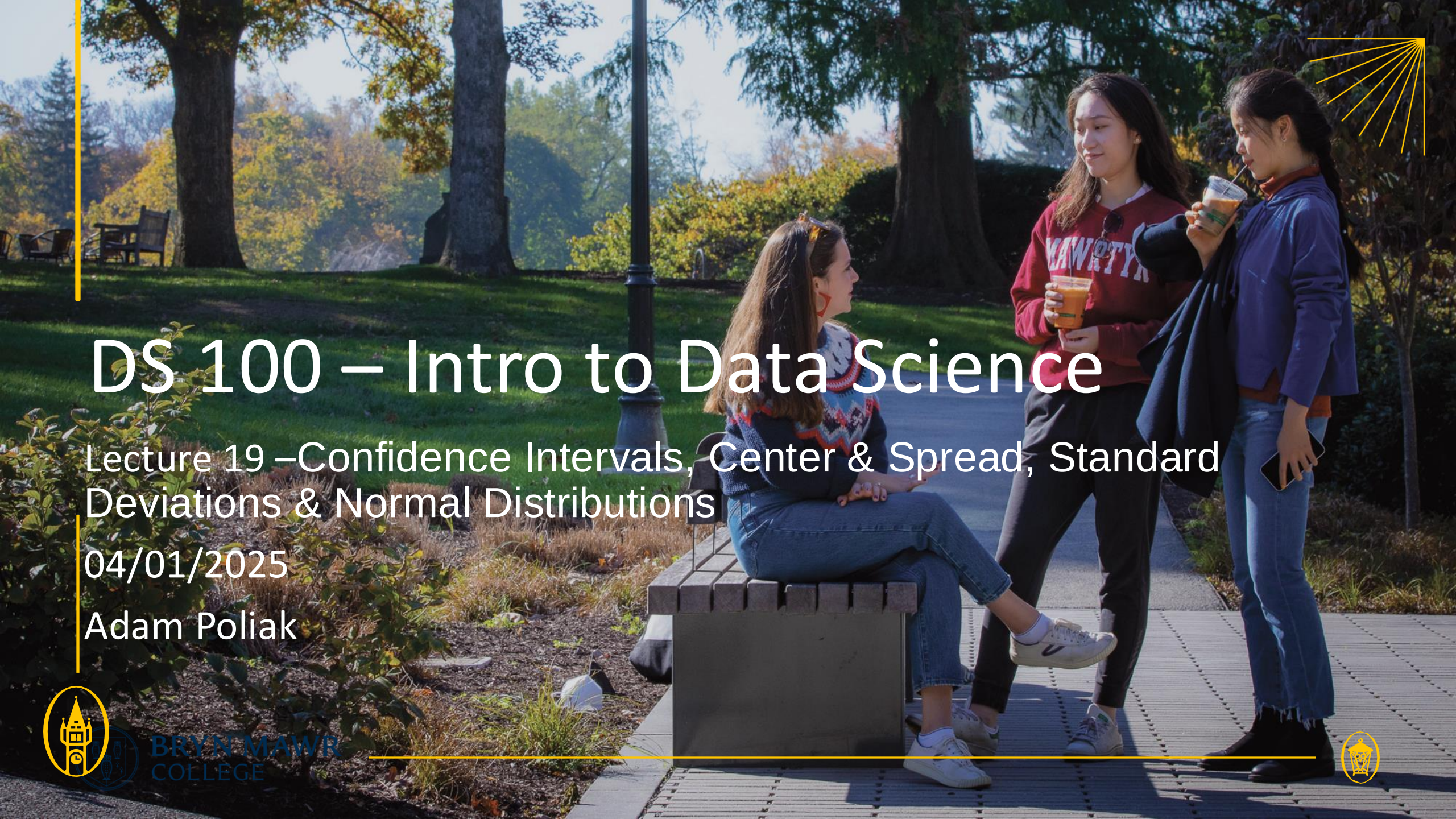


DS 100 – Intro to Data Science

Lecture 19 –Confidence Intervals, Center & Spread, Standard Deviations & Normal Distributions

