COMS BC1016
Introduction to Computational Thinking and Data Science

Lecture 18: Standard Deviation and the Normal Distribution

April 6, 2026

Data Science in this course

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

11	4/06	Inference		HW 6 - Confidence Intervals
	4/08		Lab 8 - Normal Distribution and Variability of Sample Means	HW 5 Due 11:59PM
12	4/13	Prediction		HW 7 - Sample Sizes and Confidence Intervals
	4/15		Lab 9 - Regression	HW 6 Due 11:59PM **Final Project Proposals** Due Friday, 4/17
13	4/20	Special Topics: Data Ethics (Computing Fellows Workshop)		HW 8 - Linear Regression
	4/22	Special Topics: Bias in AI (Guest lecture by Murad Megjhani)	Lab: Final Project Work Time	HW 7 Due 11:59PM
14	4/27	Prediction		**Final Project Progress Reports** HW 9 - Regression Inference (Optional, Due 5/4)
	4/29	(Either special topics or prediction)	Lab: Final Project Consultations	HW 8 Due 11:59PM
15	5/04	Wrap up!		HW 9 Due 11:59PM (Optional) **Final Project Report** Due Friday 5/8

11	4/06	Inference		HW 6 - Confidence Intervals
	4/08		Lab 8 - Normal Distribution and Variability of Sample Means	HW 5 Due 11:59PM
12	4/13	Prediction		HW 7 - Sample Sizes and Confidence Intervals
	4/15		Lab 9 - Regression	HW 6 Due 11:59PM **Final Project Proposals** Due Friday, 4/17
13	4/20	Special Topics: Data Ethics (Computing Fellows Workshop)		HW 8 - Linear Regression
	4/22	Special Topics: Bias in AI (Guest lecture by Murad Megjhani)	Lab: Final Project Work Time	HW 7 Due 11:59PM
14	4/27	Prediction		**Final Project Progress Reports** HW 9 - Regression Inference (Optional, Due 5/4)
	4/29	(Either special topics or prediction)	Lab: Final Project Consultations	HW 8 Due 11:59PM
15	5/04	Wrap up!		HW 9 Due 11:59PM (Optional) **Final Project Report** Due Friday 5/8

Final Projects Data

- Each group will choose a dataset to analyze
    - Each dataset has a starter notebook with a brief overview of the data + an outline of the final report
    - If you would like to analyze different data than the ones we provided, you must get instructor approval!

Final Project Proposal

- First milestone due next week Friday (April 17)

- You should complete the exploratory data analysis and have a clearly stated hypothesis question (2 hypothesis questions for groups of 3)

- Exploratory Data Analysis:

  - To help decide what question(s) you want to ask, you should explore the data! Are there trends you notice? Relationships you're curious about?

  - In this section you should include 1 quantitative plot, 1 qualitiative plot, a table using an aggregate function, and a table made as a result of the join

Final Project Proposal

- Proposals will be graded based on completeness and readability

- Proposals are intended to get you started thinking about what you want to do, but it's not a hard commitment to your hypothesis question (it is a commitment to the data set though)

  - You are allowed to change your hypothesis question later on

  - You also do not need to keep the same tables/charts in your final report (but you'll probably want to… It'll save you work)

- **Proposals should be submitted as a group via Gradescope**

Today’s Lecture

- Bootstrapping
- Confidence Intervals
- Standard Deviation
- Normal Distribution

Review: Bootstrap Method

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
  - To resample, draw at random with replacement the same number of times as the original sample size

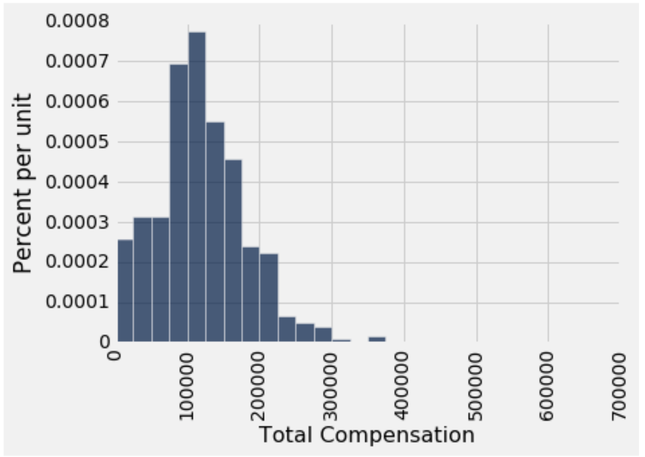
The Bootstrap

Population

?

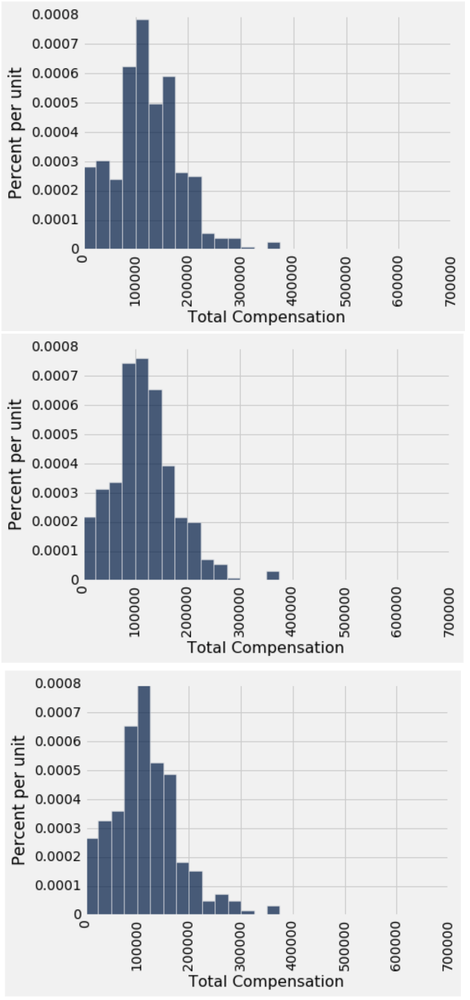
We don’t know the entire population and thus can’t calculate the parameter directly

Sample



However, we can take a single sample…

Resamples



…and generate lots of resamples

The Bootstrap Method

- Important to resample with replacement the same number of times as the original sample

  - Suppose we computed the original statistic based on n samples. We need to compare it to another statistic also based on n samples

  - Drawing without replacement gives the same sample back, so you need to sample with replacement

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

When *not* to use the Bootstrap

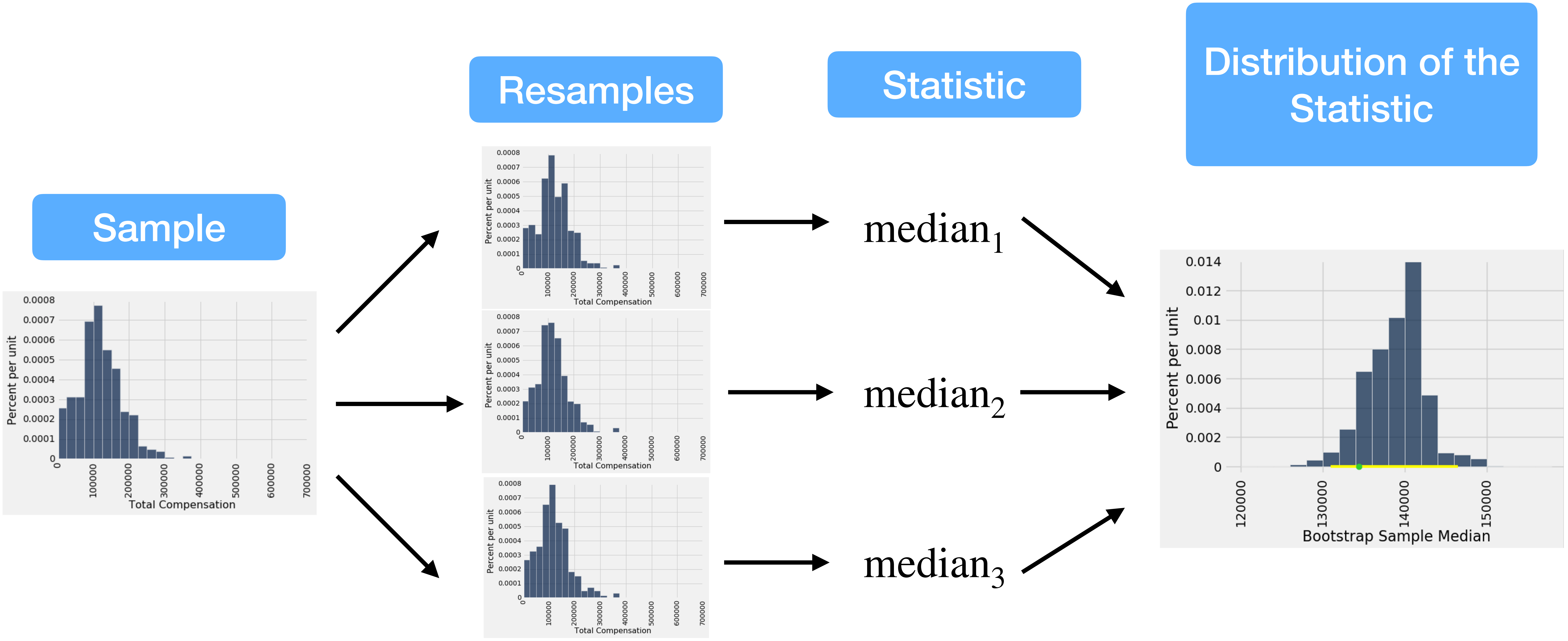
- If you're trying to estimate very high or very low percentiles
- If you're trying to estimate any parameter that's greatly affected by rare elements of the population (e.g., min or max)
- If the probability distribution of your statistic is not roughly bell shaped
    - The shape of the empirical distribution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

