COMS BC1016

Introduction to Computational Thinking and Data Science

Lecture 16: The Normal Distribution

Spring 2025

Copyright © 2026 Barnard College

November 16, 2026

Upcoming

Today →

	11/16	16 - The Normal Distribution			
11	11/18	17 - Sample Means	Lab 8 - Normal Distribution and Variability of Sample Means	HW 7 - Sample Sizes and Confidence Intervals	HW 5 Due
	11/20	-			**Final Project Proposals** Due
12	11/23	Special Topics			HW 6 Due
	11/25	*Holiday*	*No Lab*		
13	11/30	18 - Correlation and Linear Regression		HW 8 - Linear Regression	
	12/2	19 - Least Squares, Residuals, and Regression Inference	Lab 9 - Regression		HW 7 Due
14	12/7	20 - Classification		HW 9 - Classification	**Final Project Progress Reports** Due
	12/9	21 - Classification	Lab: Final Project Consultations		HW 8 Due

Last Time: Estimation and Confidence Intervals

Percentile

The p-th percentile of a collection is the smallest value in the collection that is at least as large as $p\%$ of all the values

```
percentile(p, values_array)
```

Estimation

- Parameter is a **fixed** quantity in the population
  - Example: the average age of students in 1016 has a fixed value
- If you don’t have a census:
  - Take a random sample from the population
  - Compute the statistic as an **estimate** of the parameter
- However, sample may have come out differently and affected out estimate
  - Can we quantify how different our estimate could’ve been?

Confidence Intervals

- **Confidence interval (CI)**: instead of a single estimate of a parameter, infer a **range of estimates** of the parameter
  - Consists of an upper bound, lower bound, and confidence level
  - Confidence level (e.g., 95%, 99%, 99.9%, …) indicates how confident we are in the process that we created the interval
    - E.g., 95% CI generates an interval containing the parameter 95% of the time

Where to Get Another Sample?

- Suppose we have a single sample
- We want to understand the variability of our estimate
- To get many values of an estimate, we need many random samples
- But what if we can’t go back and sample again from the population?
- Bootstrap method says we can treat it like a mini population!

The Bootstrap Method

- Suppose we have a large random sample from the population
  - By the Law of Averages, it probably resembles the population from which it's drawn
  - We can treat it like a miniature version of the population
- We can replicate sampling from the population by sampling from the sample

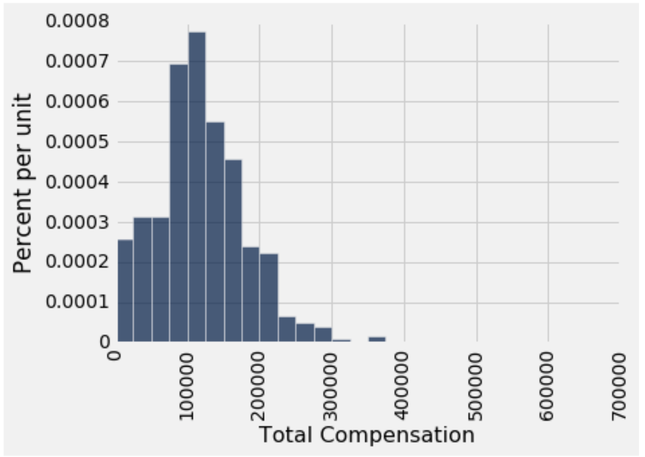
The Bootstrap

Population

?

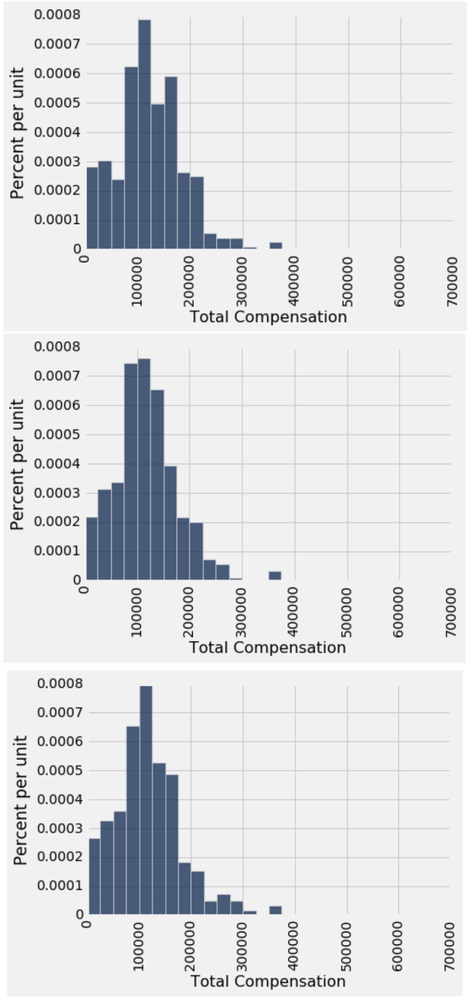
We don’t know the entire population and thus can’t calculate the parameter directly

Sample



However, we if we have a single sample…

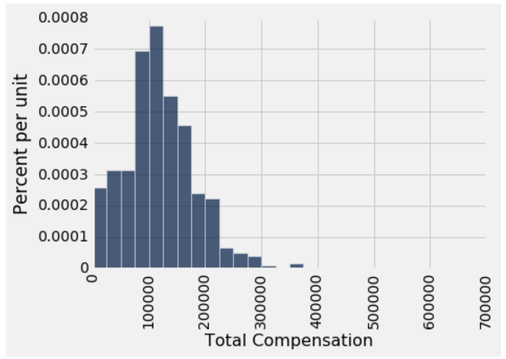
Resamples



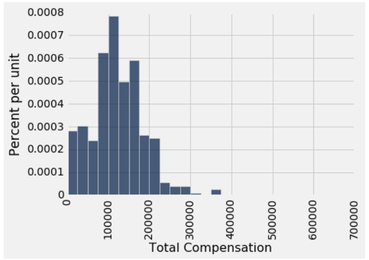
…we can generate lots of resamples

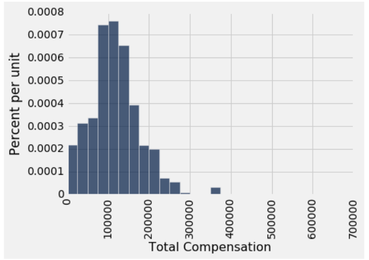
Confidence Interval

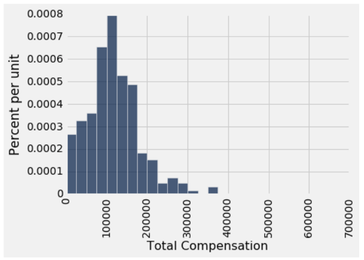
Sample



Resamples







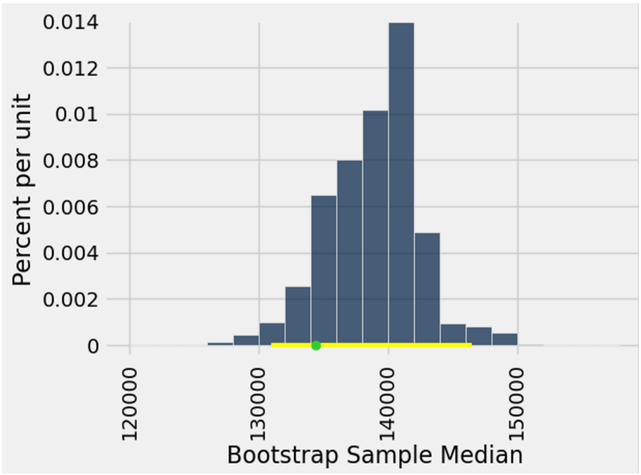
Statistic

median_1

median_2

median_3

Distribution of the Statistic



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 to ensure our two estimates are comparable

