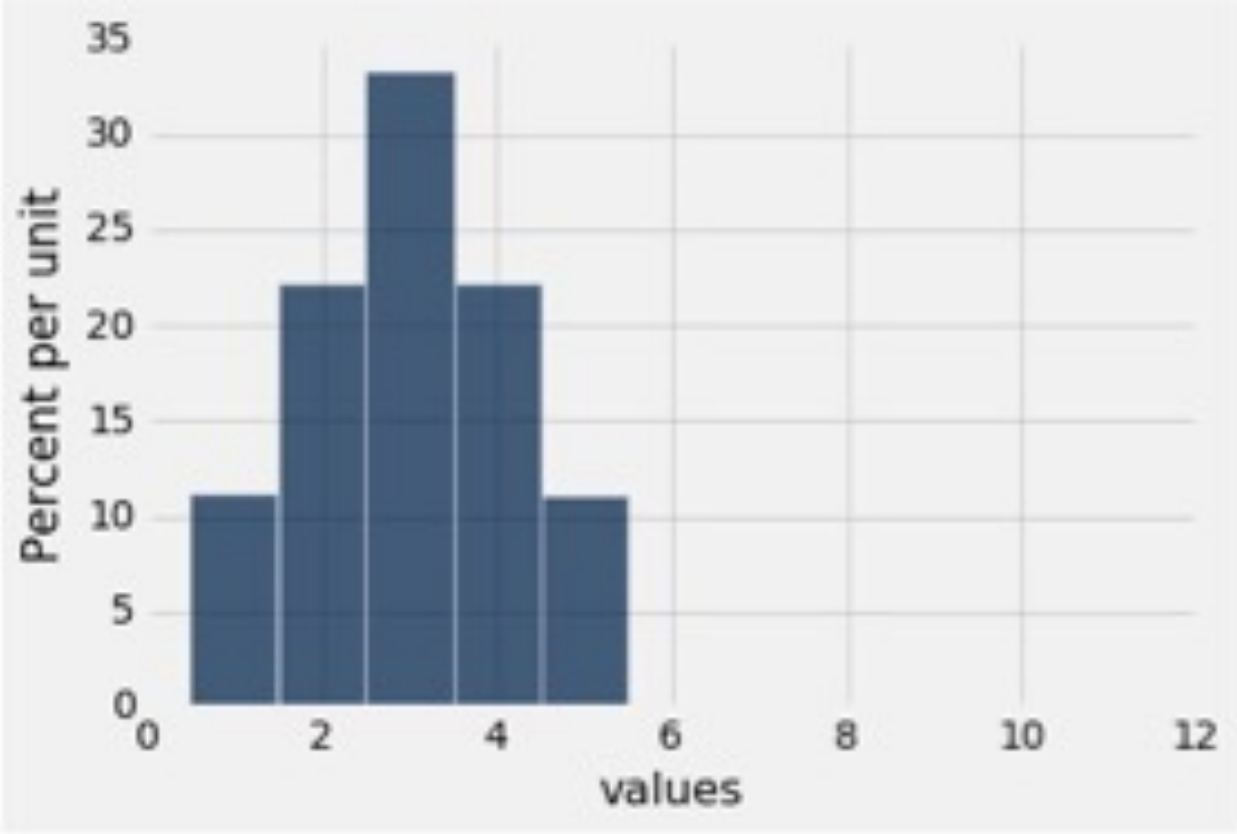
- Average is 4.25



Average and Median

Question

- What list produces this histogram?