Confidence Intervals

Confidence Interval

- Interval of **estimates of a parameter**
  - How confident we are that the parameter (the real value calculated from the population) is likely to be within this interval
    - Good if the parameter is within the interval, bad if it’s not
- The **confidence is in the process** that gives the interval
  - It generates a “good” interval about 95% of the time

Confidence Interval



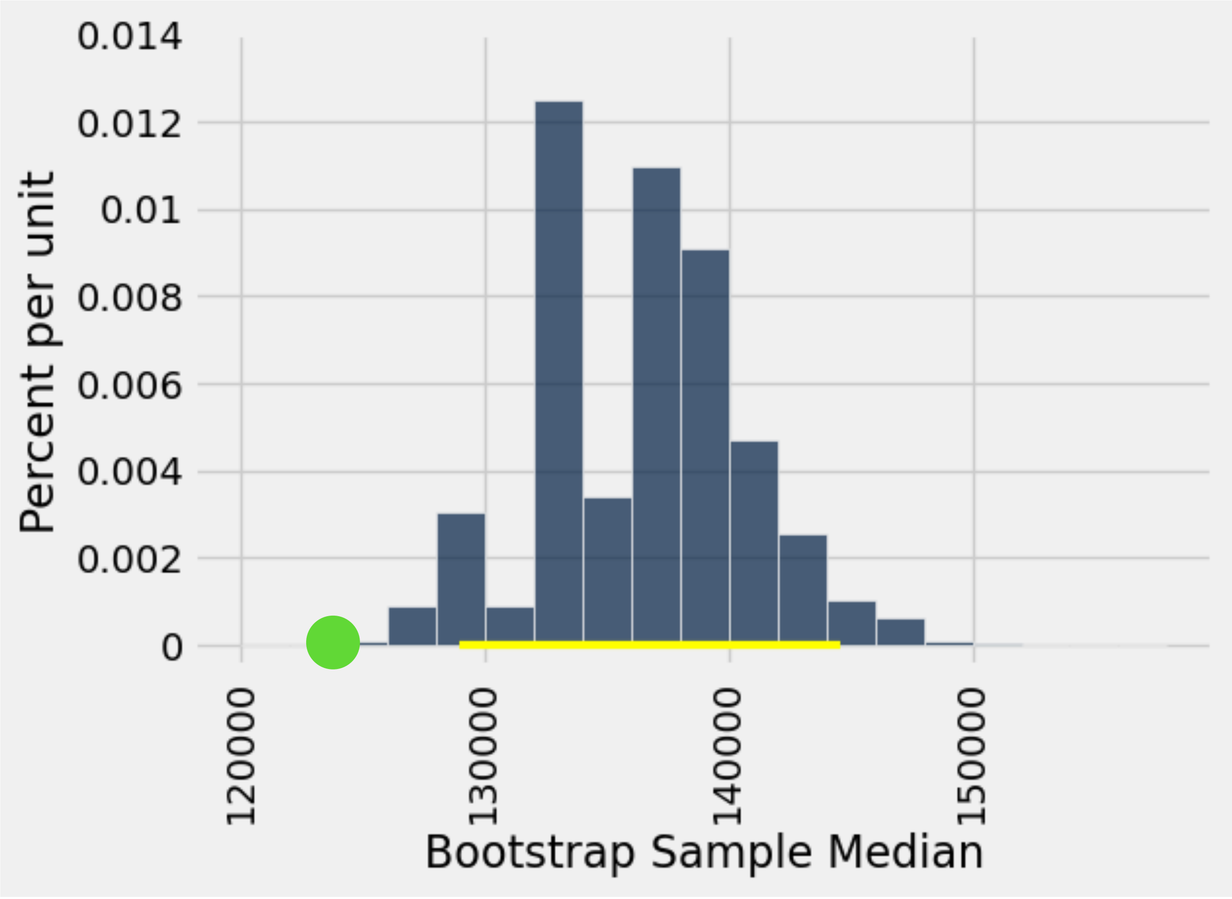
Relationship between Confidence Interval and P-value

- For a p-value cutoff of $p\%$, we reject the null hypothesis if the value is not within $(100 - p)\%$ confidence interval
  - If you use a $p\%$ cutoff for the p-value and the null hypothesis is true, then there is about a $p\%$ chance your test will conclude the alternative is true
  - $(100 - p)\%$ confidence interval says we're $(100 - p)\%$ certain the parameter is somewhere within the interval
    - If the value is outside of the interval, we reject the null hypothesis

Relationship between Confidence Interval and P-value

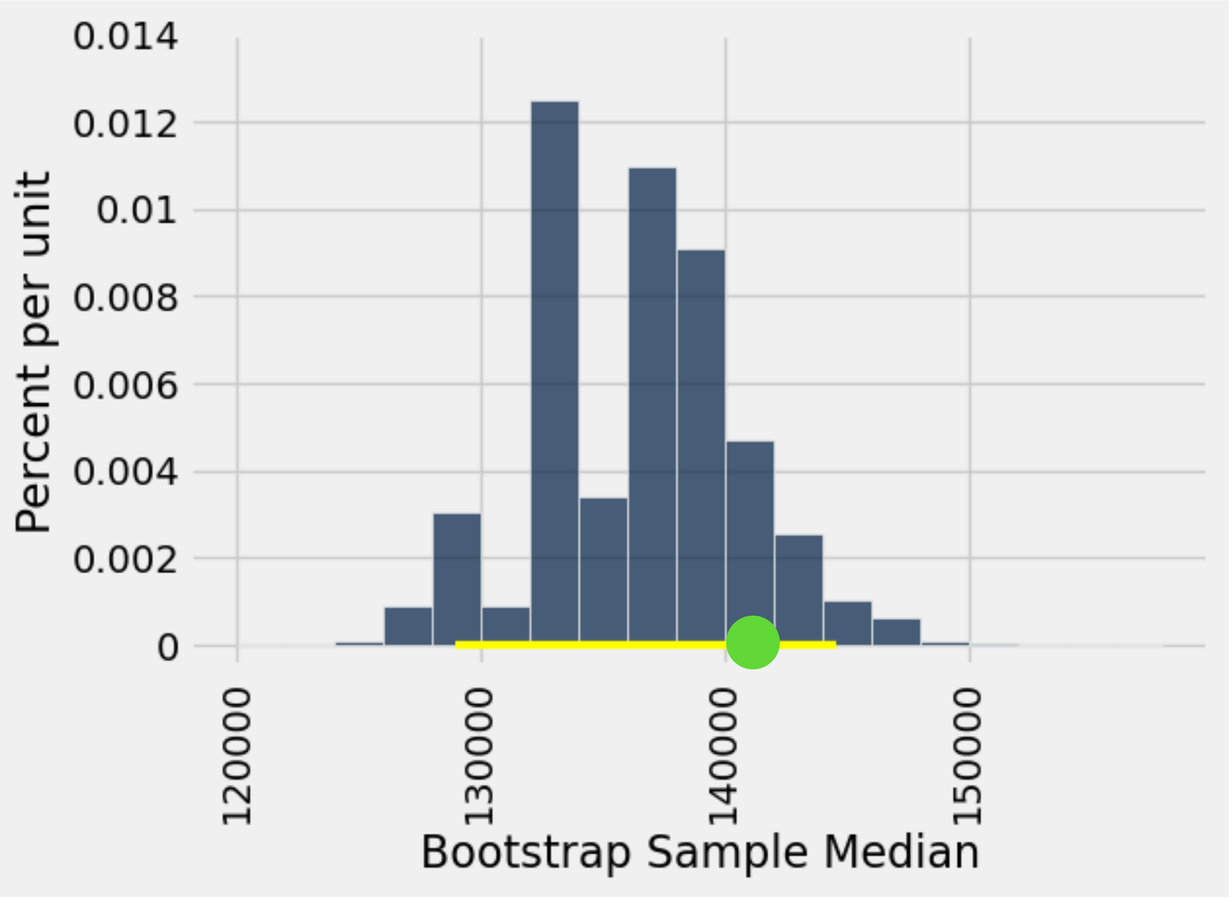
- Example:
  - Null hypothesis: Population average $= x$
  - Alternative Hypothesis: Population average $\neq x$
  - If x is not in our $(100 - p)\%$ interval, then we reject the null