The Bootstrap Method

- Note that it is not always true that these two resemble each other
    - But it’s reasonable if the sample is large enough
- Our hope is that variability of the bootstrap estimate and distribution of bootstrap errors are similar to what they are in the real world

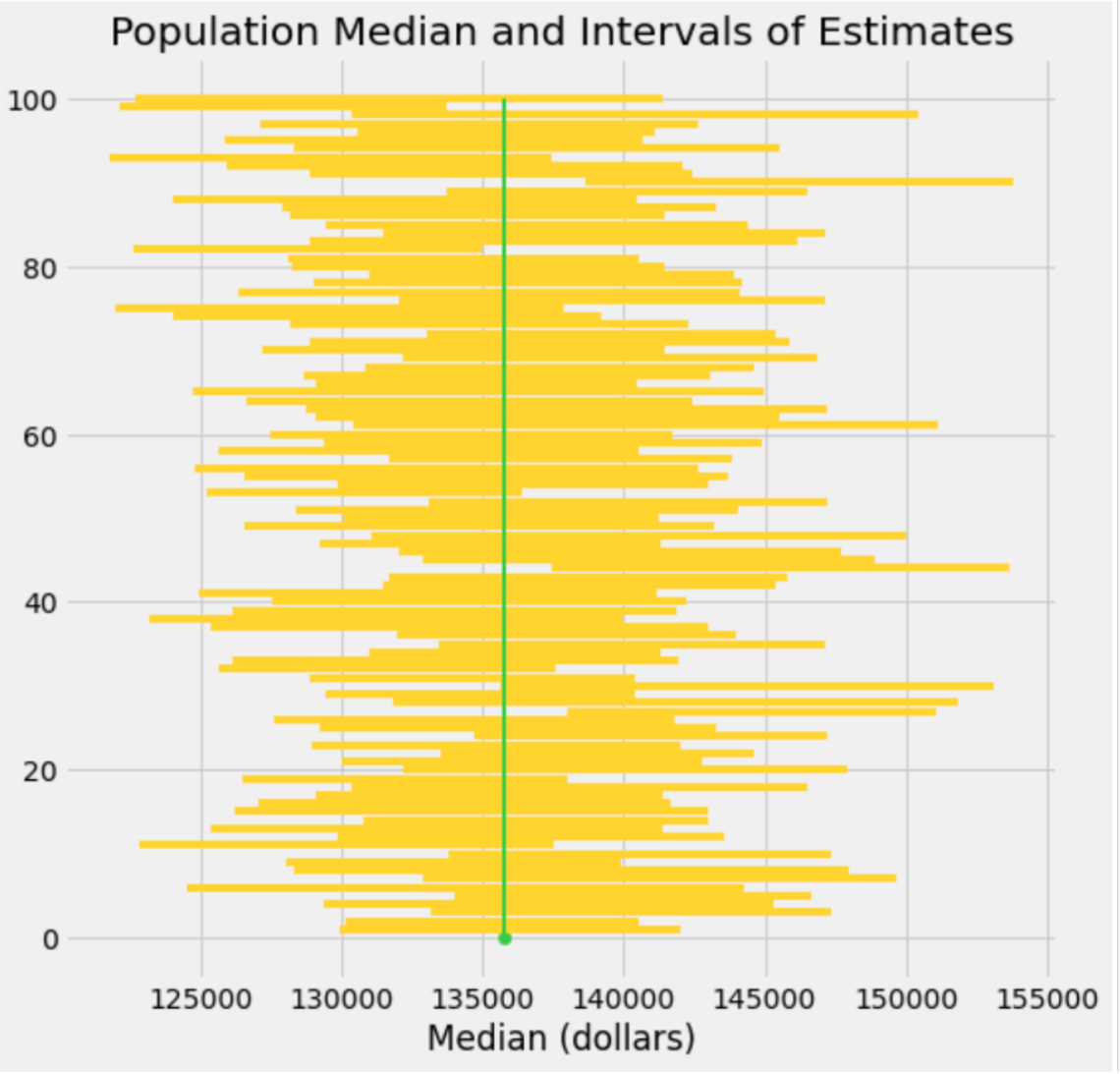
More on Confidence Intervals

95% Confidence Interval

- Interval of estimates of a parameter
- Based on random sampling
  - 95% CI is the middle 95% of the distribution of estimates
- 95% is called the confidence level
- A “good” interval is one that contains the parameter
- The confidence is in the process that creates the interval
  - It generates a “good” interval about 95% of the time

95% Confidence Interval

- Suppose we knew the parameter (green line)
  - Note this is a fixed value
- Each yellow line is a 95% confidence interval based on a fresh sample from the population
- Of the 100 intervals, about 95 of them contain the parameter



Confidence Interval Size

- If we were to create 100 confidence intervals from 100 different random samples, around x of them would contain the population parameter

- Suppose we wanted a higher confidence level. Do we expect the interval to be wider or narrower?

- What's wrong with a large interval?

Using Confidence Intervals for Testing

Estimation vs Hypothesis Testing

- Goal of **estimation**: Infer the value (or range) of an unknown population parameter
    - What fraction of the time does the coin come up with heads?
- Goal of **hypothesis testing**: Infer whether some unknown population parameter is equal to some hypothesis
    - Does the coin come up 50% of the time?
- Both kinds of inferences are made from a random sample
    - Flip the coin 100 times and record the number of heads