04/01/2025

Adam Poliak

BRYN MAWR COLLEGE

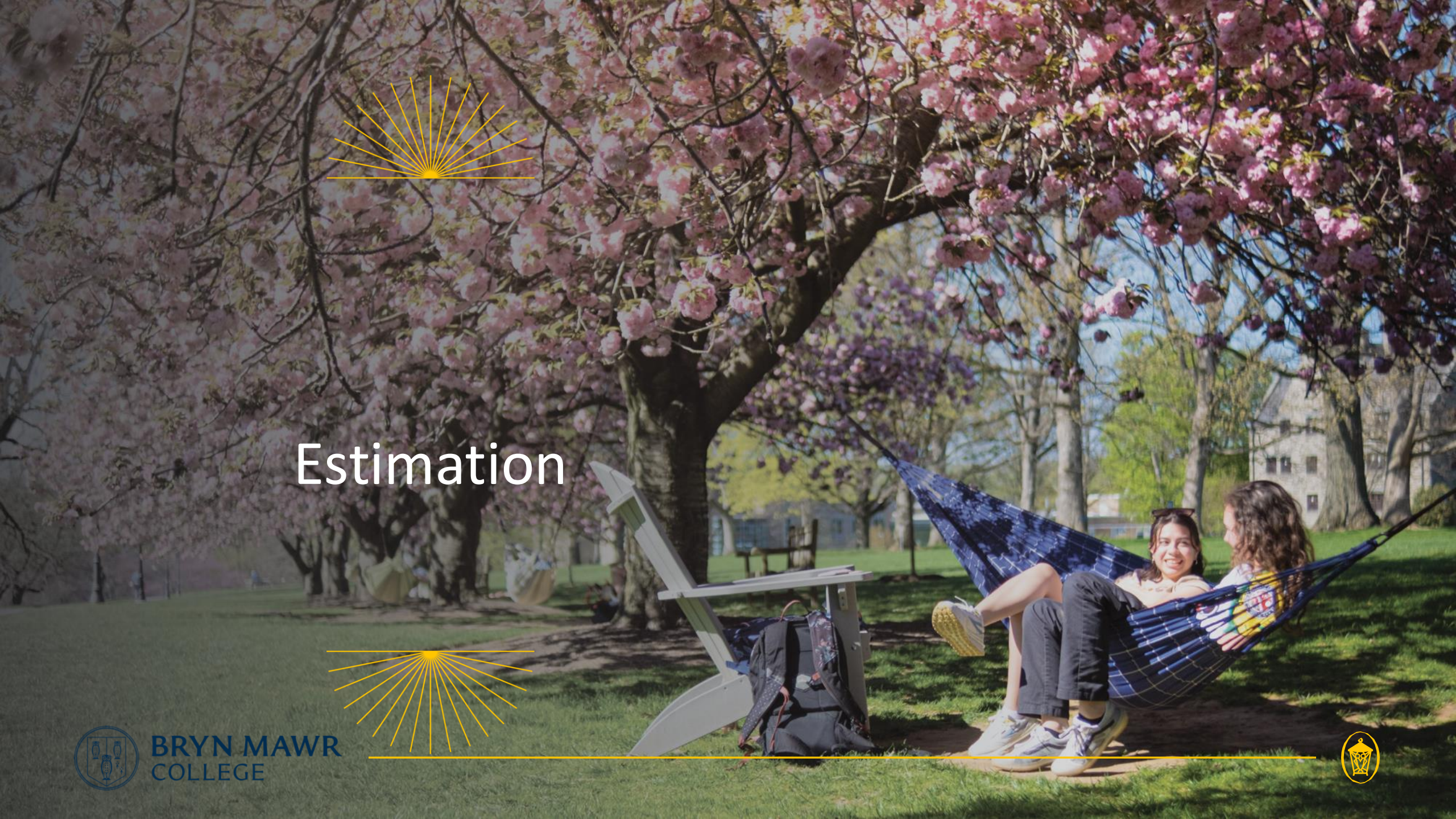
Announcements

Midsemester feedback form - https://forms.gle/M4jVdGTDgQknWAwo6

Lab 07 (due Friday April 4th)

HW07 (due Wednesday April 9th)

Project 2 (due Friday April 11th)



Estimation

BRYN MAWR COLLEGE


Estimation Variability
BRYN MAWR
COLLEGE

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

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 ≠ *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
BRYN MAWR
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?

Average & Histogram

BRYN MAWR
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

Relation to the histogram

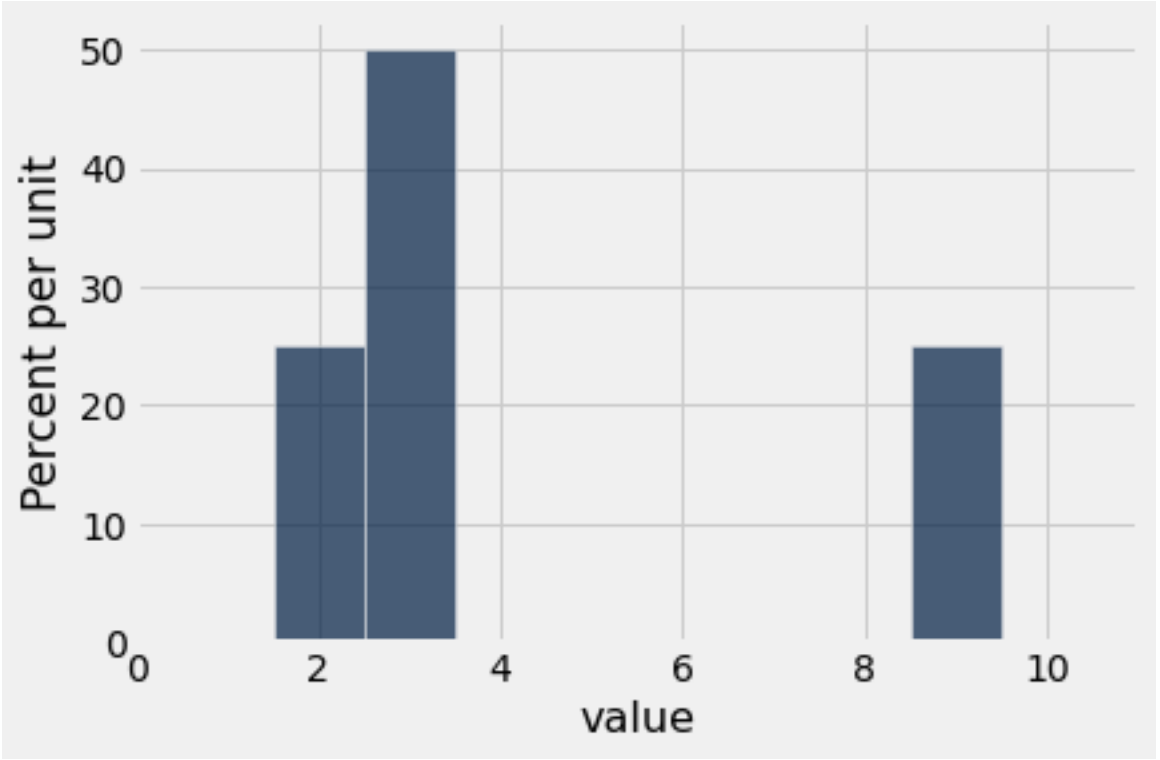
The average depends only on the **proportions** in which the distinct values appears