Rejecting the null



Our x is outside the 95% confidence interval

Cannot reject null



Our x is inside the 95% confidence interval

Confidence Interval & P-Value Example

Null Hypothesis: You have a fair coin with 50% probability of getting heads or tails

Alternative: The coin is biased

Your observed value for **% heads** is 65%

Let's say your 95% confidence interval is [45, 60]

Confidence Interval & P-Value Example

Null Hypothesis: You have a fair coin with 50% probability of getting heads or tails

Alternative: The coin is biased

Your observed value for **% heads** is 65%

Let's say your 95% confidence interval is [45, 60]

Questions:

1. For a 5% p-value cutoff, can we reject the null?

2. For a 10% p-value cutoff, can we reject the null?

Confidence Interval & P-Value Example

Null Hypothesis: You have a fair coin with 50% probability of getting heads or tails

Alternative: The coin is biased

Your observed value for **% heads** is 65%

Let's say your 95% confidence interval is [45, 60]

Questions:

1. For a 5% p-value cutoff, can we reject the null?
   - Yes, 65% is outside our confidence interval
2. For a 10% p-value cutoff, can we reject the null?

Confidence Interval & P-Value Example

Null Hypothesis: You have a fair coin with 50% probability of getting heads or tails

Alternative: The coin is biased

Your observed value for **% heads** is 65%

Let's say your 95% confidence interval is [45, 60]

Questions:

1. For a 5% p-value cutoff, can we reject the null?
   - Yes, 65% is outside our confidence interval
2. For a 10% p-value cutoff, can we reject the null?
   - Yes, we expect the confidence interval to be even narrower, so 65% would still be outside the confidence interval

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

Can you use a confidence interval like this?

Suppose our 95% confidence interval for the average age of mothers is 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.

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

Can you use a confidence interval like this?

Suppose our 95% confidence interval for the average age of mothers is 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.

    - **False**. We are estimating the **average age** is in this interval

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

Can you use a confidence interval like this?

Suppose our 95% confidence interval for the average age of mothers is 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.

    - **False**. We are estimating the **average age** is in this interval

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

    - **False**. The average age is unknown but **constant**. It is not random.

Average and Histograms

Review: Averages/Means

Computing the average of [2, 3, 3, 9]:

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

Notice:

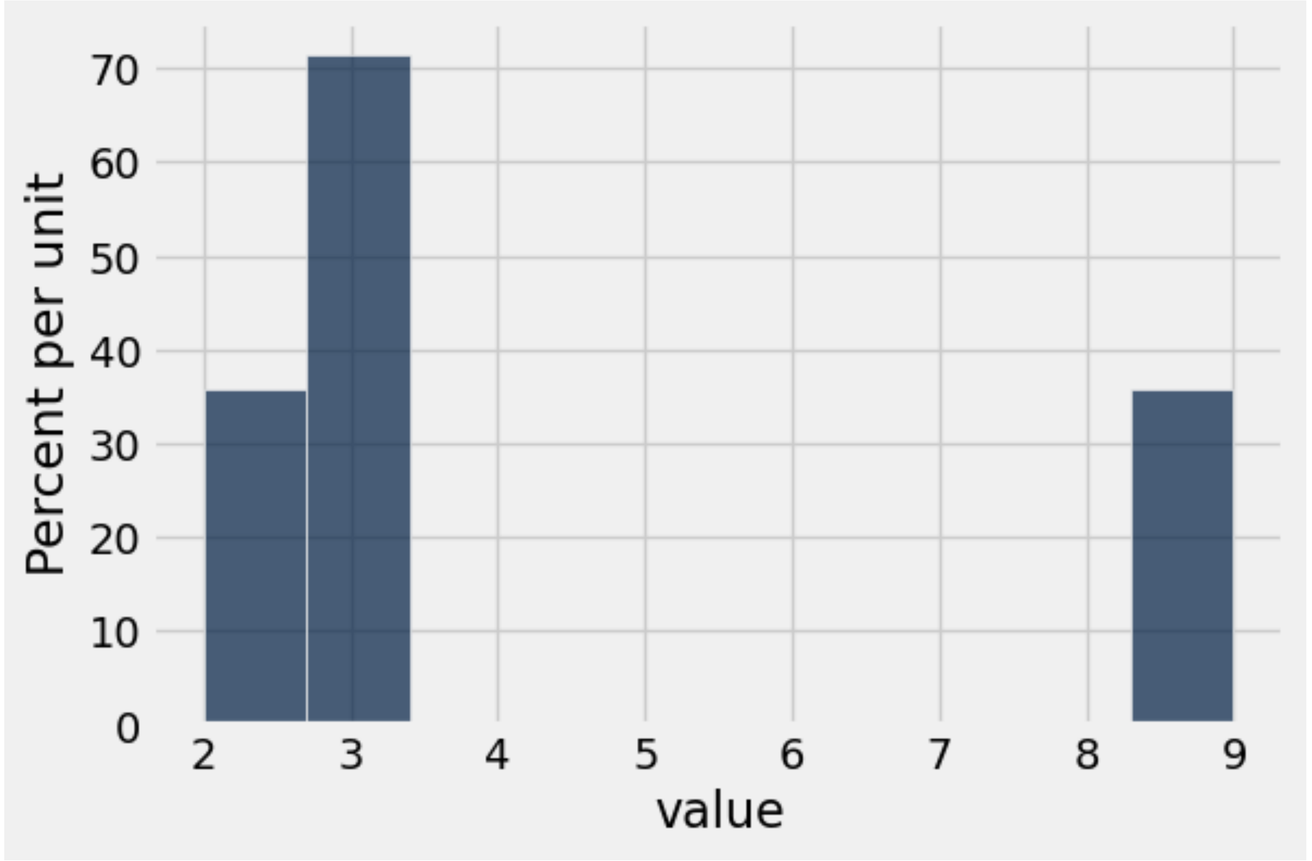
- Need not be a value in the list
- Need not be an integer even if the data consists of integers
- Somewhere between the min and max, but not necessarily the halfway between min and max
- Same units as the data
- Smoothing operator: collects all contributions in a big pot and splits evenly