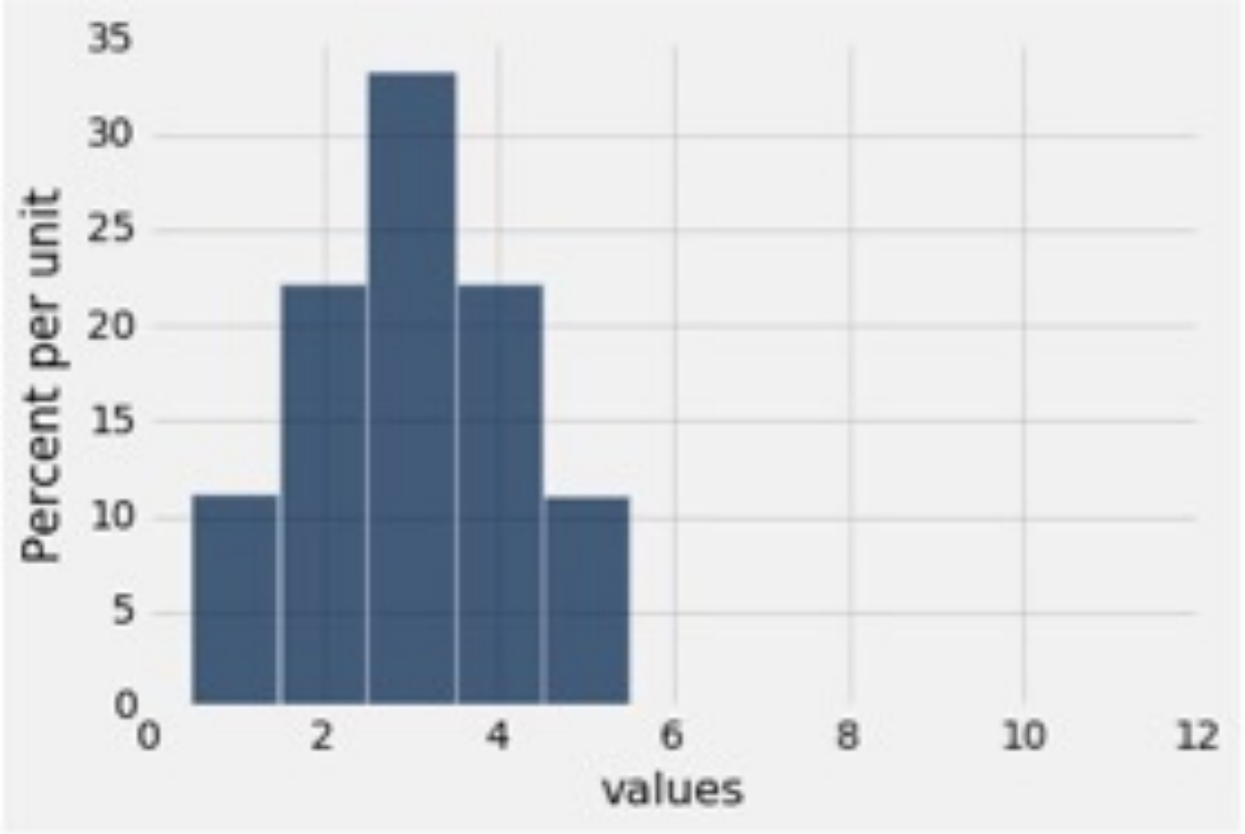


Question

- What list produces this histogram?

1, 2, 2, 3, 3

3, 4, 4, 5



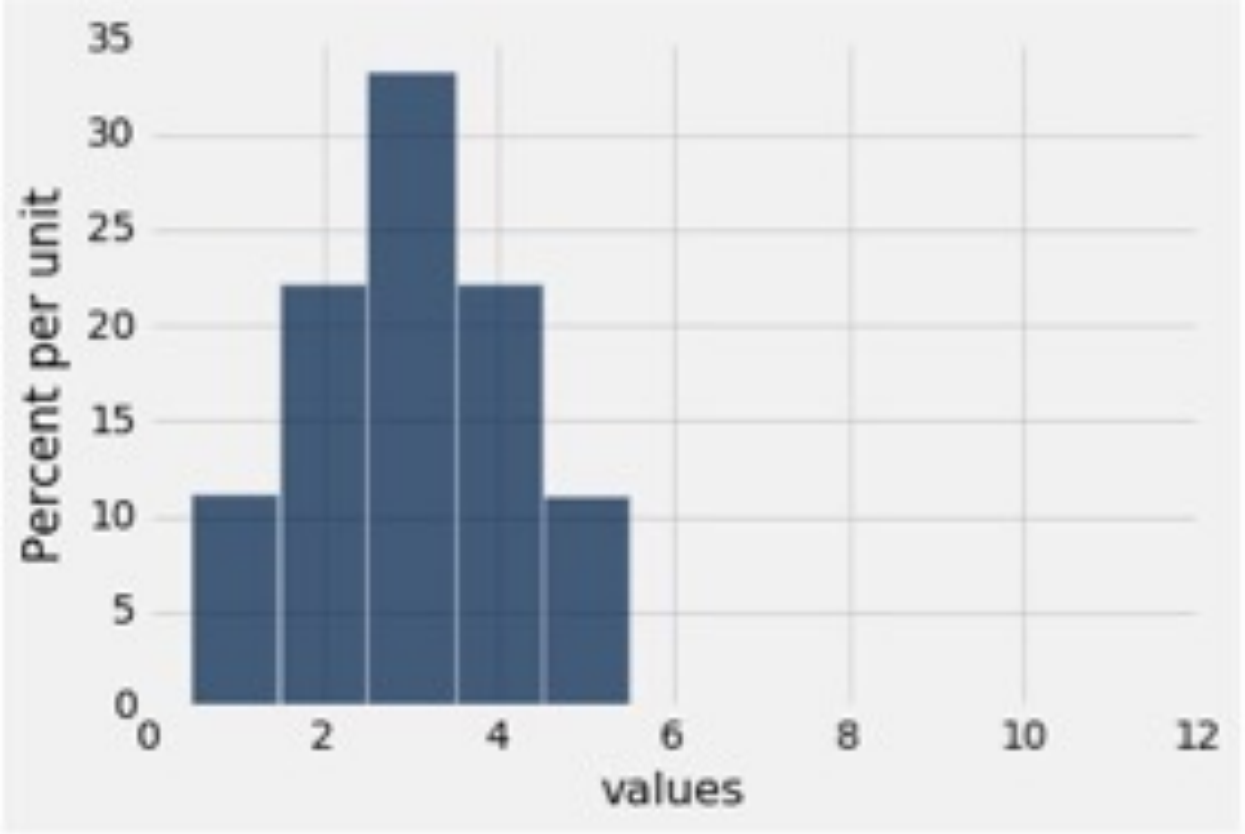
Question

- What list produces this histogram?

1, 2, 2, 3, 3

3, 4, 4, 5

- Average?