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

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

Average as a balance point

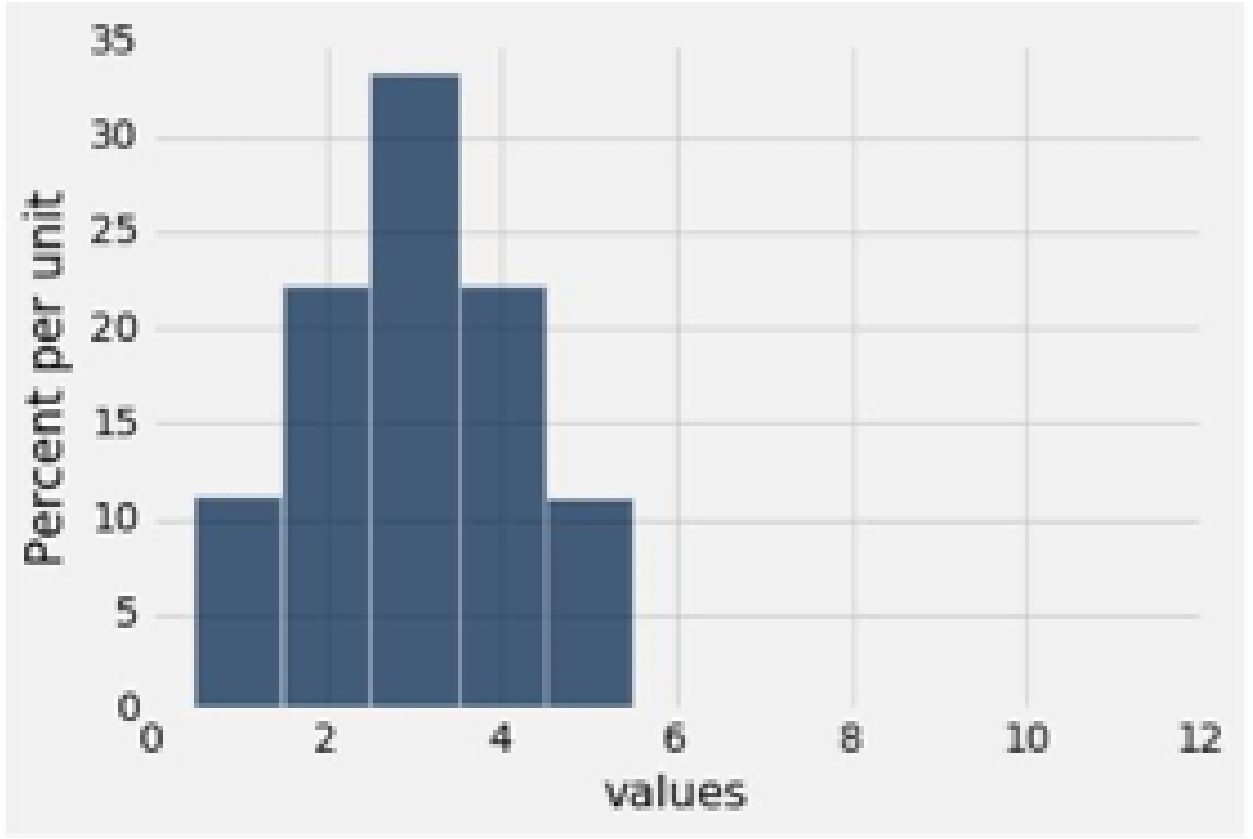
Average is 4.25



Average & Median

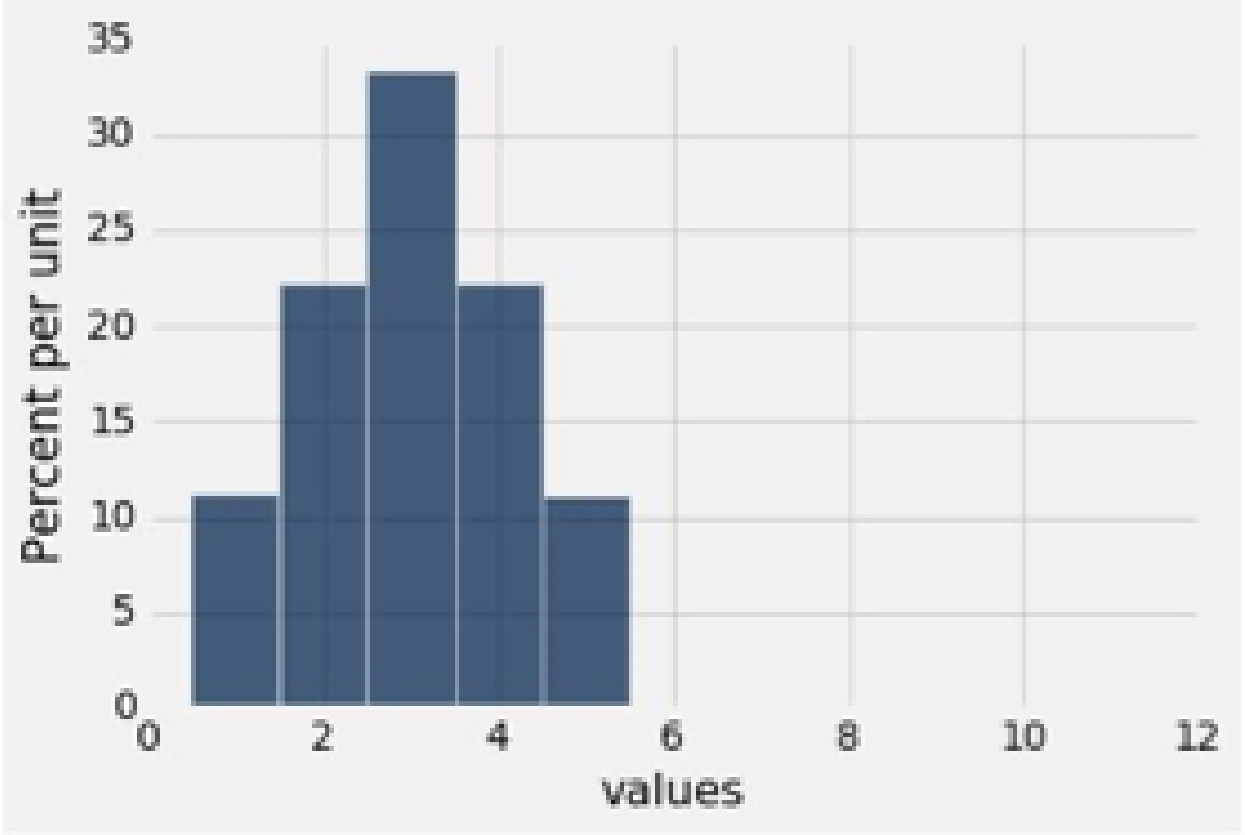
BRYN MAWR
COLLEGE

What list produces this histogram?



What list produces this histogram?