Relation to Histograms

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

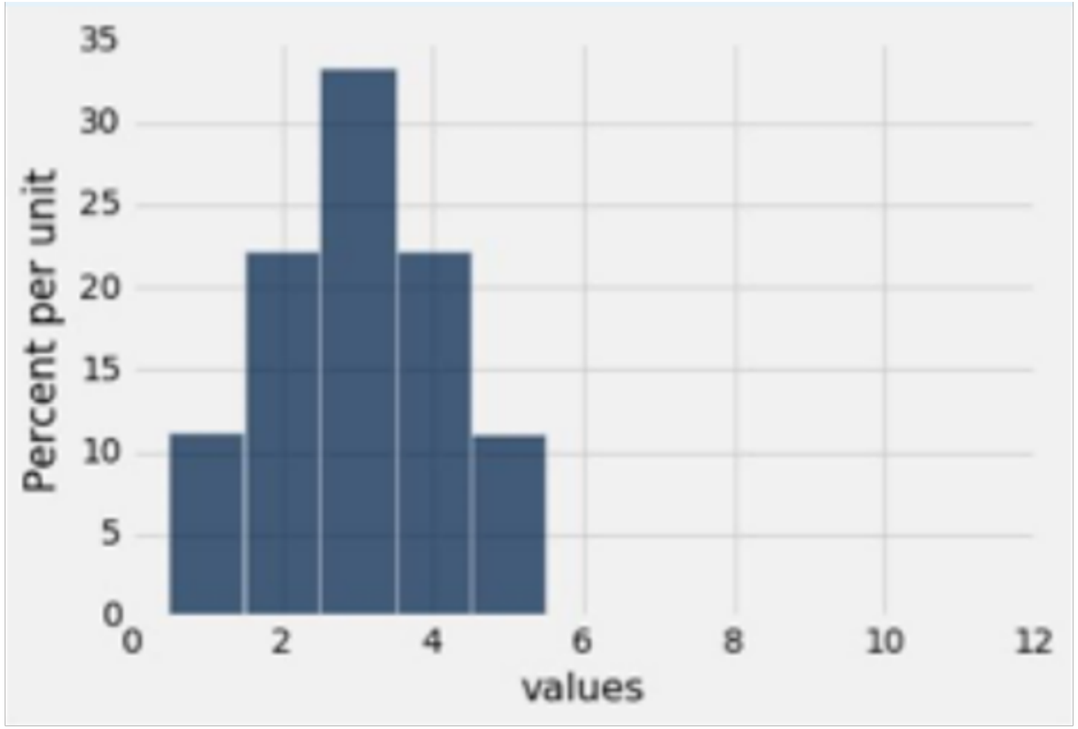
- The average is the **center of gravity** of the histogram

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



Average and Median

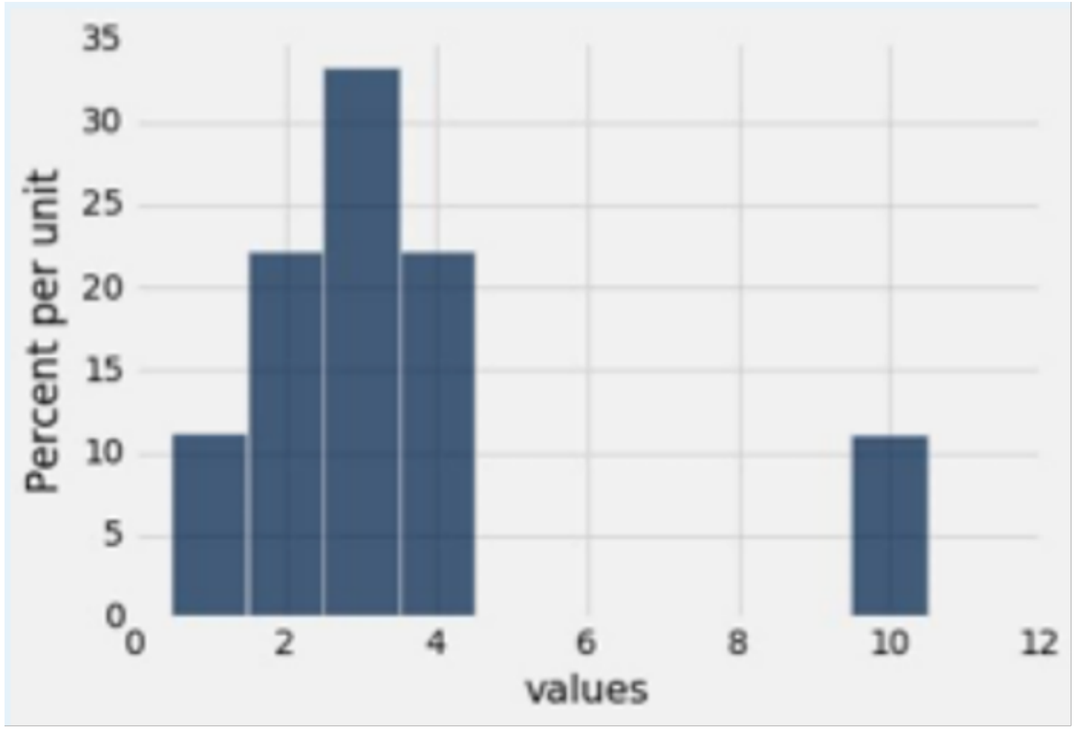
- [1,2,2,3,3,3,4,4,5]
- What is the average?
  - 3
- What is the median?
  - 3

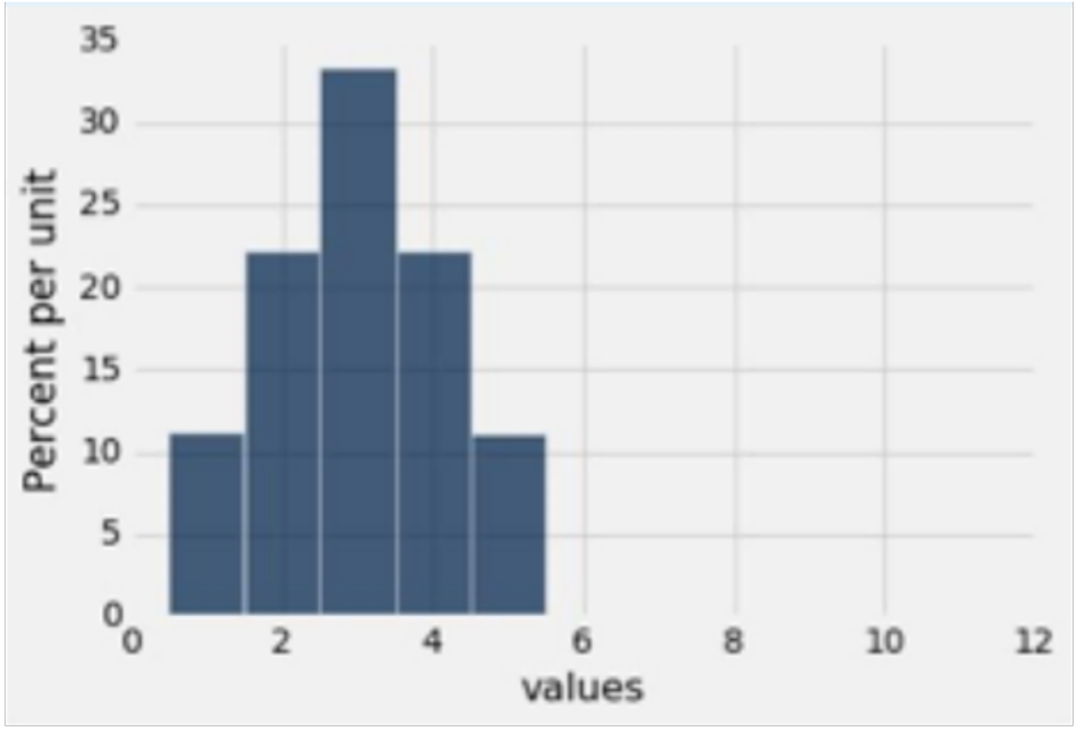


Average and Median

- Are the medians of these two the same or different?

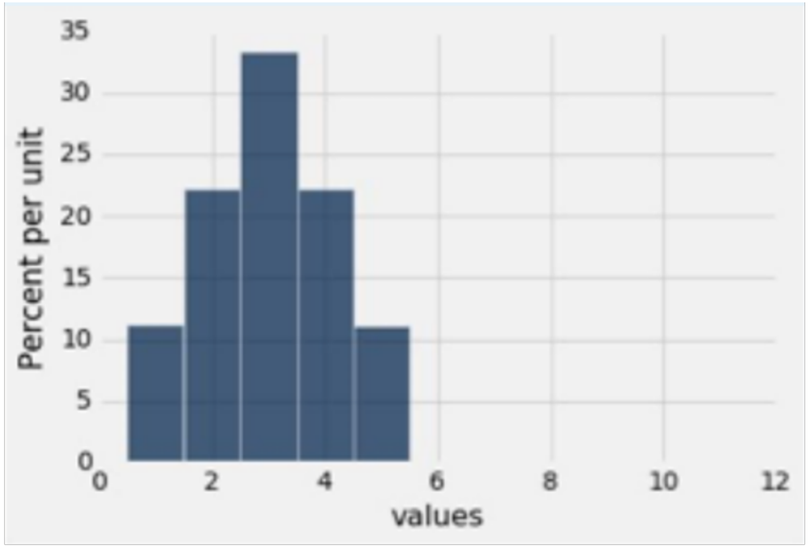
- Are the means the same or different?