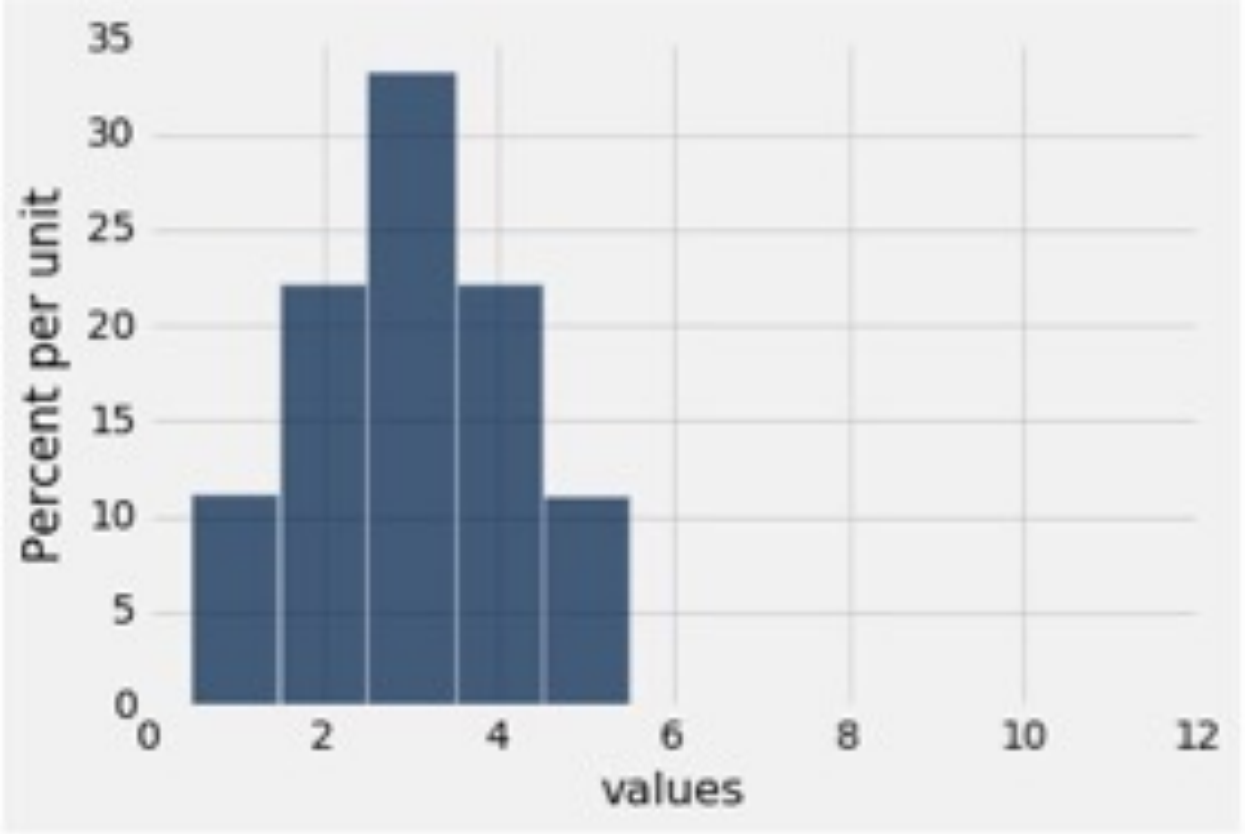
Question

- What list produces this histogram?