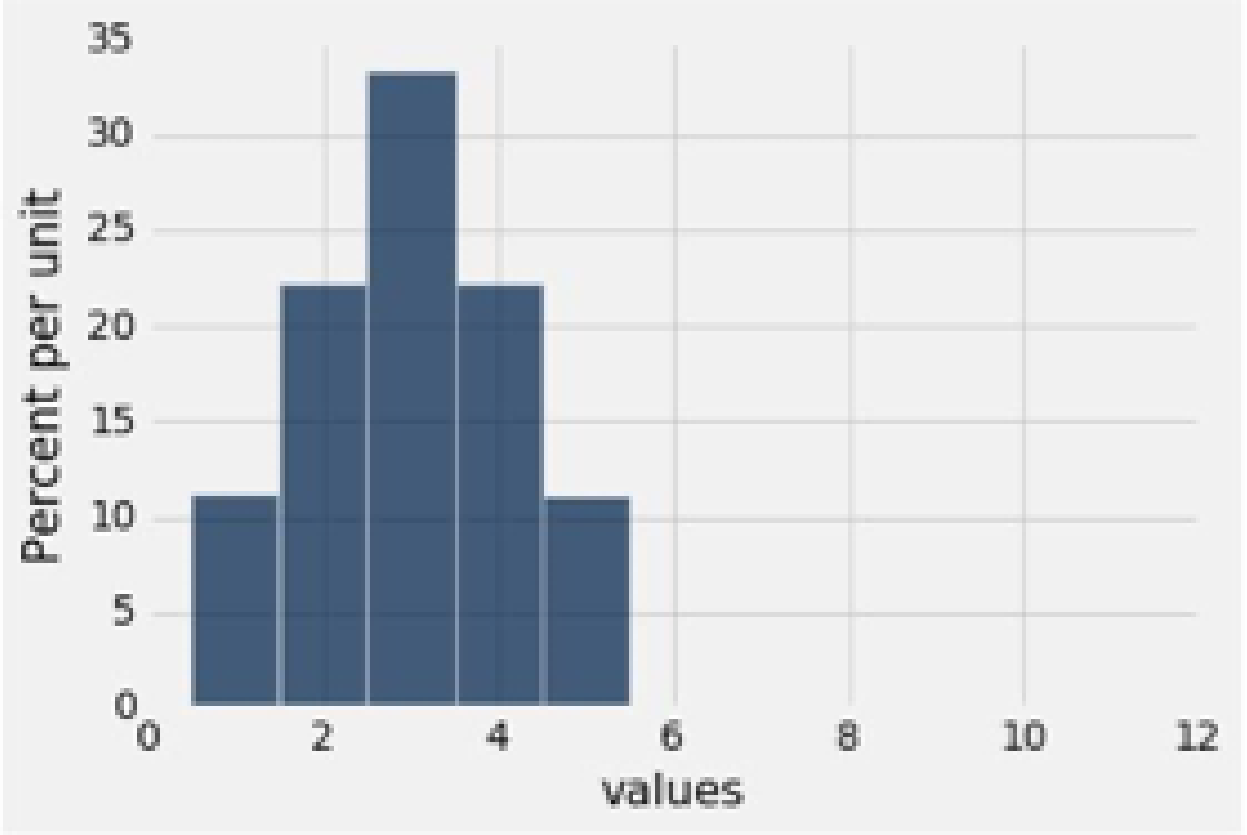
1, 2, 2, 3, 3
3, 4, 4, 5



What list produces this histogram?

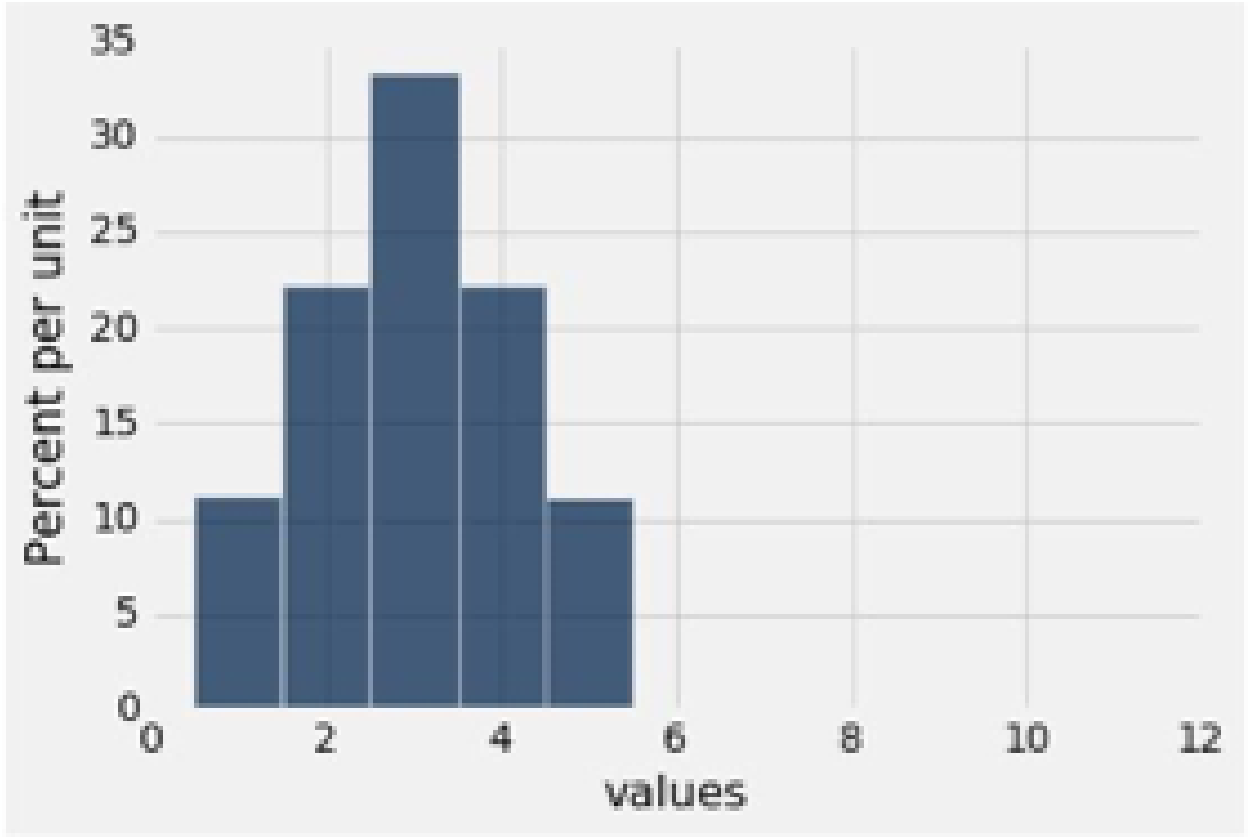
1, 2, 2, 3, 3
3, 4, 4, 5

Average?



What list produces this histogram?