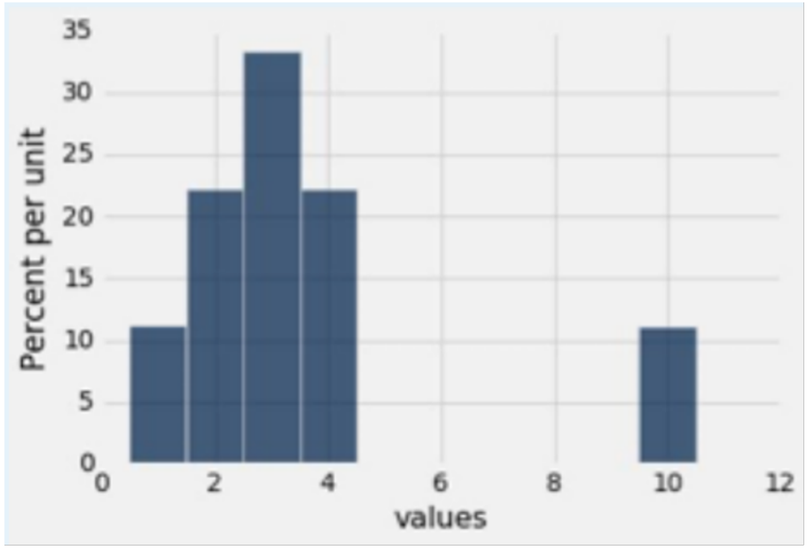


Average and Median

- List 1: [1,2,2,3,3,3,4,4,5]
  - Median =
  - Average =

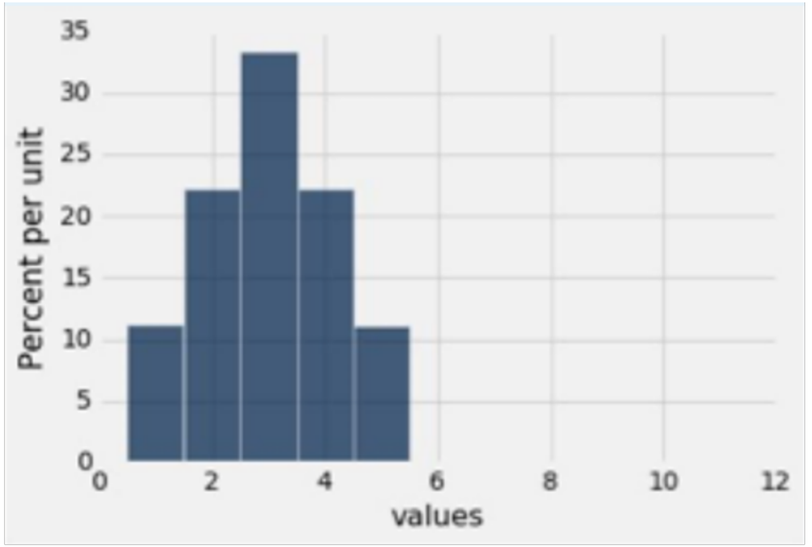


- List 2: [1,2,2,3,3,3,4,4,10]
  - Median =
  - Average =

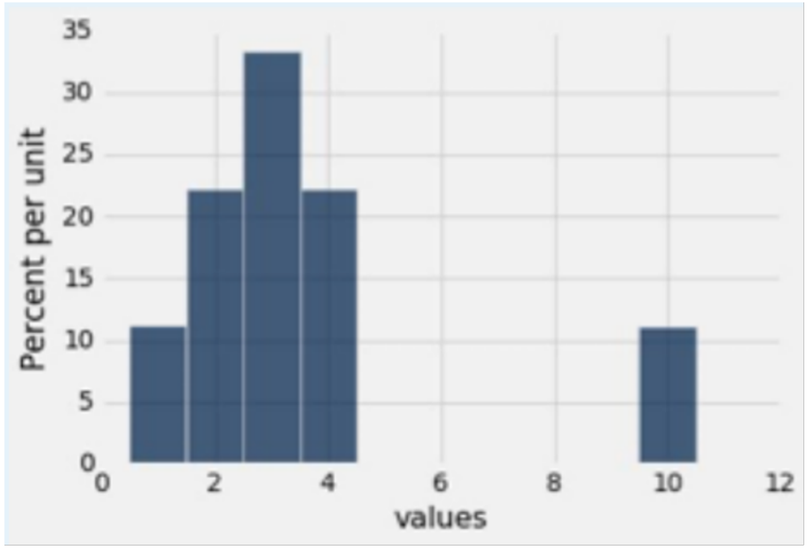


Average and Median

- List 1: [1,2,2,3,3,3,4,4,5]
  - Median = 3
  - Average =

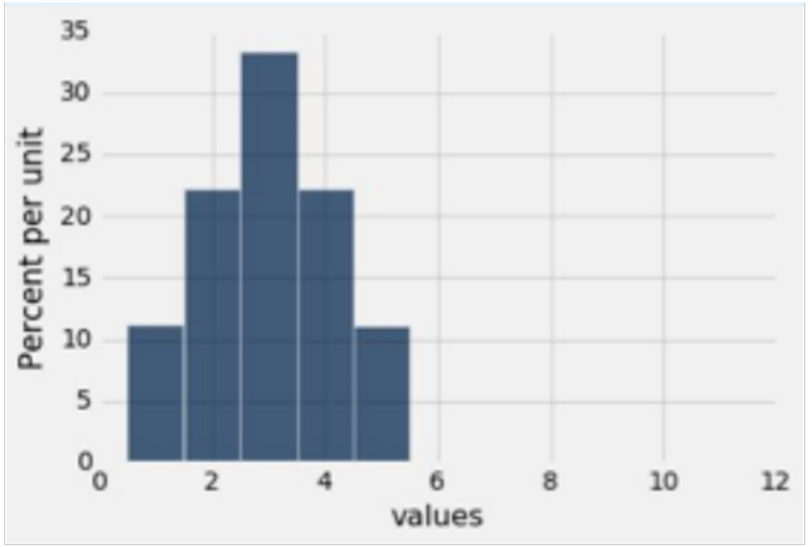


- List 2: [1,2,2,3,3,3,4,4,10]
  - Median = 3
  - Average =

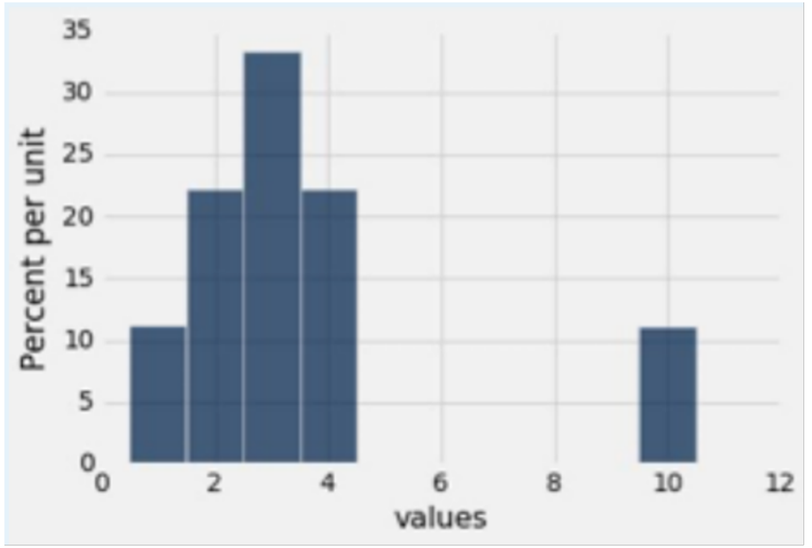


Average and Median

- List 1: [1,2,2,3,3,3,4,4,5]
  - Median = 3
  - Average = 3



- List 2: [1,2,2,3,3,3,4,4,10]
  - Median = 3
  - Average = 3.55556



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 tail

Standard Deviation

Variability

- Center of gravity of a histogram is the mean
    - What about the values on either side?
- Variability is how we describe how far apart values are spread away from the center (mean)