1, 2, 2, 3, 3

3, 4, 4, 5

- Average?
  - 3

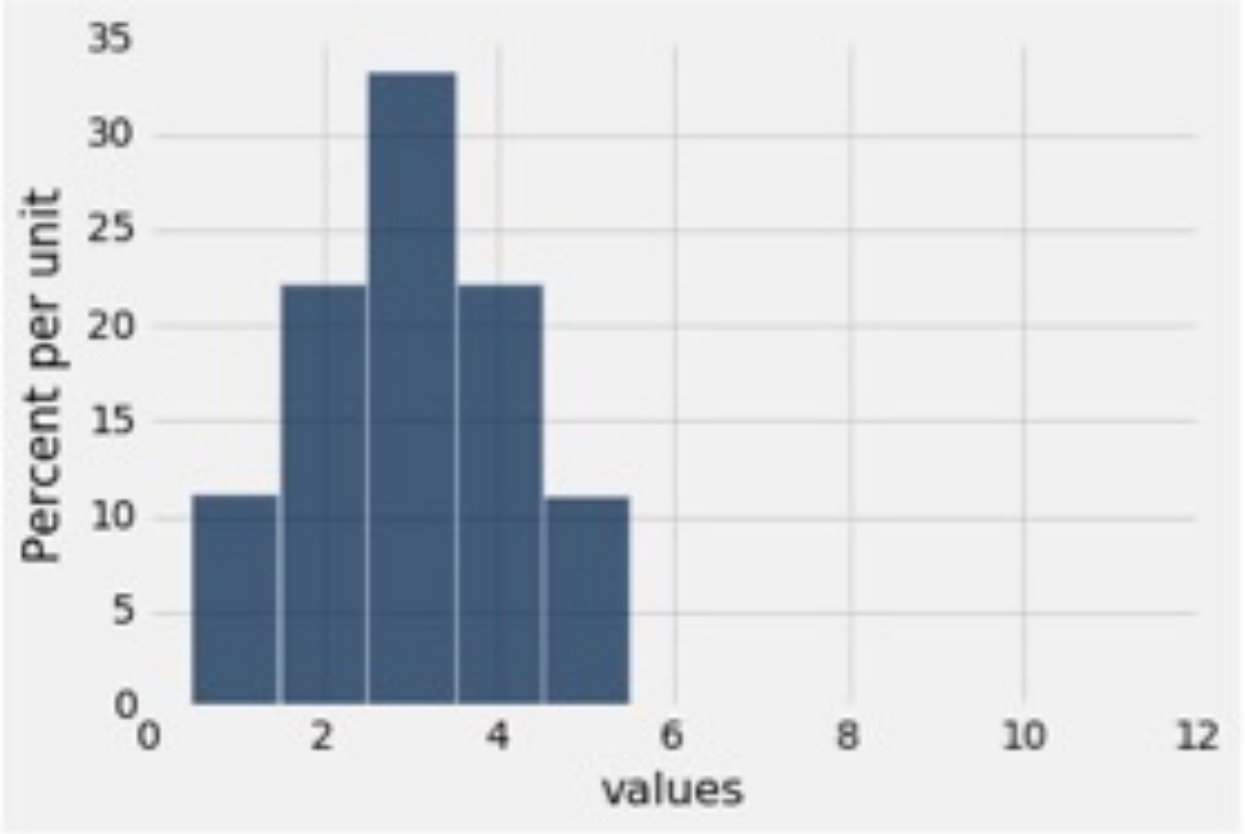


Question

- What list produces this histogram?

1, 2, 2, 3, 3

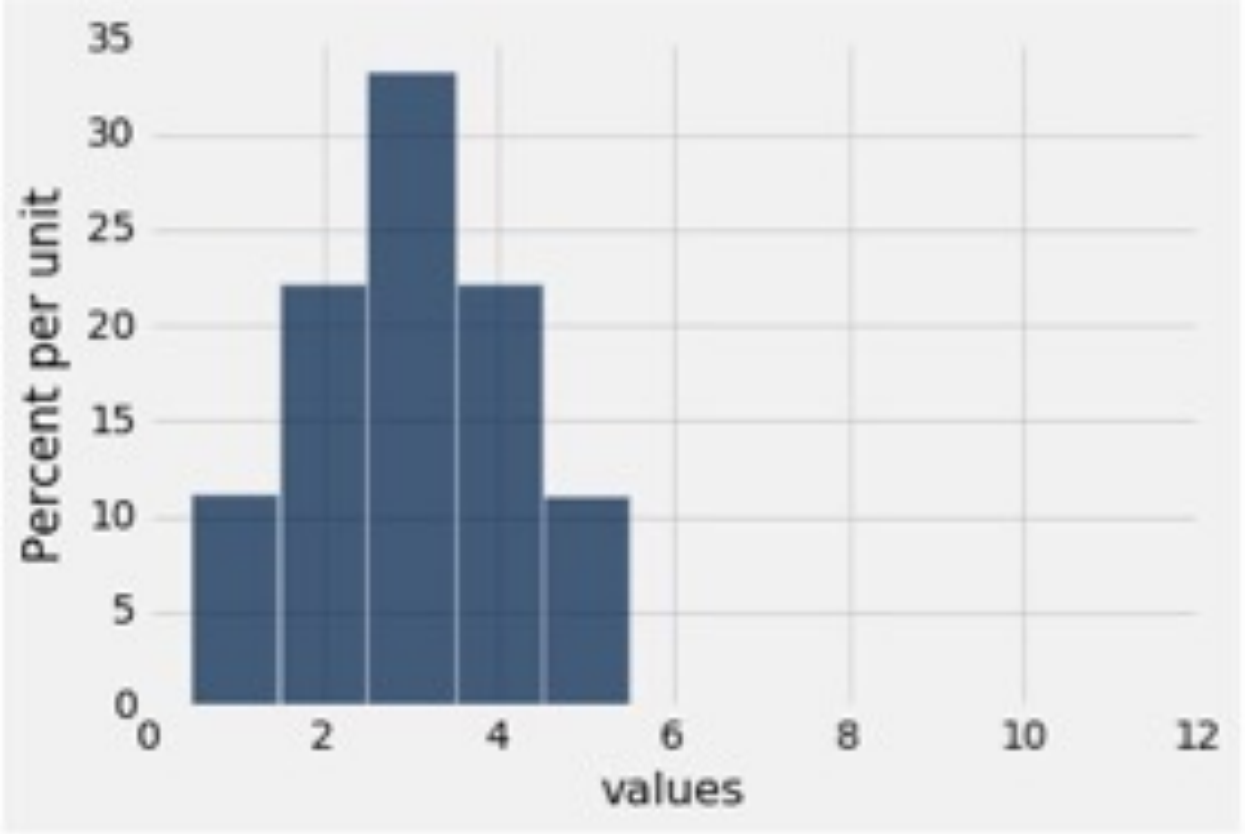
3, 4, 4, 5



- Average?
  - 3
- Median?

Question

- What list produces this histogram?