Using a CI for Testing

Null hypothesis: Population average $= x$

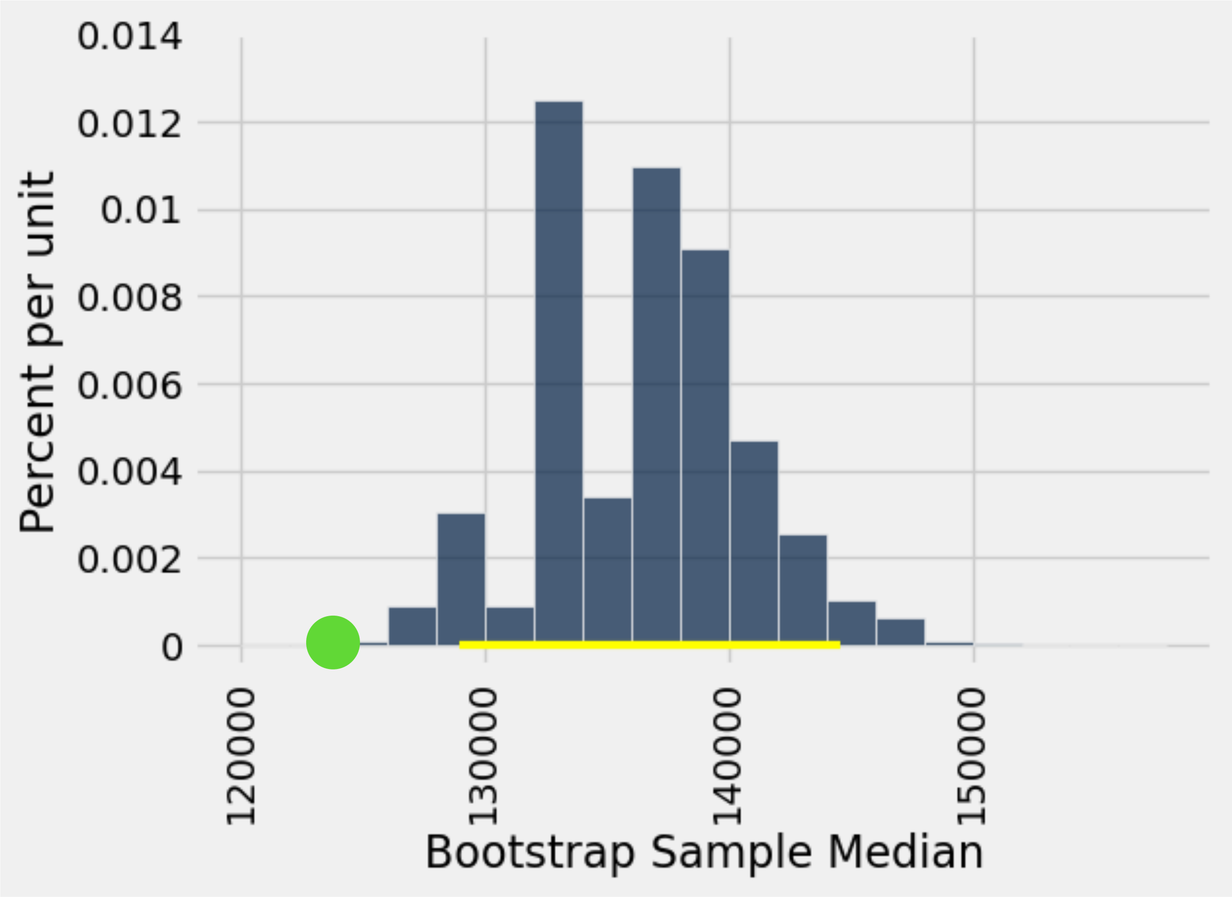
Alternative hypothesis: Population average $\neq x$

Cutoff for p-value: $p\,\%$

Method:

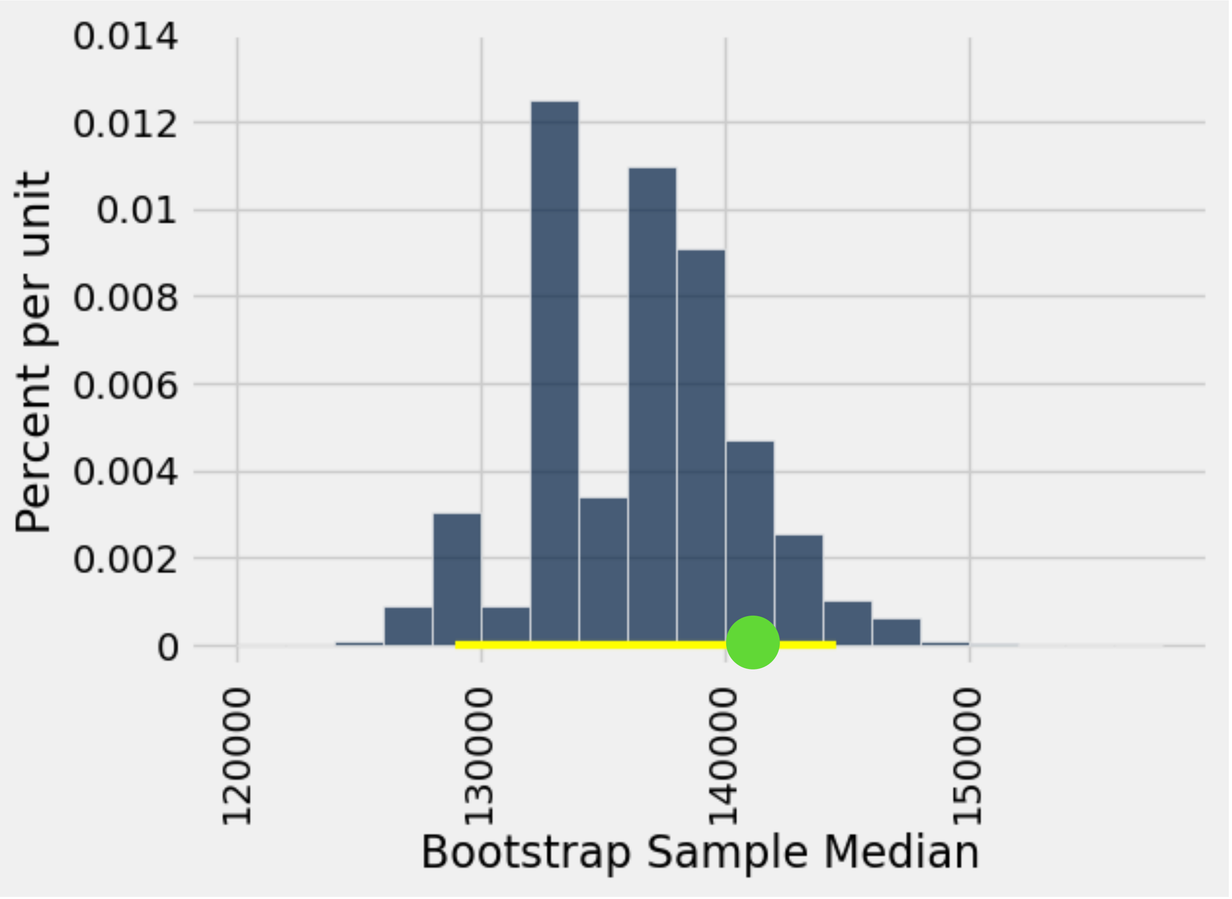
- Construct a $(100 - p)\,\%$ confidence interval for the population average
- If x is not in the interval, reject the null
- If x is in the interval, can’t reject the null

Rejecting the null



Our x is outside the 95% confidence interval

Cannot reject null



Our x is inside the 95% confidence interval

Using Methods Appropriately

When *Not* to Use Bootstrapping

- If you're trying to estimate any parameter that's greatly affected by rare elements of the population
  - Ex: Very high or very low percentiles; min or max
- If the probability distribution of your statistic is not roughly bell shaped
- If the original sample is very small

DEMO

Can you use a CI like this?