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

- To convert our units back to our original units, we need to take the square root
  - This gives us the **standard deviation**

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

- To compute the standard deviation of `arr`:
  - `np.std(arr)`

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

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

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

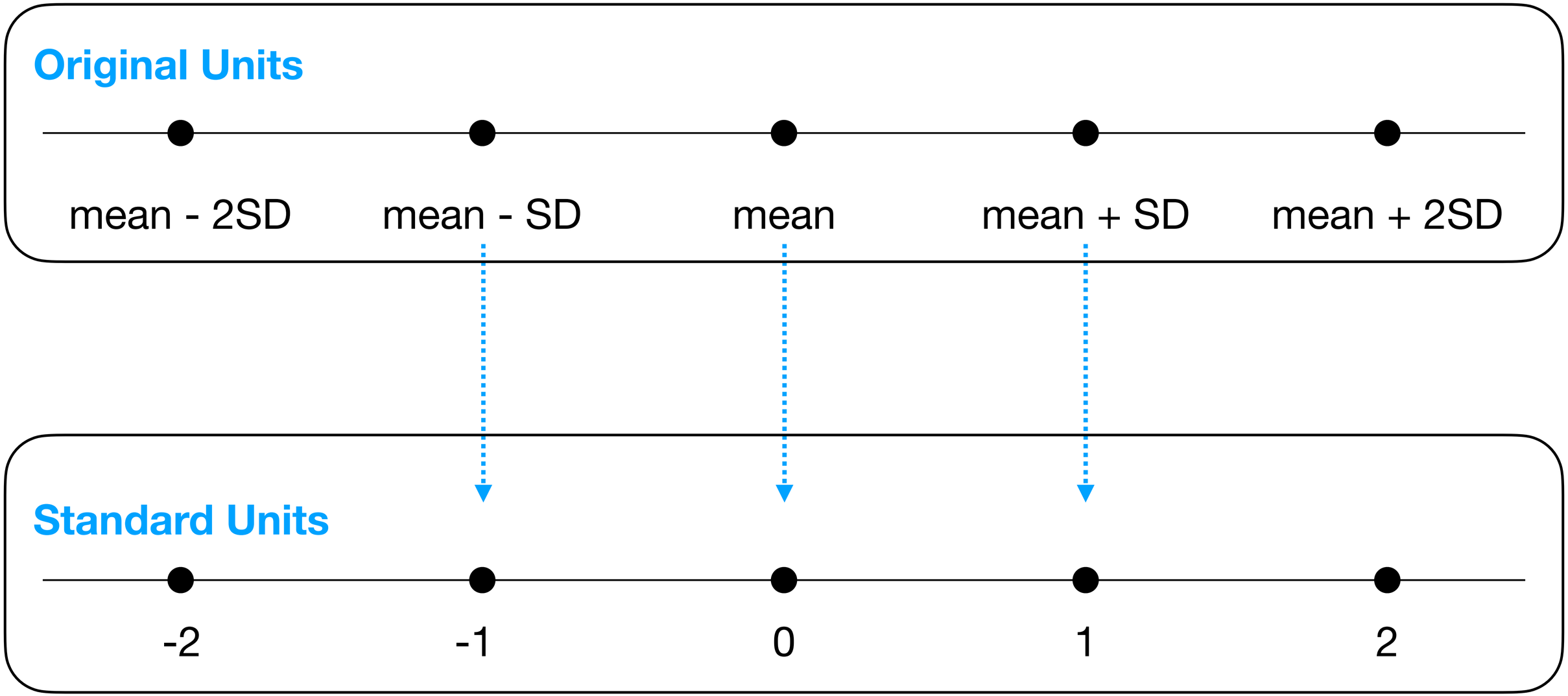
- The quantity z (from "$\text{average} \pm z\ \text{SDs}$" in Chebychev's inequality) measures **standard units**

  - **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

Example

What whole numbers are closest to:

- Average age?

- 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

Example

What whole numbers are closest to:

- Average age?
  - **27**. The standard unit is close to 0
- 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

Example

What whole numbers are closest to:

- Average age?
  - **27**. The standard unit is close to 0
- The SD of ages?
  - About **6** years. The standard unit at 33 is close to 1 and $33 - 27 = 6$

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

Normal Distribution

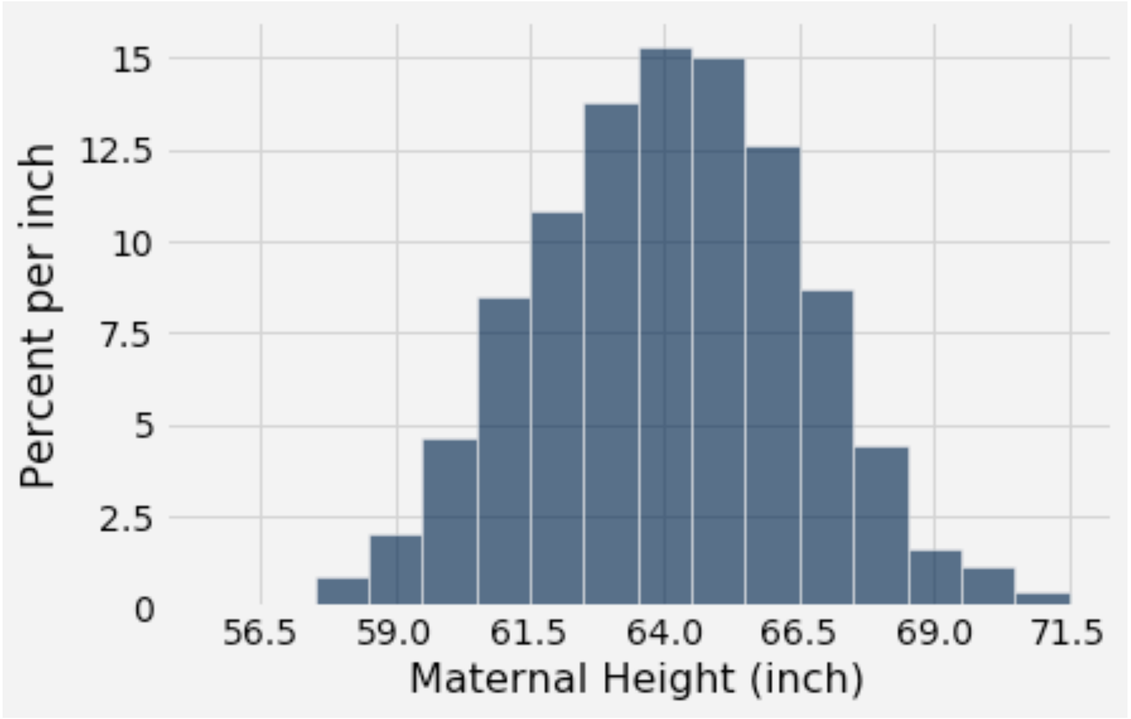
Bell Shaped Curves

- The normal curve / bell-curve is a very common distribution

- For bell-shaped (aka Gaussian distribution):