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

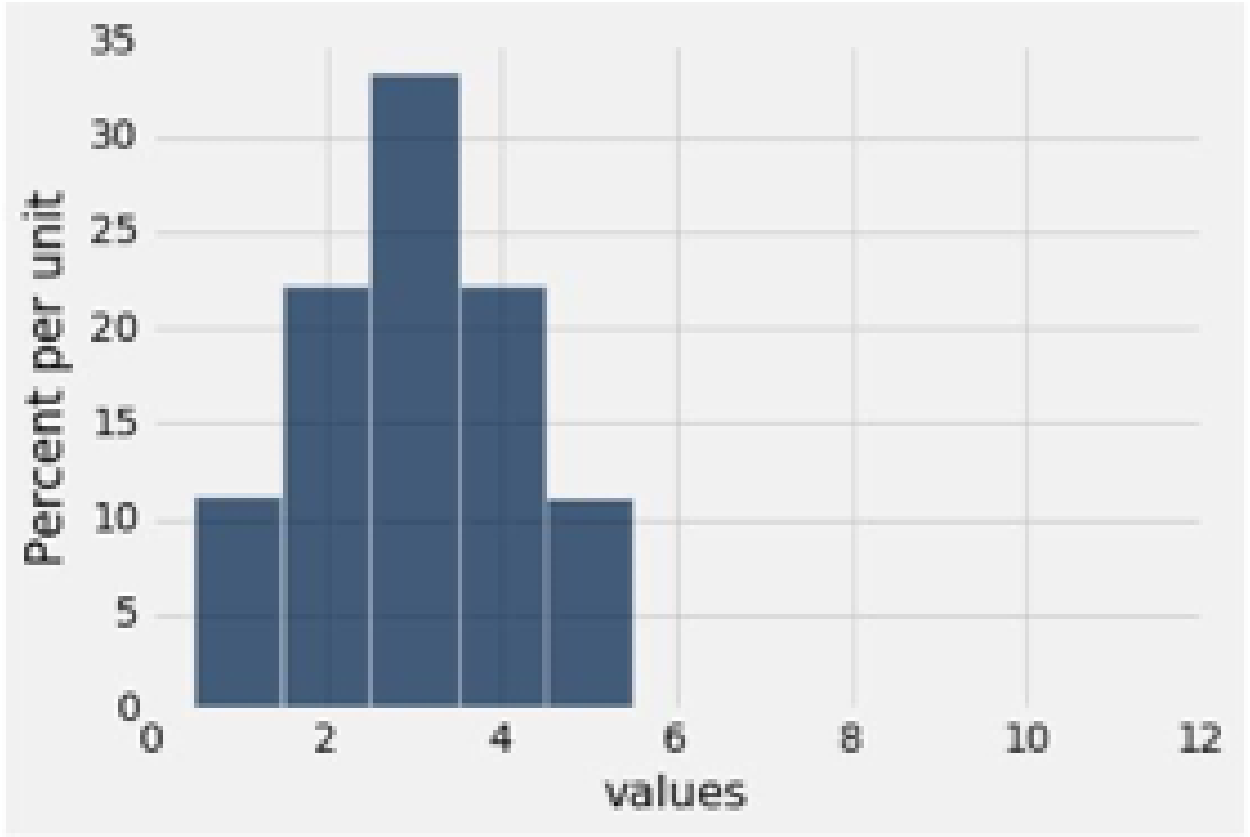


Average?
3

Median?

What list produces this histogram?

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

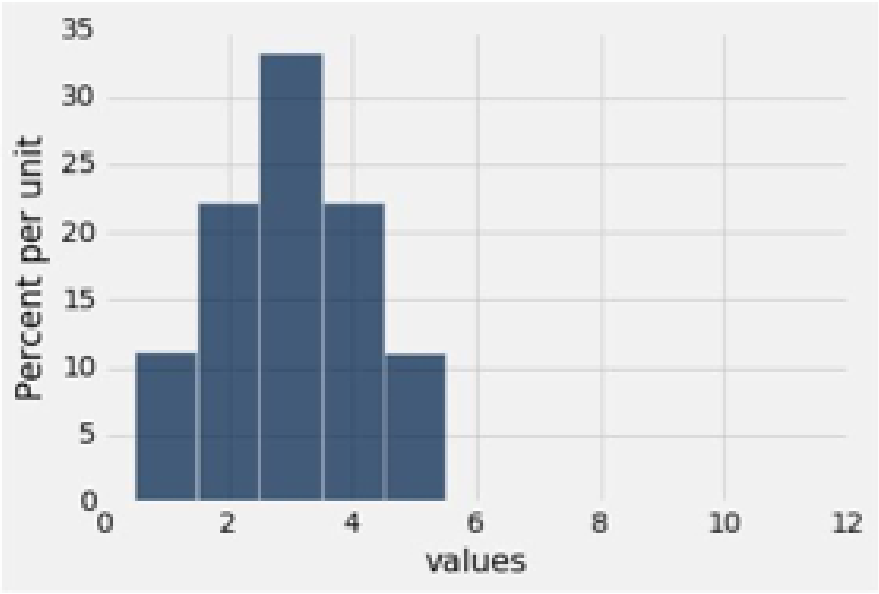


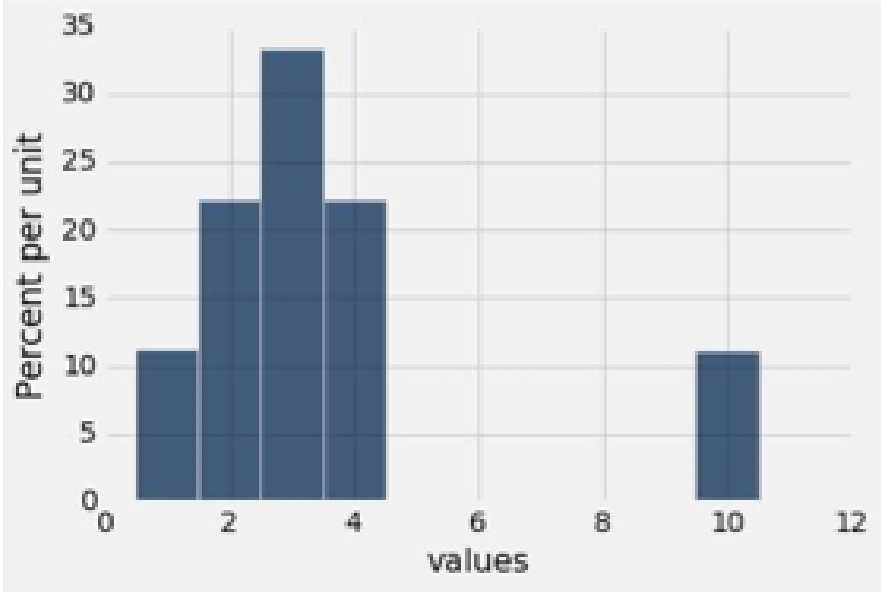
Average?
3

Median?
3

Question 2

Are the medians of these two distributions the same or different? Are the mean 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