Suppose an appropriate 95% confidence interval for the average age of the mothers in a population is (26.9, 27.6)

- True or False: About 95% of the mothers in the population were between 26.9 years and 27.6 years old

Is this what a CI means?

Suppose an appropriate 95% confidence interval for the average age of the mothers in a population is (26.9, 27.6)

- True or False: There is a 95% probability that the average age of mothers in the population is in the range 26.9 years to 27.6 years old

Mean and Median

Review: Averages/Means

The average of [2, 3, 3, 9]:

$$\frac{2+3+3+9}{4} = 4.25$$

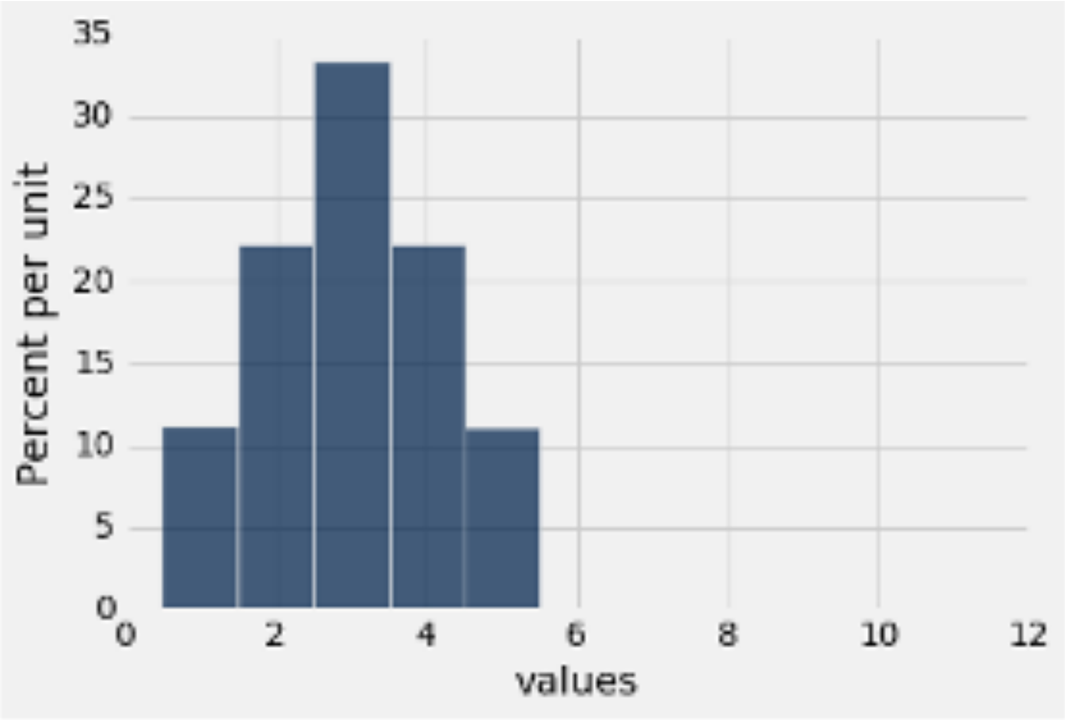
Notice:

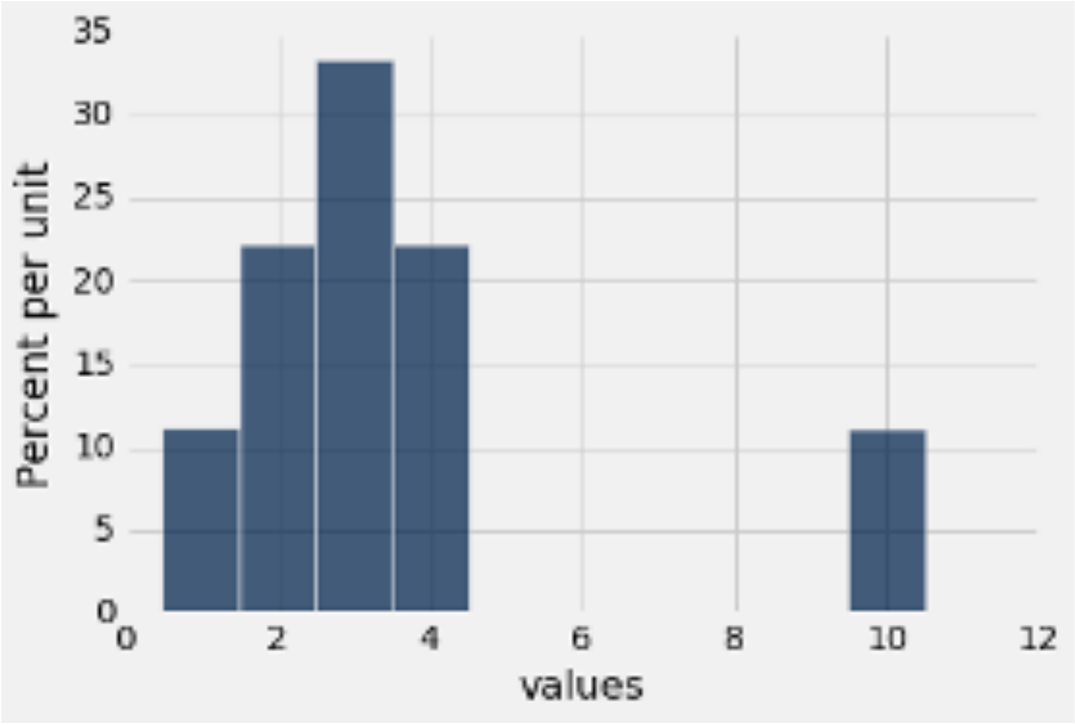
- Need not be a value in the list
- Need not be an integer even if the data consists of integers
- Somewhere between the min and max, but not necessarily halfway between
- Same units as the data

Average and Median

Are the medians of these two the same or different? What about the mean?

- If they’re different, which is bigger?





Comparing Mean and Median

- **Mean**: Balance point of the histogram

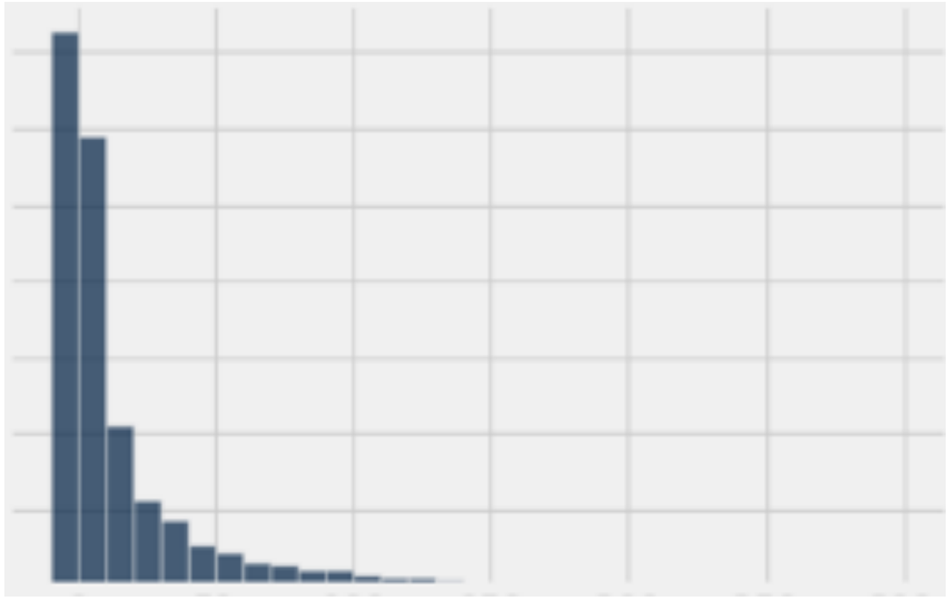
- **Median**: Half-way point of the data. Half of the area of the histogram is on either side of the median