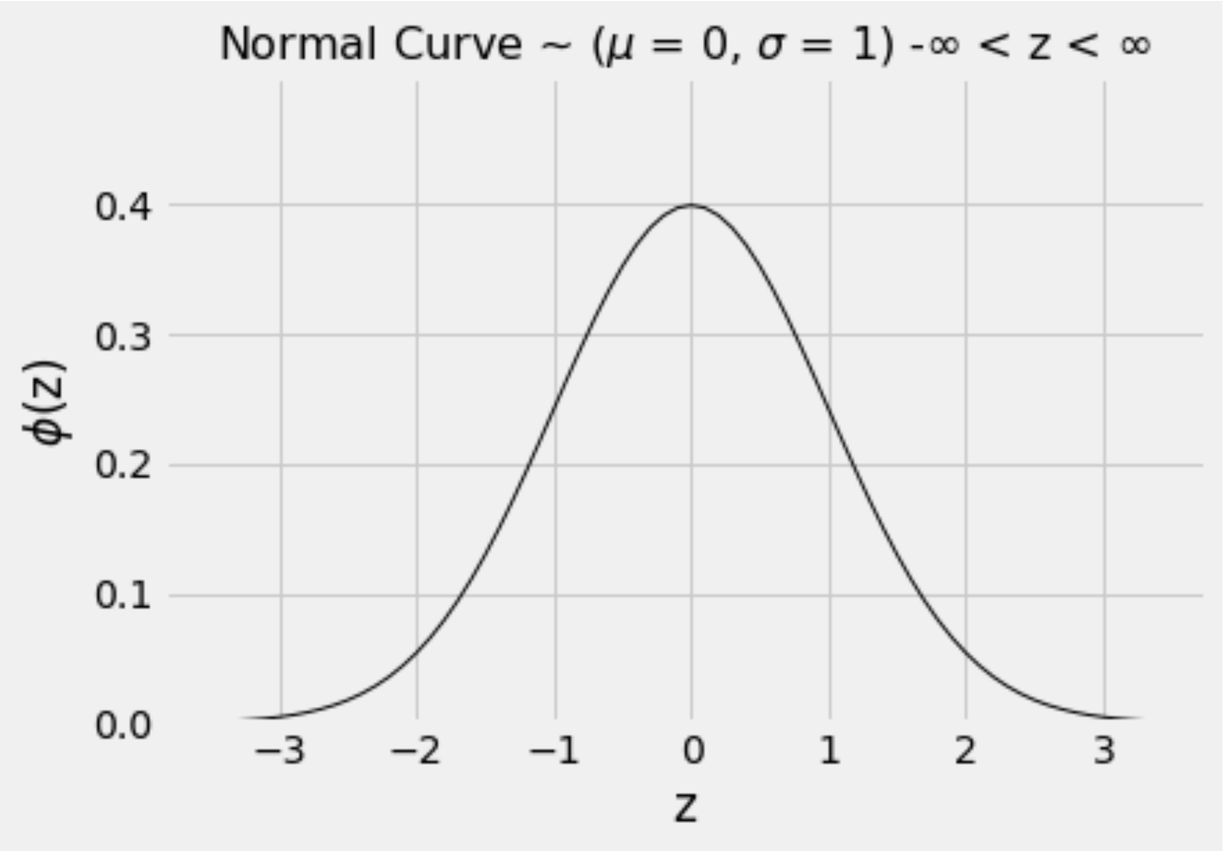
  - Average is at the center

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



Normal Distribution

- On a standard normal curve, x-axis units are standard units
- Total area of the curve is 1
- Curve is symmetric around 0 (mean and median are both 0)
- Points of inflection are -1 and 1
    - Standard deviation is 1

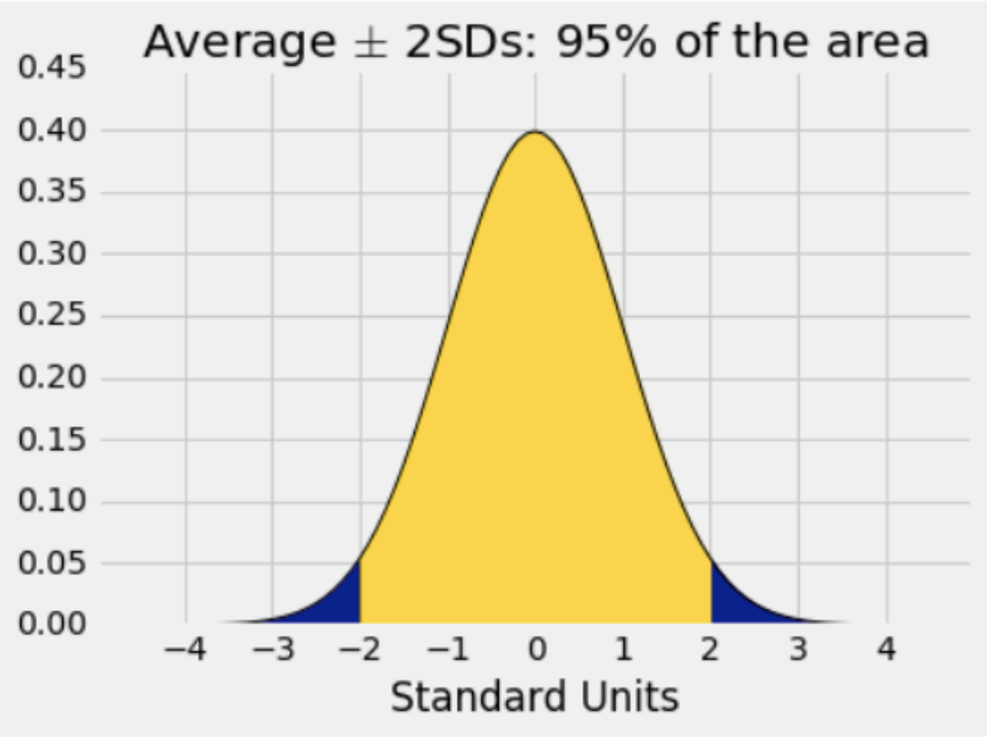


Application to Normal Distributions

- If a histogram is bell-shaped (normal), then 95% of the data is in the range
  $\text{average} \pm 2\ \text{SDs}$

- Note this is much higher than Chebychev's bound of 75%

  - 75% is a lower bound that applies to *all* distributions



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%

Central Limit Theorem

- Describes how a normal distribution is connected to random sample averages (the average of a sample we collect)

- We calculate sample averages because they can help us estimate population averages

Central Limit Theorem

Definition:

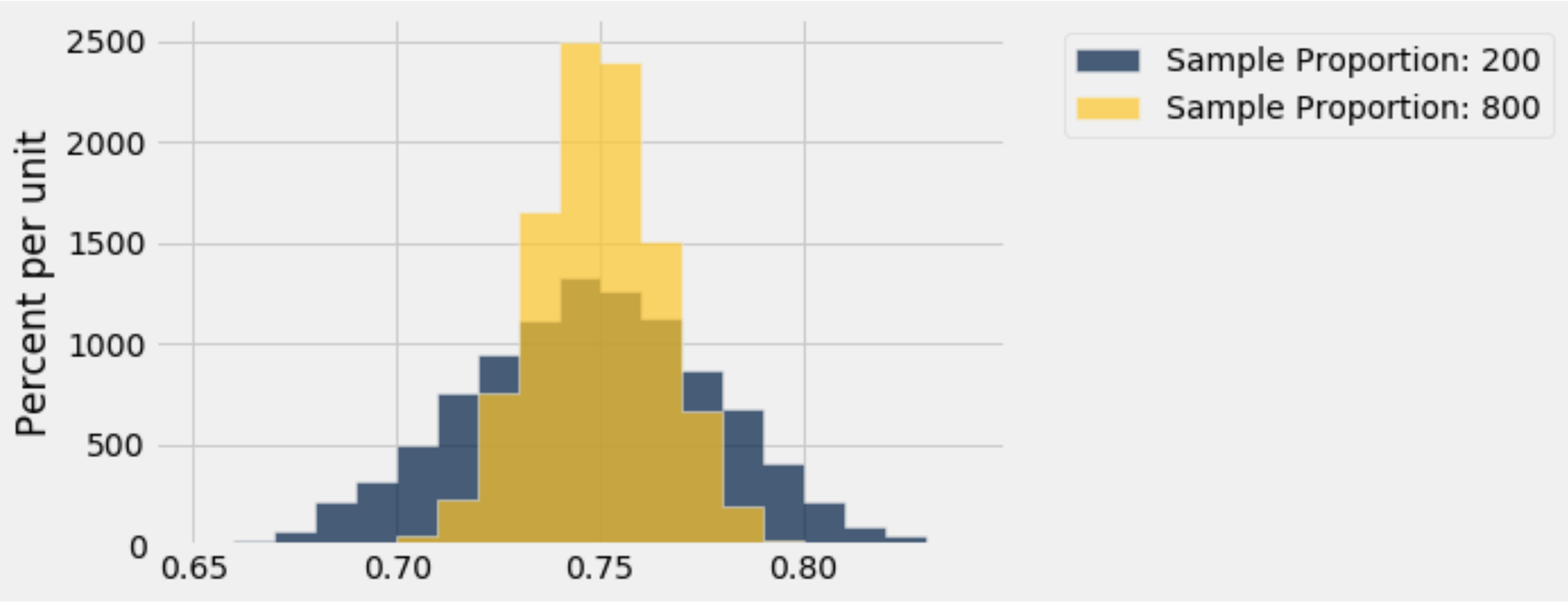
Is a sample is large and drawn at random with replacement

Then regardless of the distribution,

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

Central Limit Theorem

- Next time: how can we use this property to help us determine the sample size we need to draw useful conclusions?