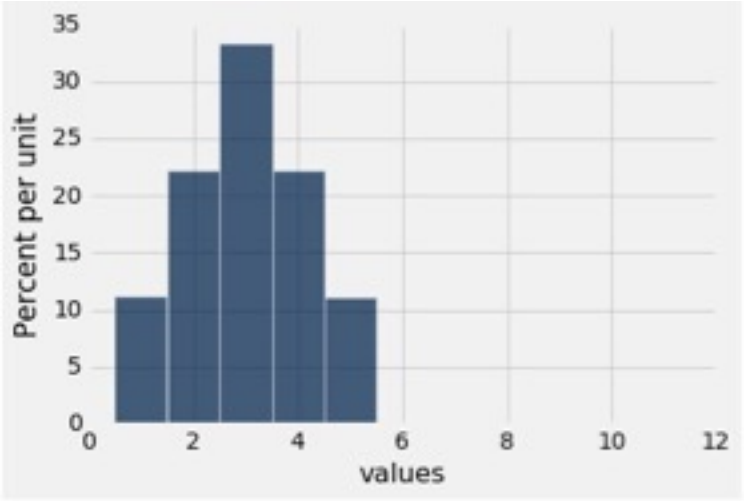
1, 2, 2, 3, 3

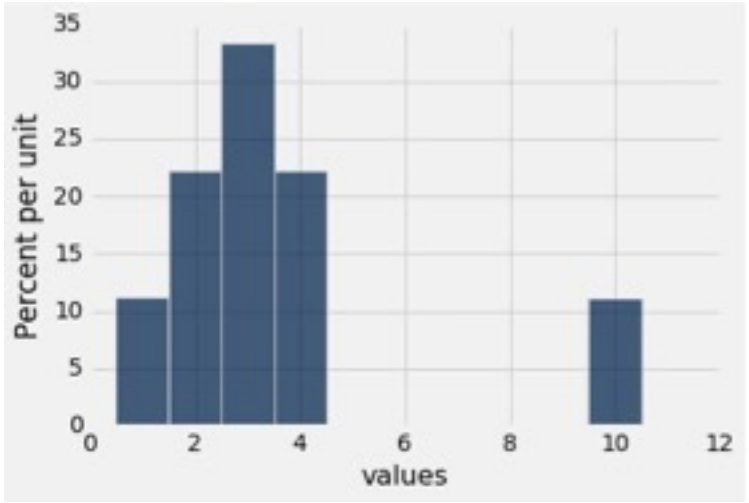
3, 4, 4, 5

- 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

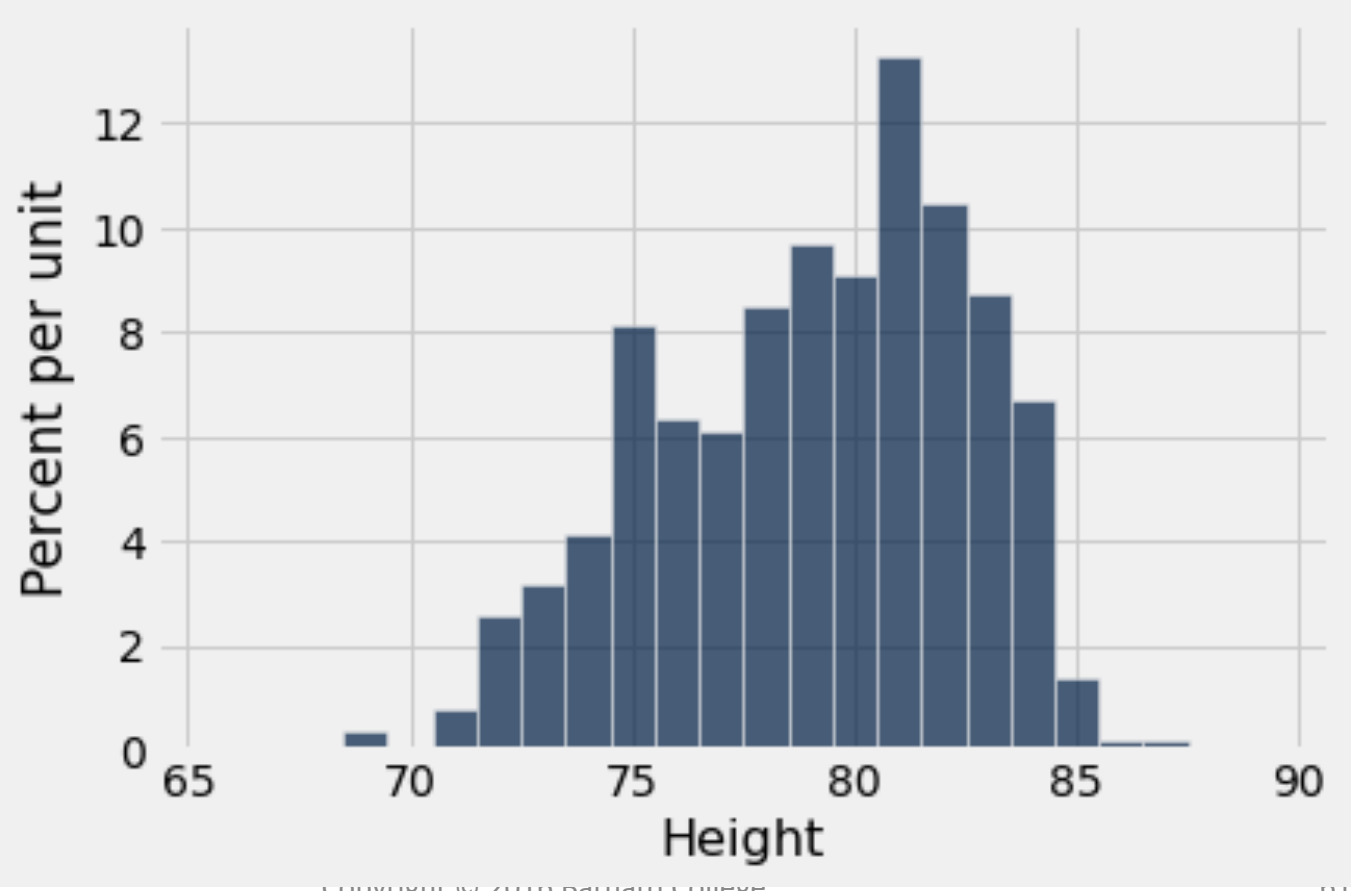
- List 1
  - 1, 2, 2, 3, 3, 3, 4, 4, 5
- List 2
  - 1, 2, 2, 3, 3, 3, 4, 4, 10
- Medians = 3
- Mean(List1) = 3
- Mean (List 2) = 3.55556

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 in the lecture

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

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/*z*2

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/*z*2

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/*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%)

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

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

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.