- 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 skew (tail)

Discussion Question

Suppose we're given this distribution



Which of the following is True:

A. Less than 50% of the values are above the average

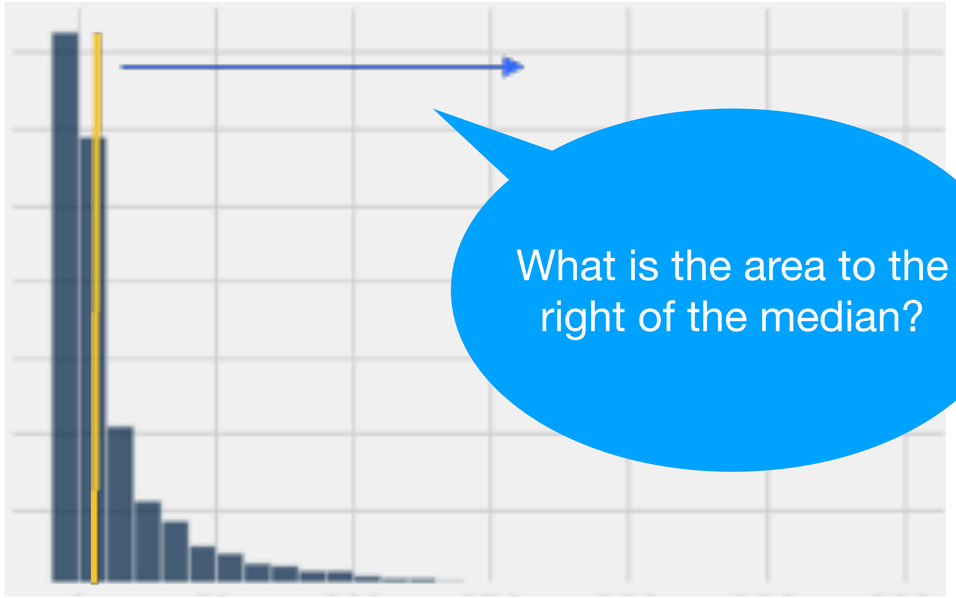
B. 50% of the values are above the average

C. More than 50% of the values are above the average

Discussion Question Hint

Suppose we're given this distribution

The gold line is at the median.



Which of the following is True:

A. Less than 50% of the values are above the average

B. 50% of the values are above the average

C. More than 50% of the values are above the average

Standard Deviation

Defining Variability

Variability: How spread out the distribution is

- How do we define it?

- What if we define it as “biggest value - smallest value”?
  - Doesn’t tell us about the shape of the distribution

- What if we define it as how far a typical observation is from the mean?
  - How do you quantify this?

Deviation from the Average

We can compute the deviation from the average of a value from this list as:

$$\text{deviation} = \text{value} - \text{mean}$$

- If a value is above the mean, the deviation is **positive**
- If a value is below the mean, the deviation is **negative**
- Deviations tell us the direction and size of the difference

How can we use this to define variability?

How to Define Variability?

- To measure for how far the numbers are spread from the mean:
  - Compute the average

Let $\overrightarrow{V}$ be a collection of values and $\mu = \text{avg}\left(\overrightarrow{V}\right)$

How to Define Variability?

- To measure for how far the numbers are spread from the mean:
  - Compute the average
  - Compute each value's deviation from the average

Let $\overrightarrow{V}$ be a collection of values and $\mu = \text{avg}\left(\overrightarrow{V}\right)$

$$v - \mu \quad \text{for } v \in \overrightarrow{V}$$

How to Define Variability?

- To measure for how far the numbers are spread from the mean:
  - Compute the average
  - Compute each value's deviation from the average
  - Square the deviations

Let $\overrightarrow{V}$ be a collection of values and $\mu = \text{avg}\left(\overrightarrow{V}\right)$

$$(v - \mu)^2 \text{ for } v \in \overrightarrow{V}$$

How to Define Variability?

- To measure for how far the numbers are spread from the mean:
  - Compute the average
  - Compute each value's deviation from the average
  - Square the deviations
  - Compute the mean of the these squared deviations

Let $\overrightarrow{V}$ be a collection of values and $\mu = \text{avg}\left(\overrightarrow{V}\right)$

$$\text{avg}\left((v-\mu)^2 \text{ for } v \in \overrightarrow{V}\right)$$

How to Define Variability?

- To measure for how far the numbers are spread from the mean:
  - Compute the average
  - Compute each value's deviation from the average
  - Square the deviations
  - Compute the mean of the these squared deviations

Let $\overrightarrow{V}$ be a collection of values and $\mu = \text{avg}\left(\overrightarrow{V}\right)$

Variance of $\overrightarrow{V}$
$= \text{avg}\left((v-\mu)^2 \text{ for } v \in \overrightarrow{V}\right)$

Standard Deviation (SD)

Standard deviation is the root mean square of deviations from the average

$$\sigma = \sqrt{\text{avg}\left((v-\mu)^2 \text{ for } v \in \vec{V}\right)}$$

Why we like standard deviation:

- No matter the shape of the distribution, the bulk of the data is in the range “average plus or minus a few standard deviations”

- It has a nice relation with the bellcurve (to be discussed later)

Chebyshev's Inequality

Chebyshev's Inequality: *No matter what the shape of the distribution*, the proportion of values in the range "$\text{average} \pm z$ SDs" is **at least** $1 - \frac{1}{z^2}$

- Note this is a lower bound, not an exact answer
  - The proportion of entries within the range "$\text{average} \pm z$ SDs" could be much larger than $1 - \frac{1}{z^2}$, but it can't be smaller

Chebyshev's Bounds

Range	Proportion
average $\pm$ 2 SDs	at least $1 - \frac{1}{4} = 75\,\%$
average $\pm$ 3 SDs	at least $1 - \frac{1}{9} \approx 89\,\%$
average $\pm$ 4 SDs	at least $1 - \frac{1}{16} = 93.75\,\%$
average $\pm$ 5 SDs	at least $1 - \frac{1}{25} = 96\,\%$

True no matter what the distribution looks like!