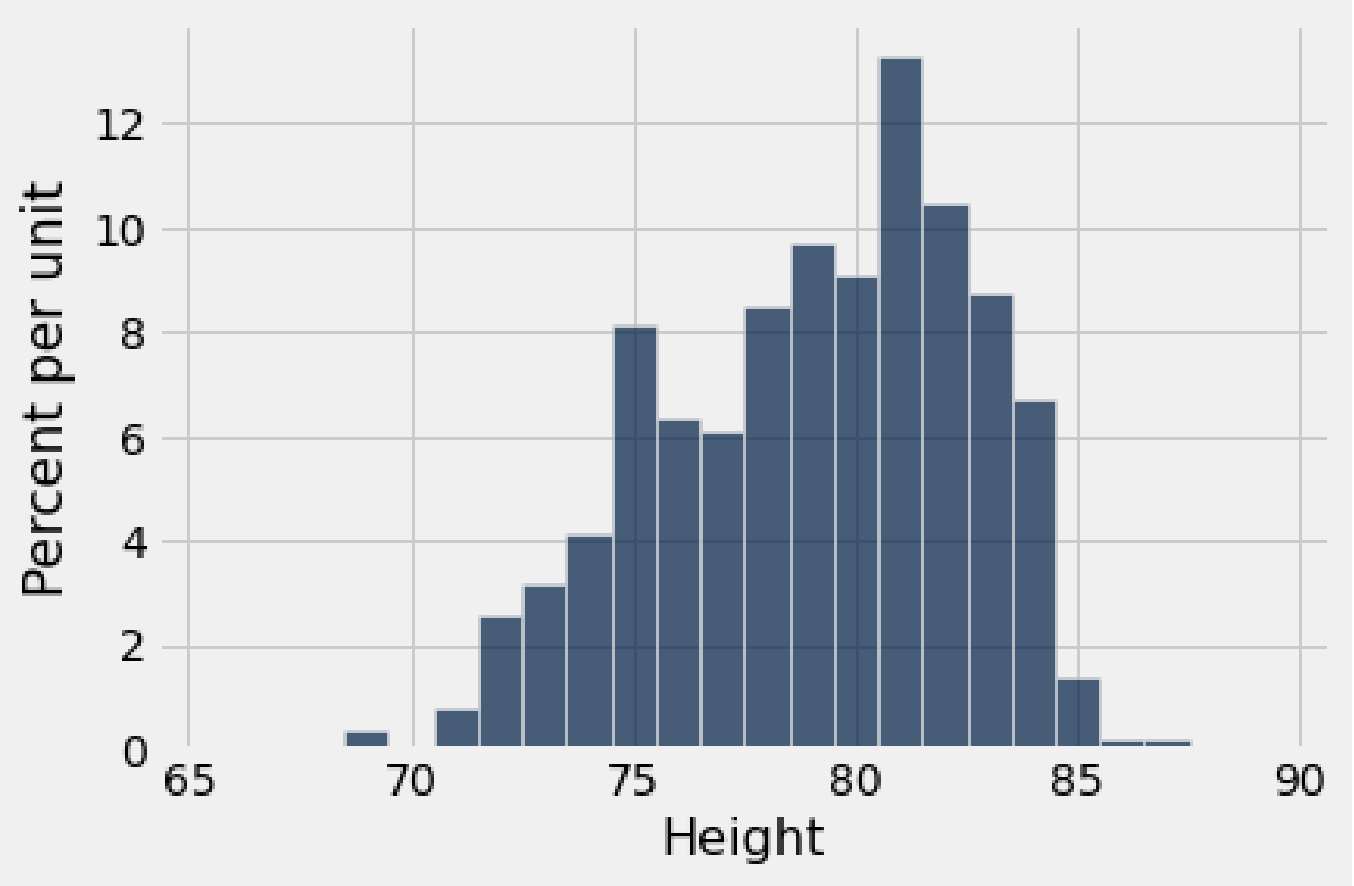
Mean: Balance point of the histogram

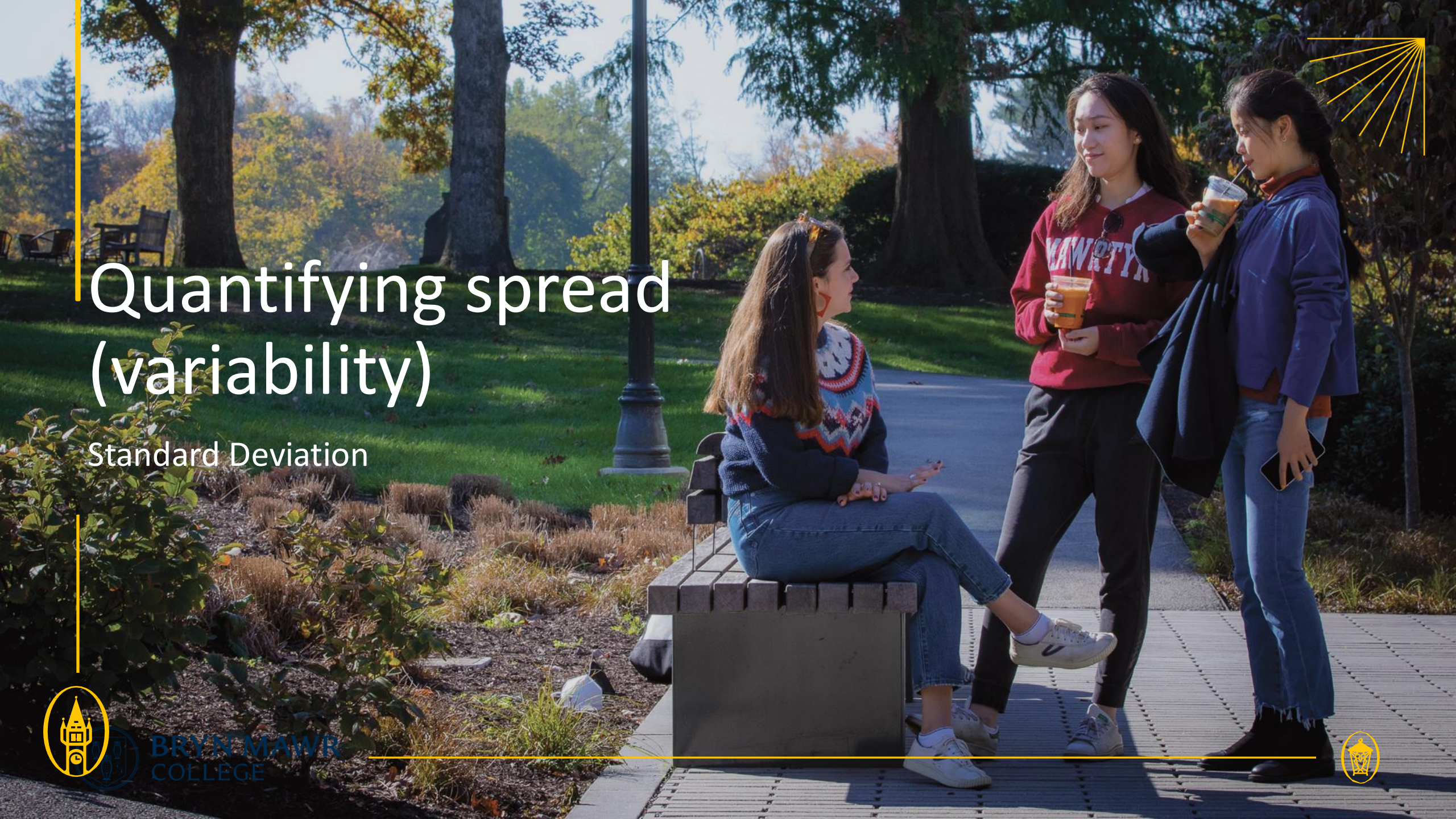
Median: Half-way point of data; half the area of histogram is on either side of median

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.