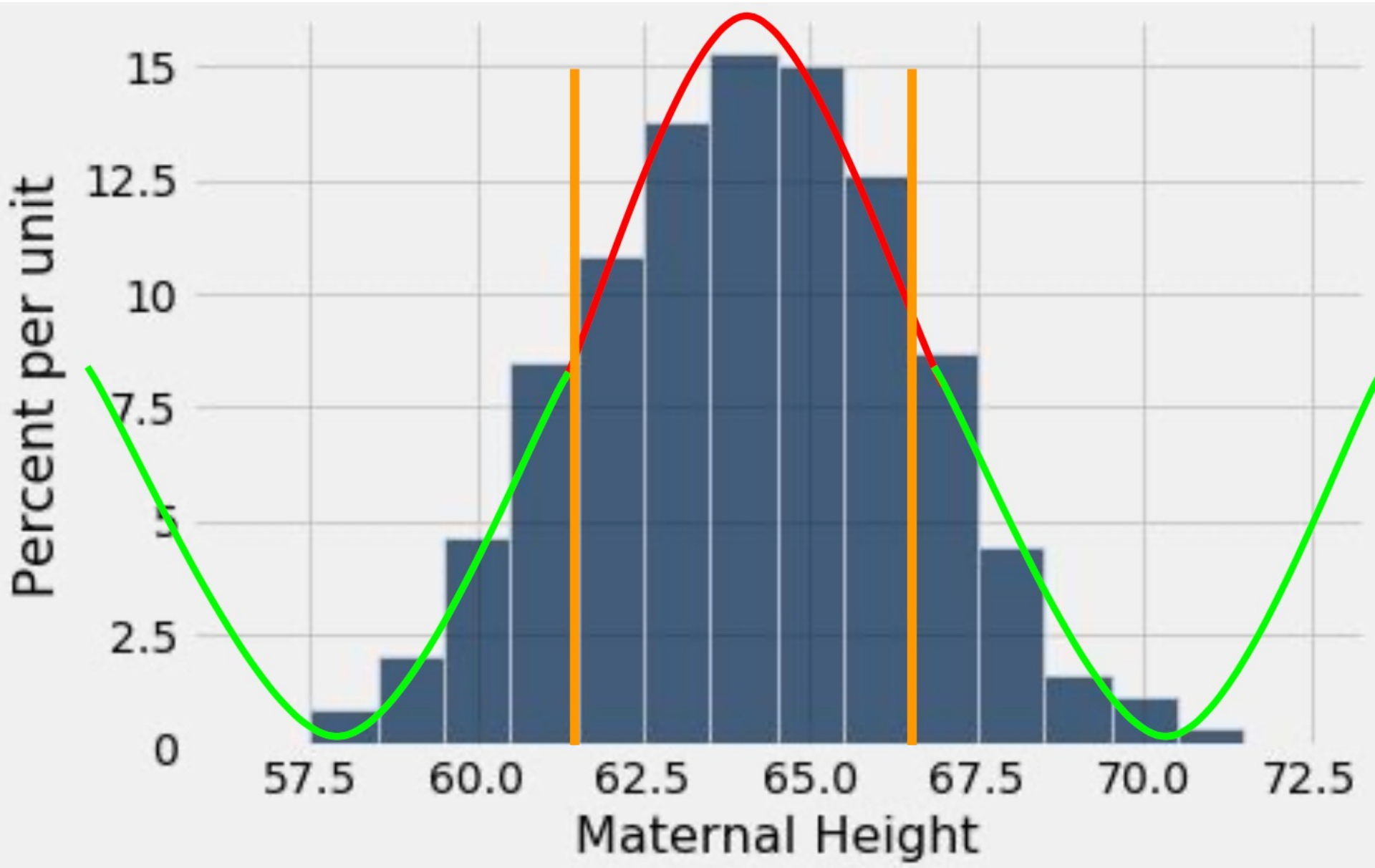
- But if the histogram has a bell shape, then you can