Standard Units

Standard Units

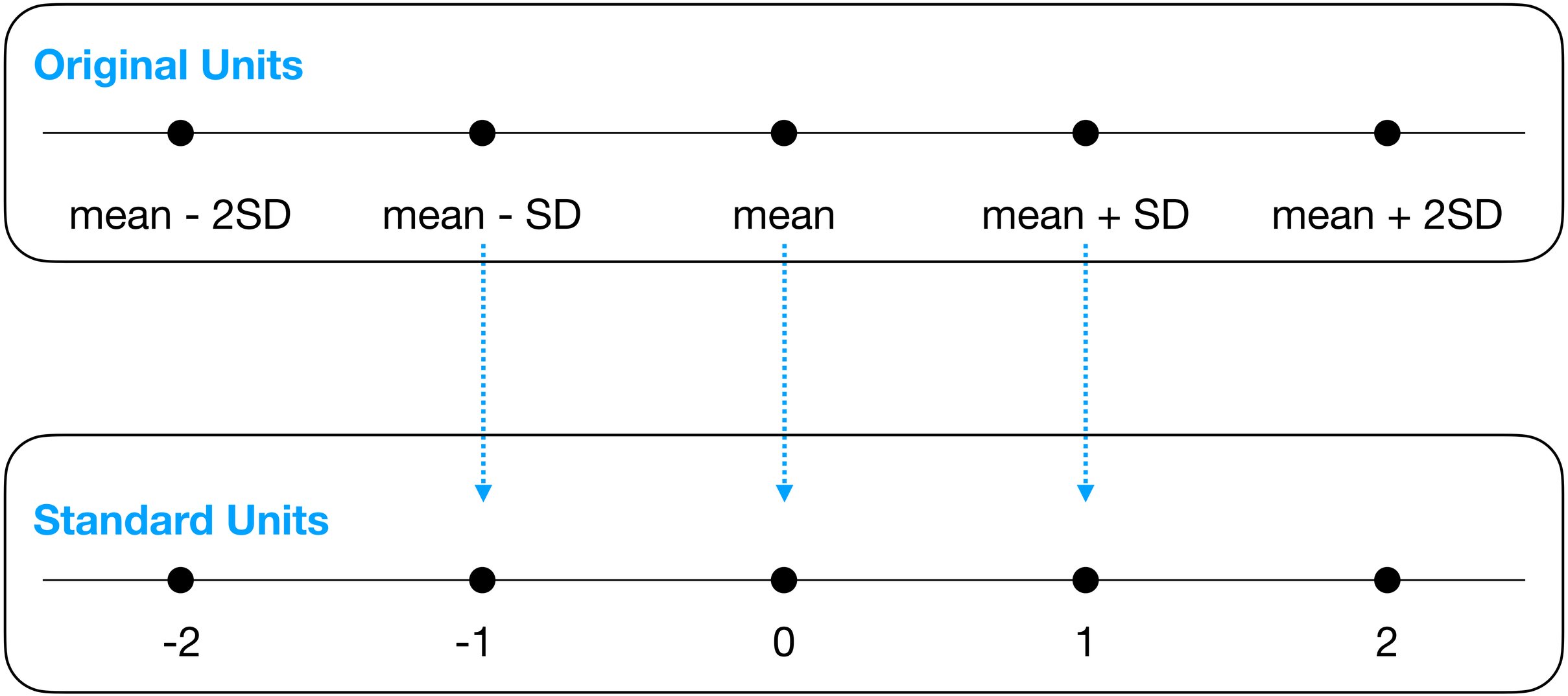
- The quantity z (from "$\text{average} \pm z\ \text{SDs}$" in Chebychev's inequality) measures **standard units**

  - **Standard units** is the number of standard deviations away from the average

- To convert a value (v) to standard units, compare the deviation from the average (μ) with the standard deviation (SD):

$$z = \frac{v - \mu}{\text{SD}}$$

Converting to Standard Units



Interpreting Standard Units

$$z = \frac{v - \mu}{\text{SD}}$$

- When z is **negative**, the value v is **below** average
- When z is **positive**, the value v is **above** average
- When z is **0**, the value v **is** the average

When values are in standard units, average = 0, SD = 1

Discussion

What whole numbers are closest to:

A. Average age?

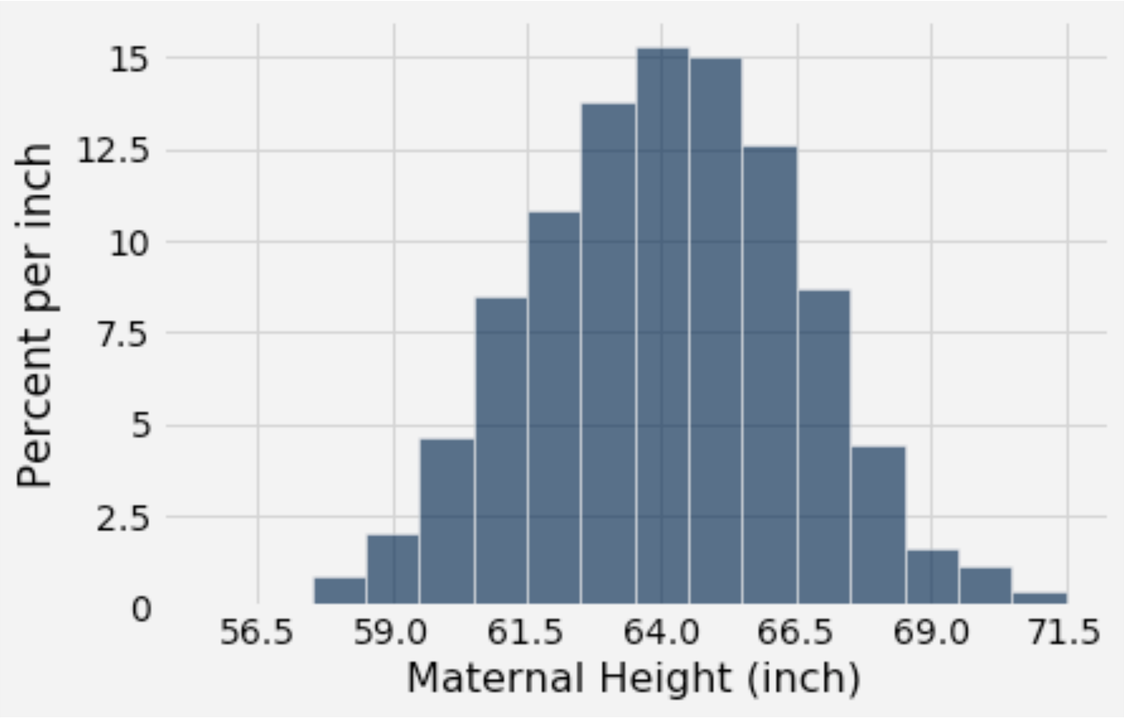
B. The SD of ages?