Which is bigger? Median or Mean?





Quantifying spread (variability)

Standard Deviation

BRYN MAWR 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

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

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 $1 - 1/z2$

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

Understand the Midterm

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

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

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

The SD and the Histogram

Usually, it's not easy to estimate the SD by looking at a histogram.

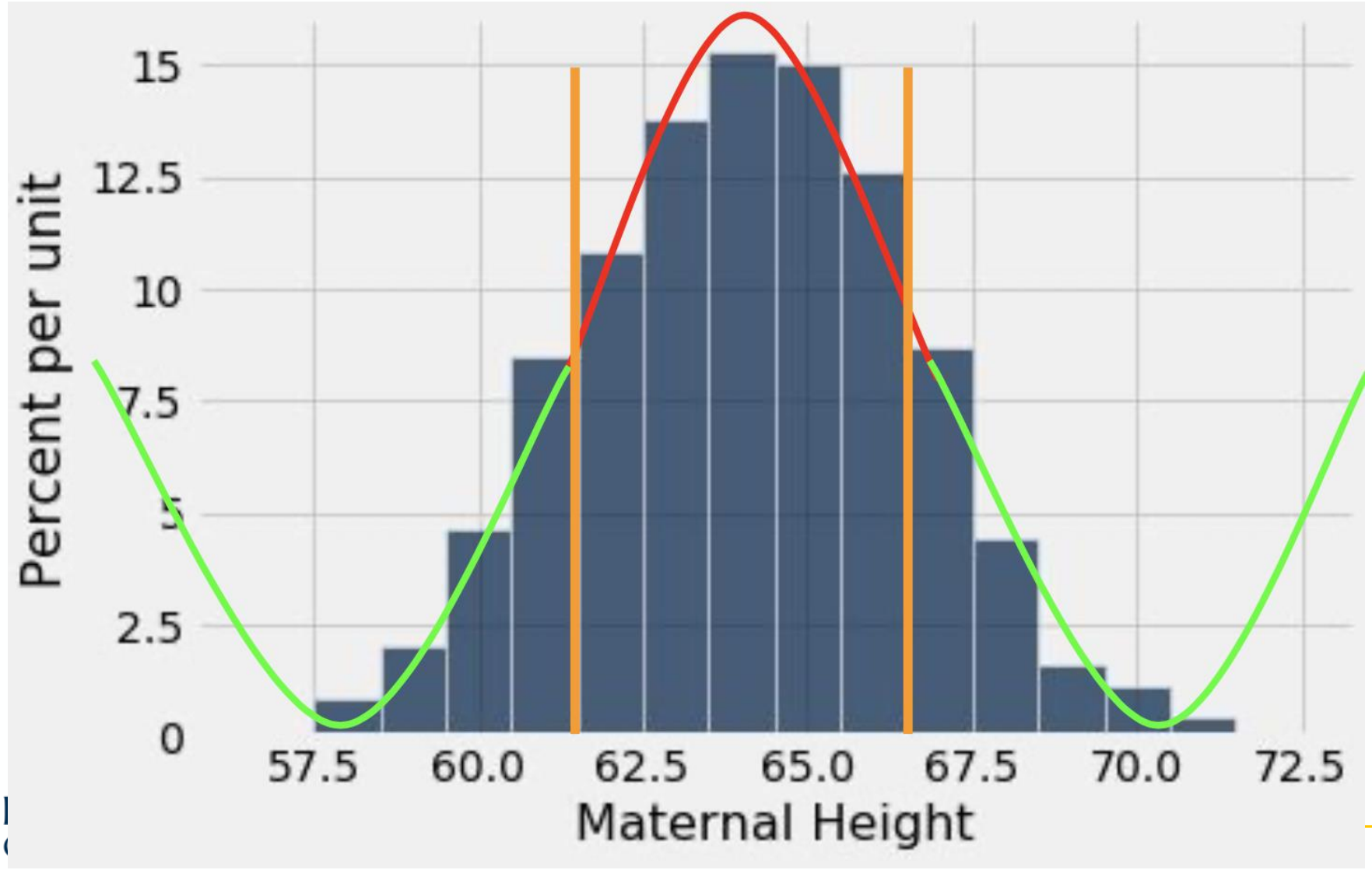
But if the histogram has a bell shape, then you can