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



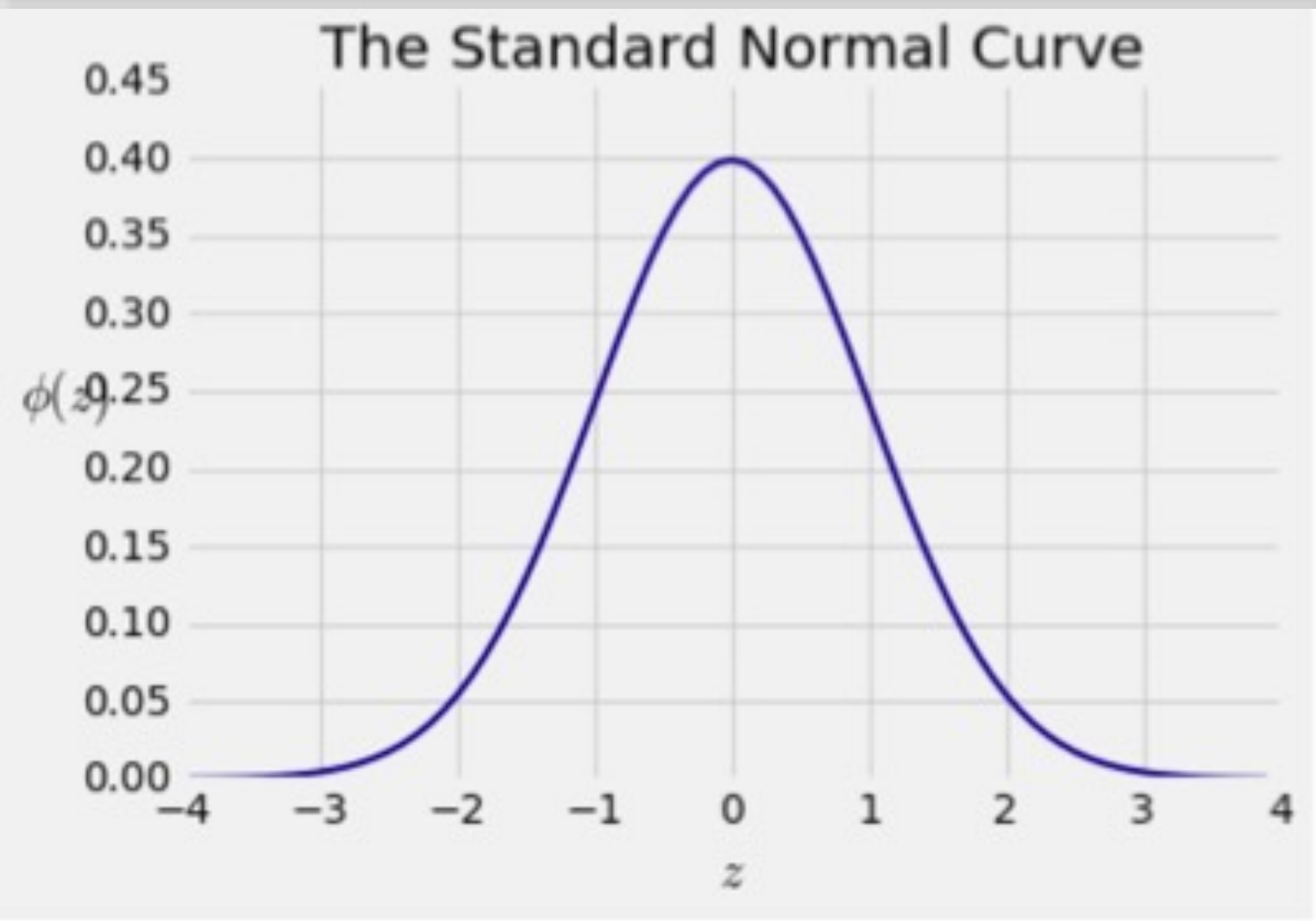
Normal Distribution

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

Equation for the normal curve

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

Bell Curve



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"