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

Standard Deviation and Histograms

- Usually it’s not easy to estimate the standard deviation by looking at a histogram

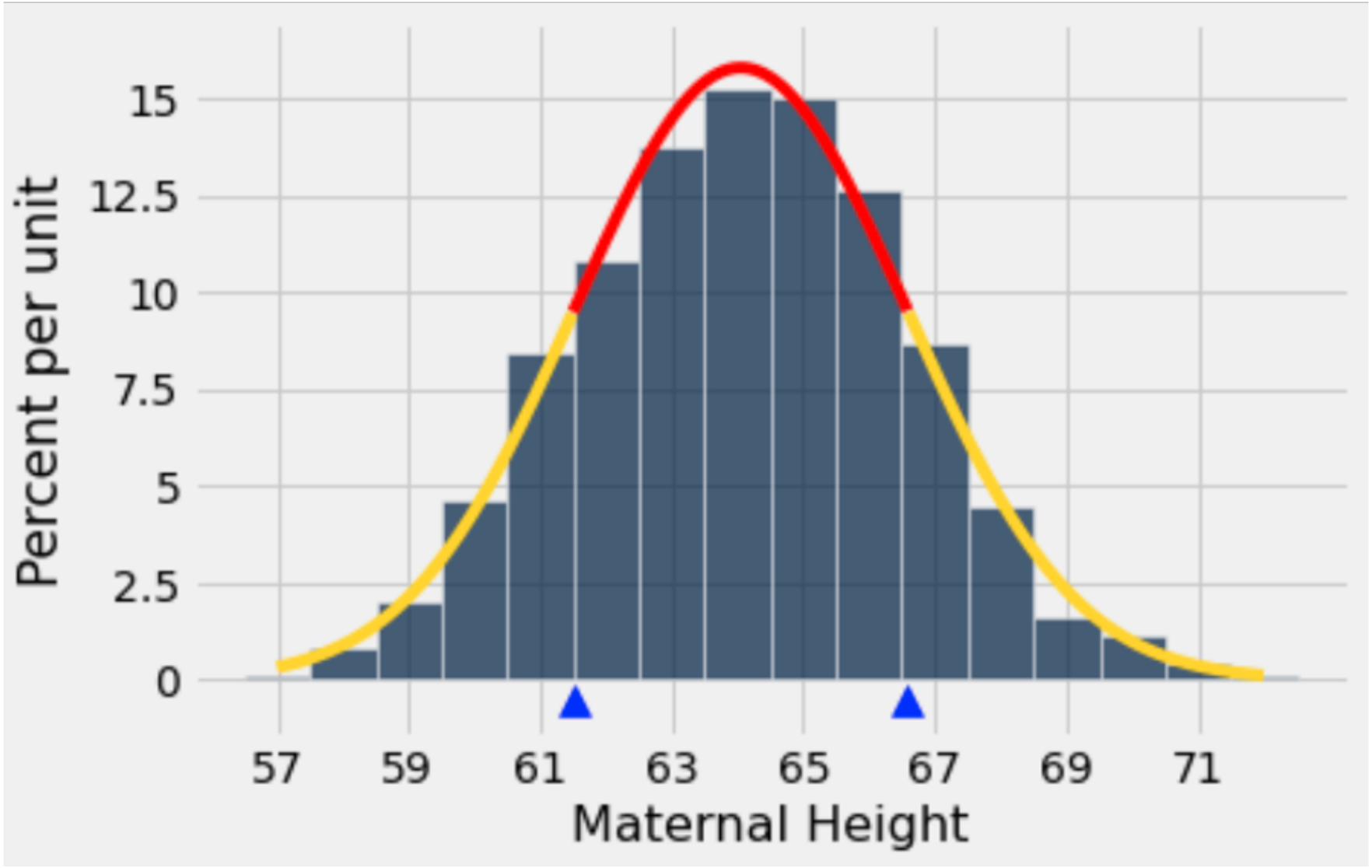
  - … But you can if the histogram has a bell shape!