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



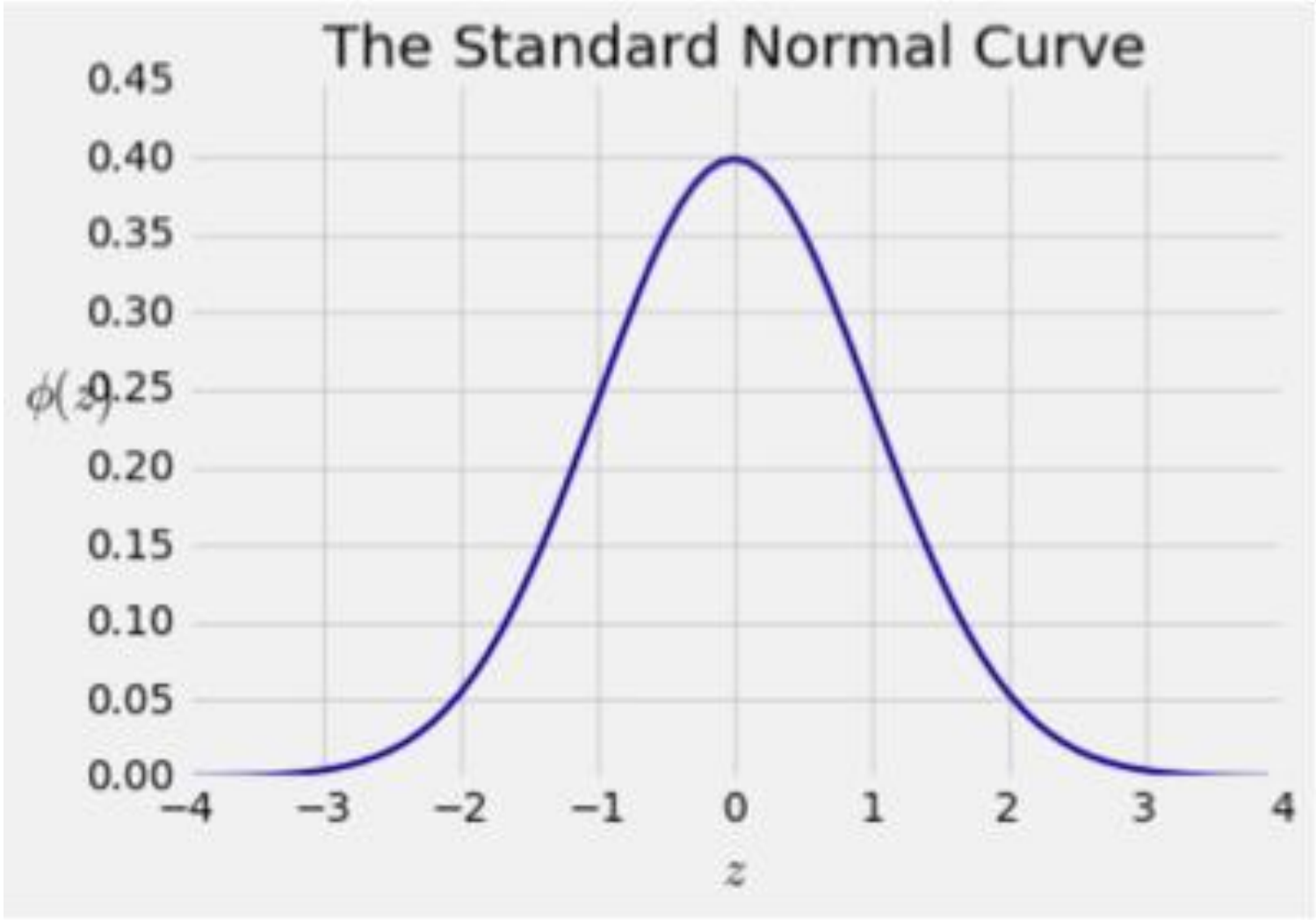
Normal Distribution

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

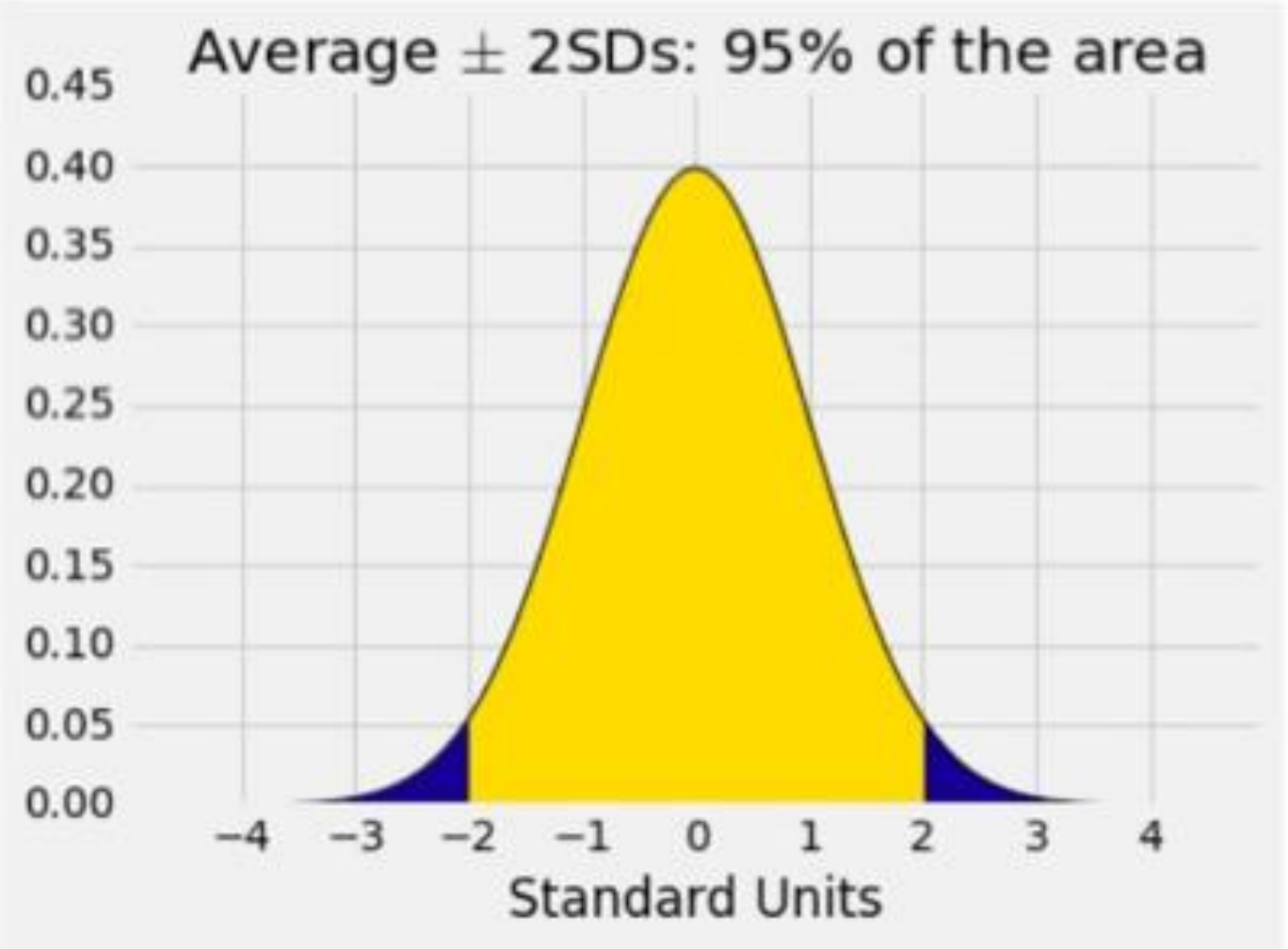
No matter what the shape of the distribution, the bulk of the data are in the range "average ± a few SDs"

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