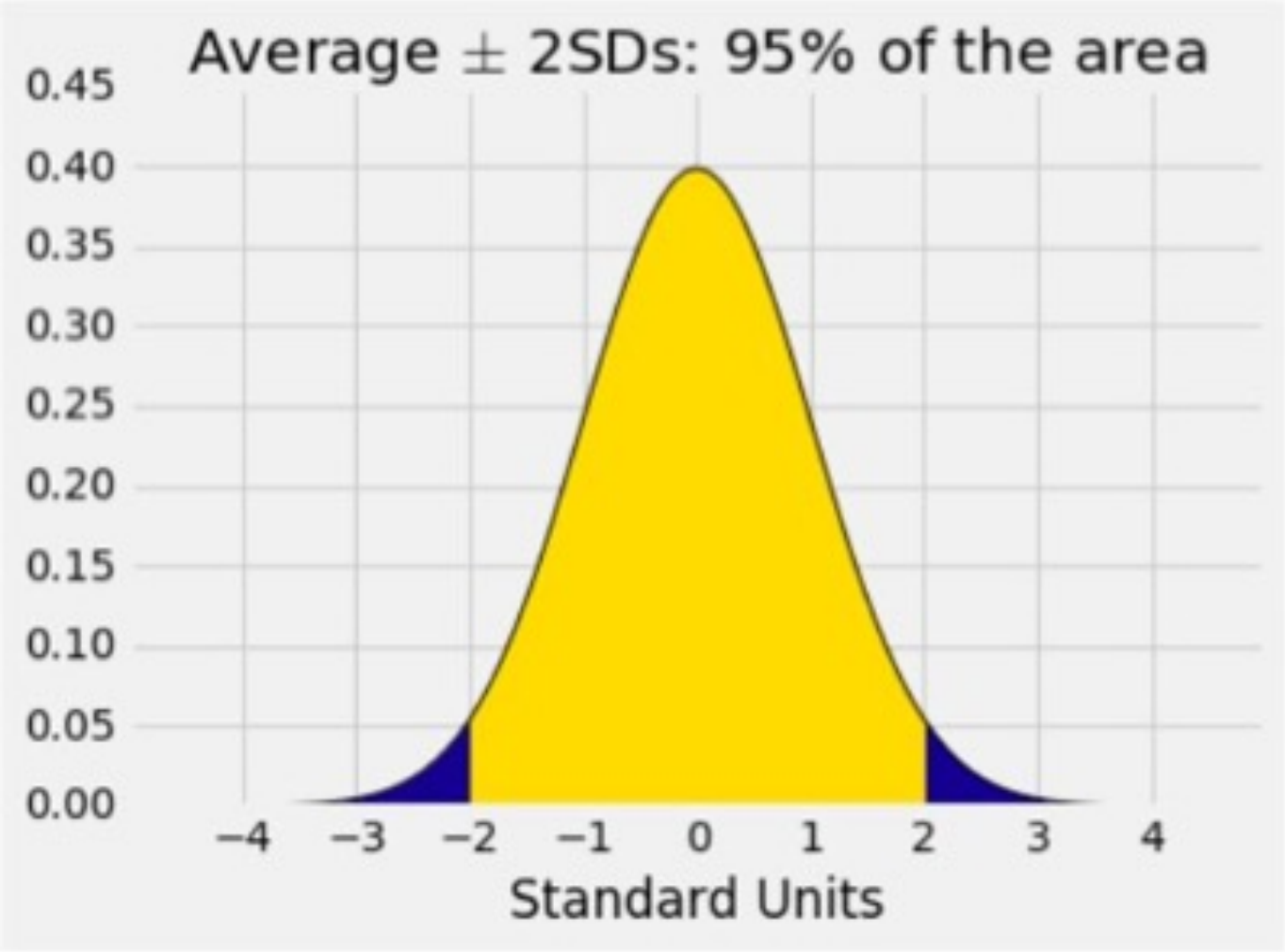
If a histogram is bell-shaped, then

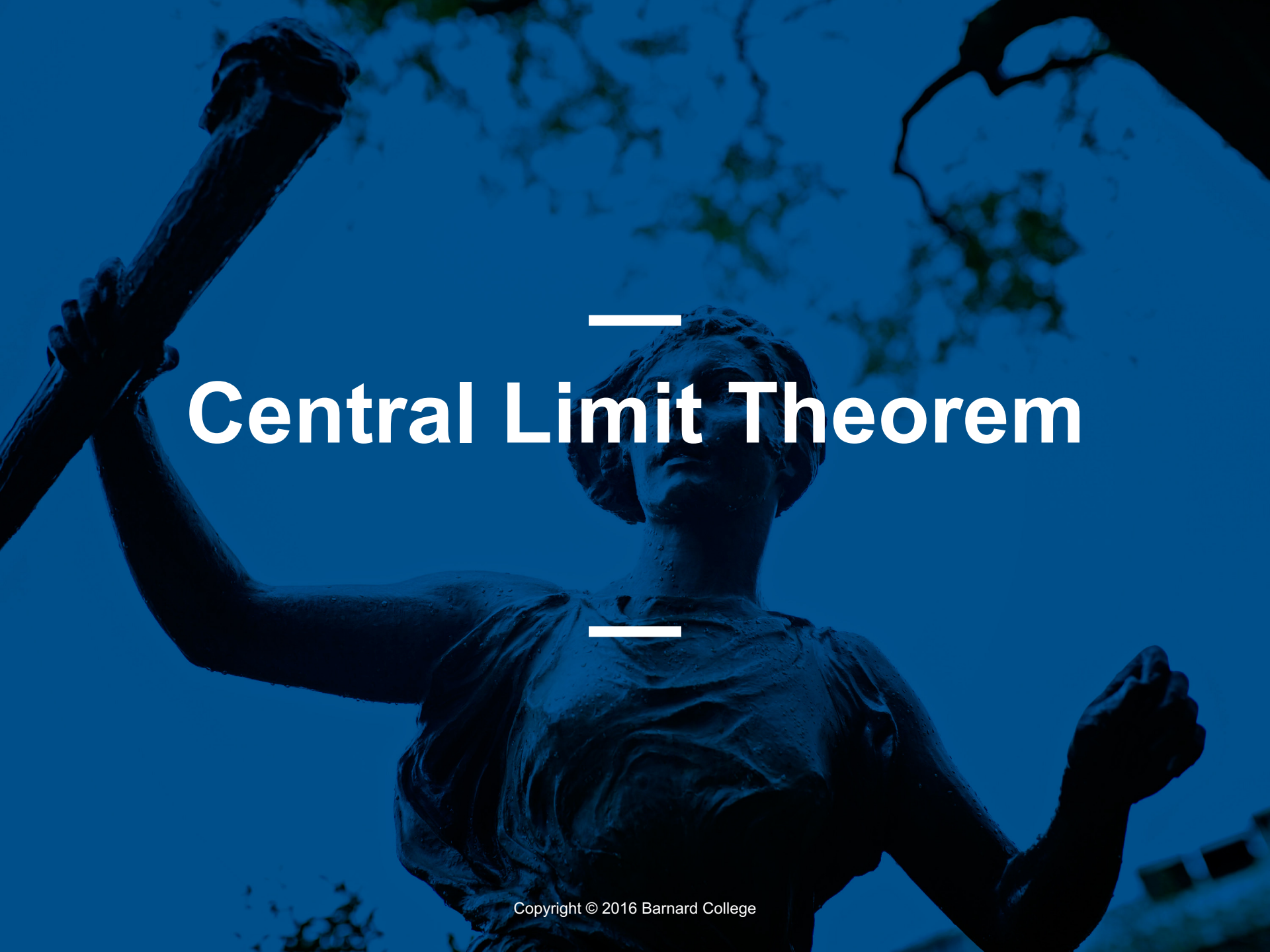
- Almost all of the data are in the range "average ± 3 SDs

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