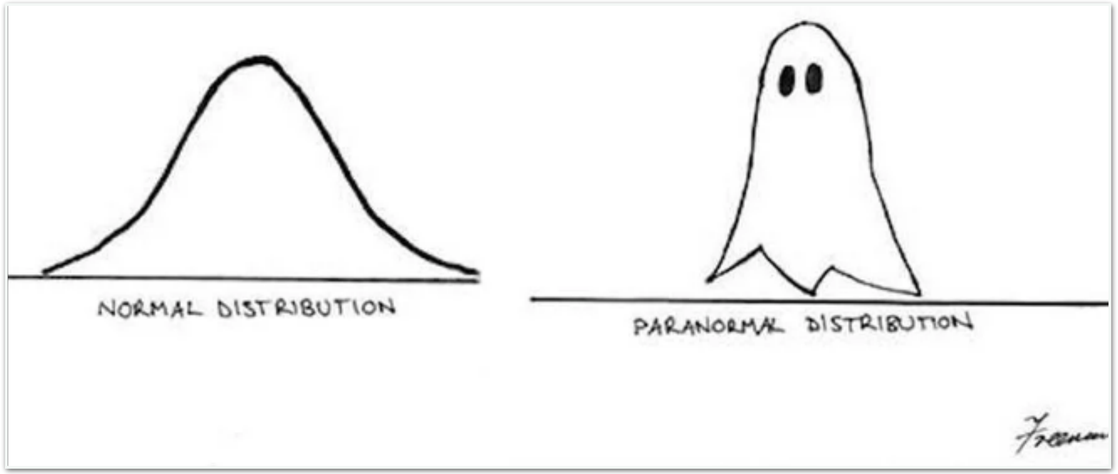
Standard Deviation and Histograms

- If the 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



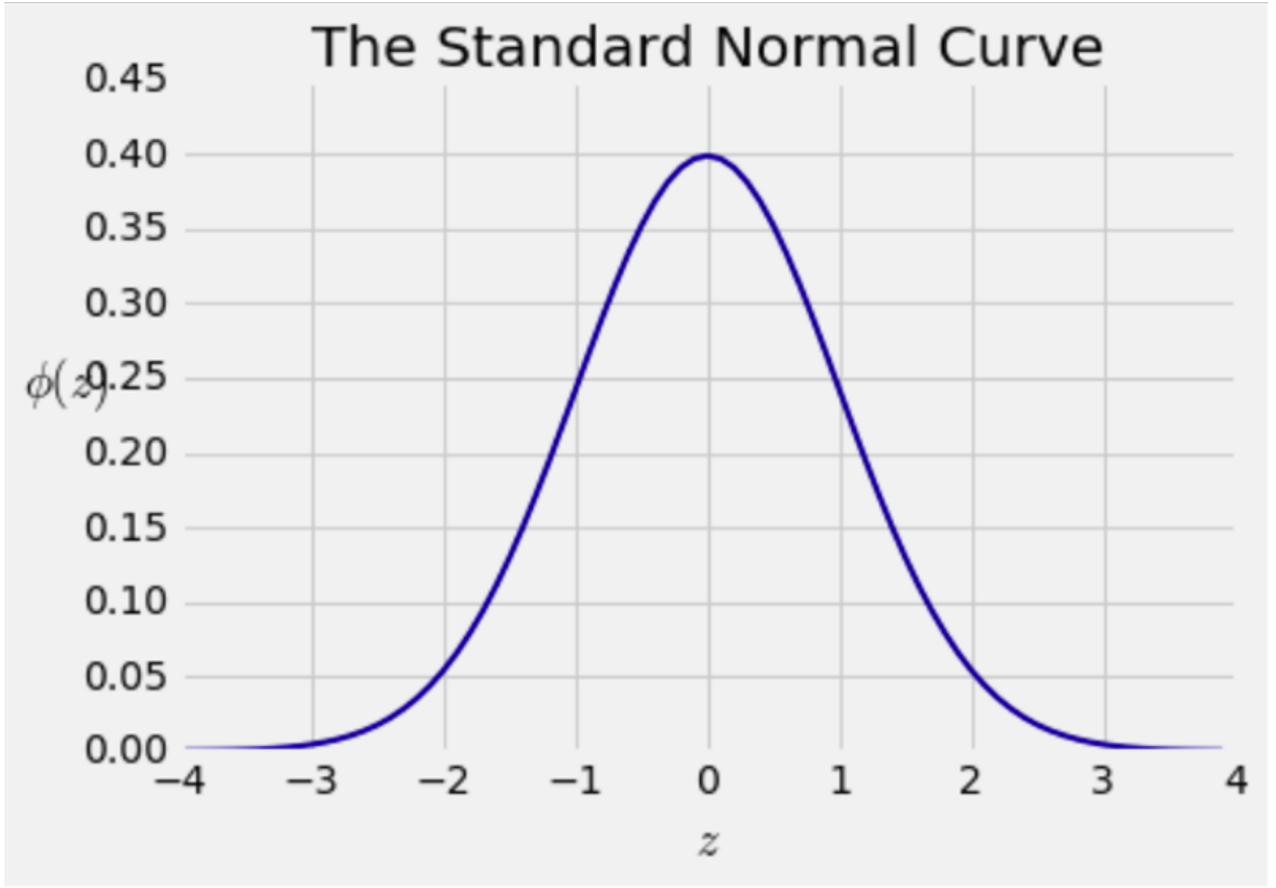
Normal Distribution



The Standard Normal Curve

Here's a beautiful formula we won't use at all:

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



How Big are Most of the Values?

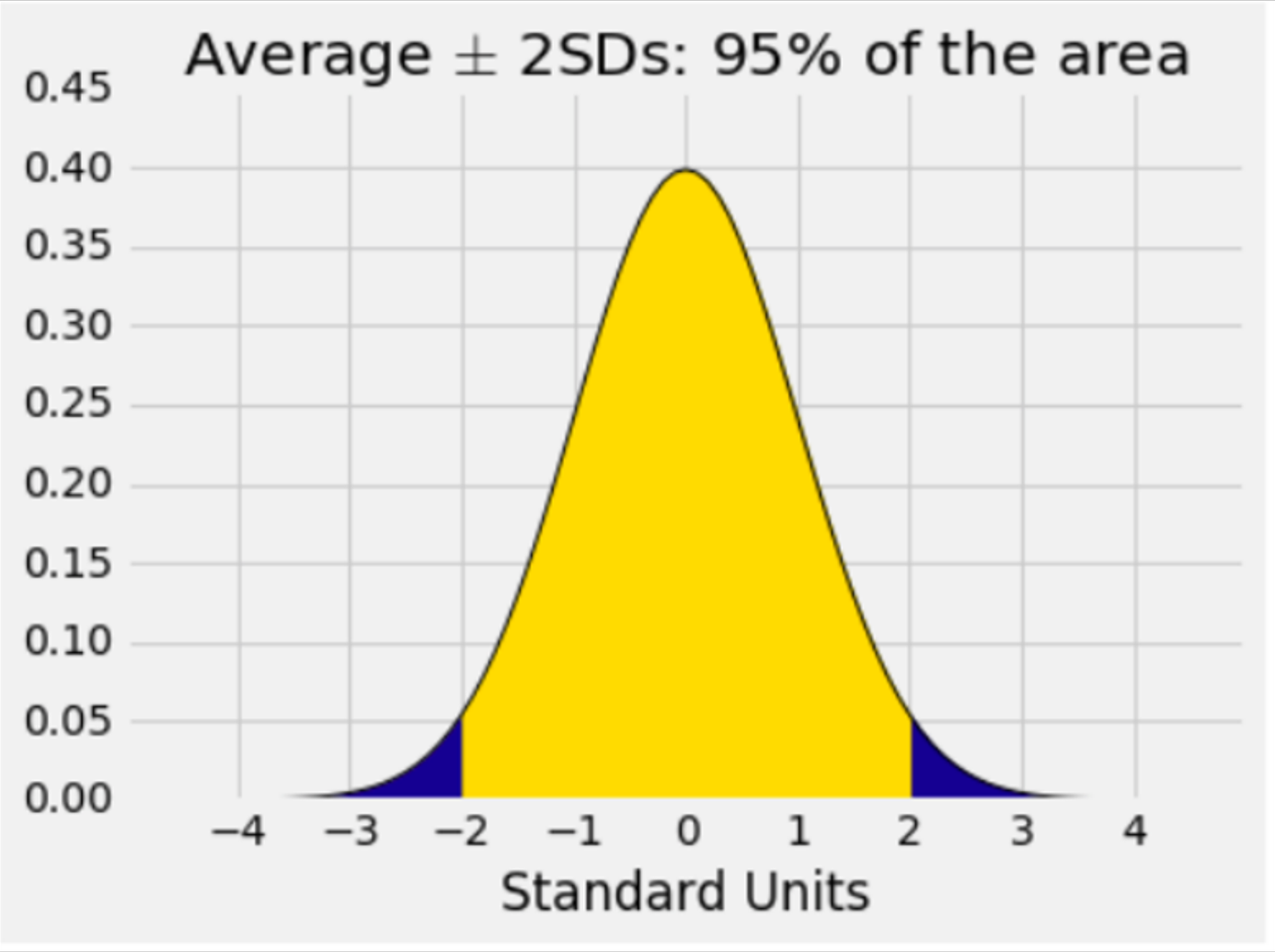
No matter the shape of the distribution, the bulk of the data is in the range "average ± a few SDs"

If the histogram is bell-shaped, then almost all of the data is in the range "average ± 3 SDs"

Normal vs All Distributions

Range	All Distributions (Chebyshev's)	Normal Distribution
mean $\pm$ 1 SDs	At least 0%	At least 68%
mean $\pm$ 2 SDs	At least 75%	At least 95%
mean $\pm$ 3 SDs	At least 89%	At least 99%

A “Central” Area



Central Limit Theorem

Sample Averages

- Central Limit Theorem describes how a normal distribution (bell-shape curve) is connected to random sample averages

- We care about sample averages because they can help us estimate population averages

Central Limit Theorem

Definition:

Is a sample is large and drawn at random with replacement

Then regardless of the distribution,

The **probability distribution of the sample average** is roughly normal

Discussion

After rolling 1,000,000 fair 6-sided dice, which of the following histograms do you expect to have a bell shape?

A. The histogram of outcomes of these million rolls

B. The histogram that results from computing the average outcome of these million rolls

C. The histogram that results from splitting the outcomes into 1,000 groups of 1,000 (in the order they occurred) and computing the average outcome of each group

Next Class

- Today
  - More on Confidence Intervals
  - Standard Deviation
  - Normal Distribution
- Wednesday
  - Central Limit Theorem