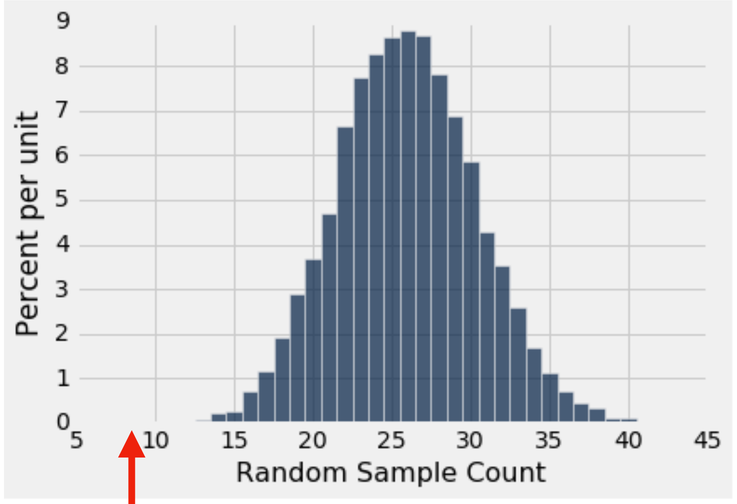


Summary of Stats so far…

Hypothesis Testing

- Modeling expected outcomes under the null and comparing it to our observed outcome

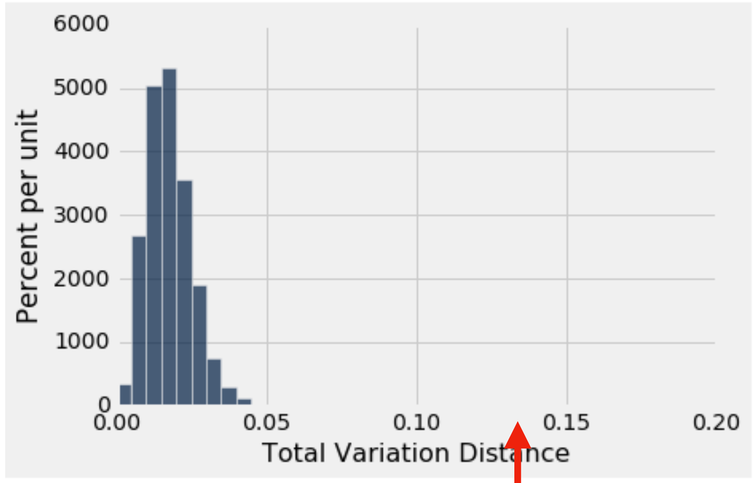
Swain v Alabama



Observed Number

2 Categories

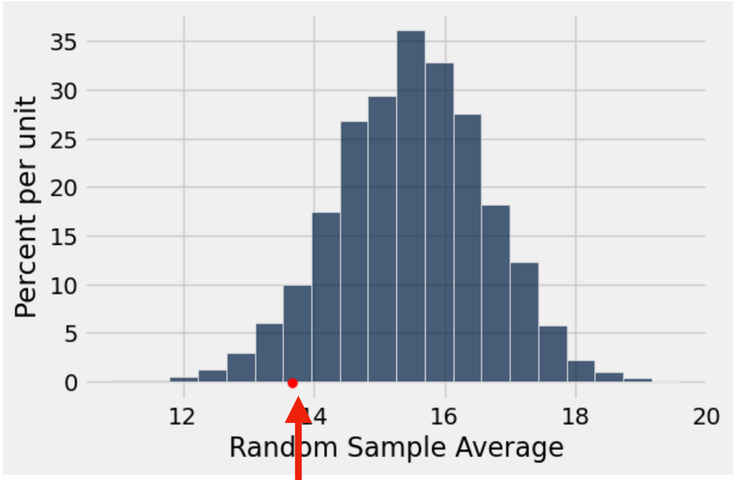
Alameda Jury



Observed TVD

3+ Categories

Midterm Exam Scores



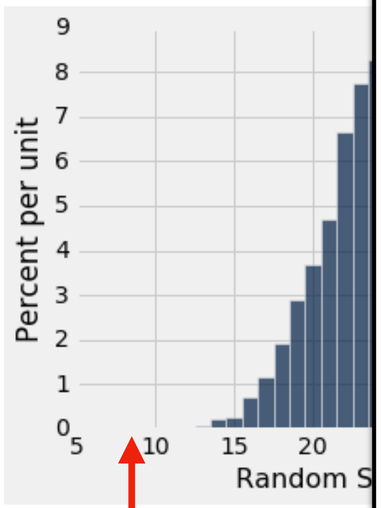
Observed Average

Numerical Data

Hypothesis Testing

- Modeling expected outcomes under the null and comparing it to our observed outcome

Swain v



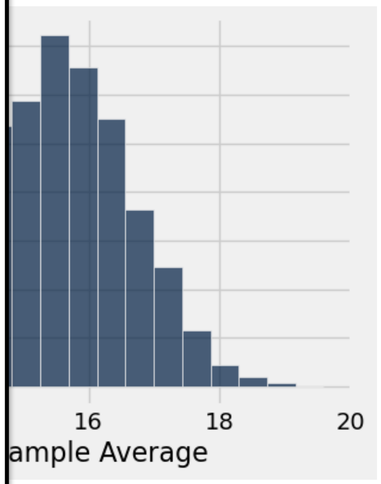
Observed N

It's often not easy to say whether the observed outcome falls within our expectations…

How can we more precisely characterize the likelihood of observing an expected outcome?

p-value!

kam Scores



Average

2 Categories

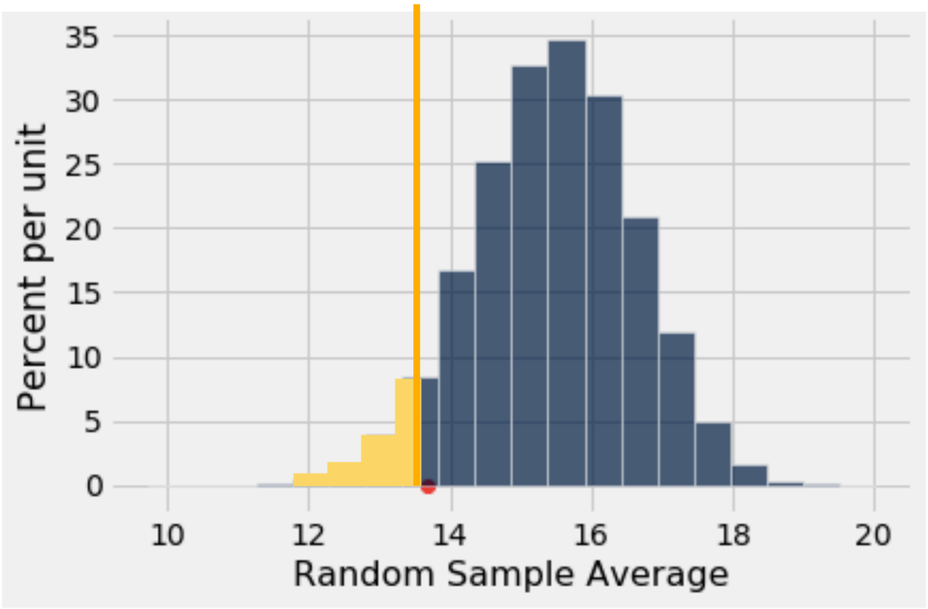
3+ Categories

Numerical Data

Hypothesis Testing

- Modeling expected outcomes under the null and comparing it to our observed outcome

Midterm Exam Scores



p-value = 0.058

p-value & **statistical significance**

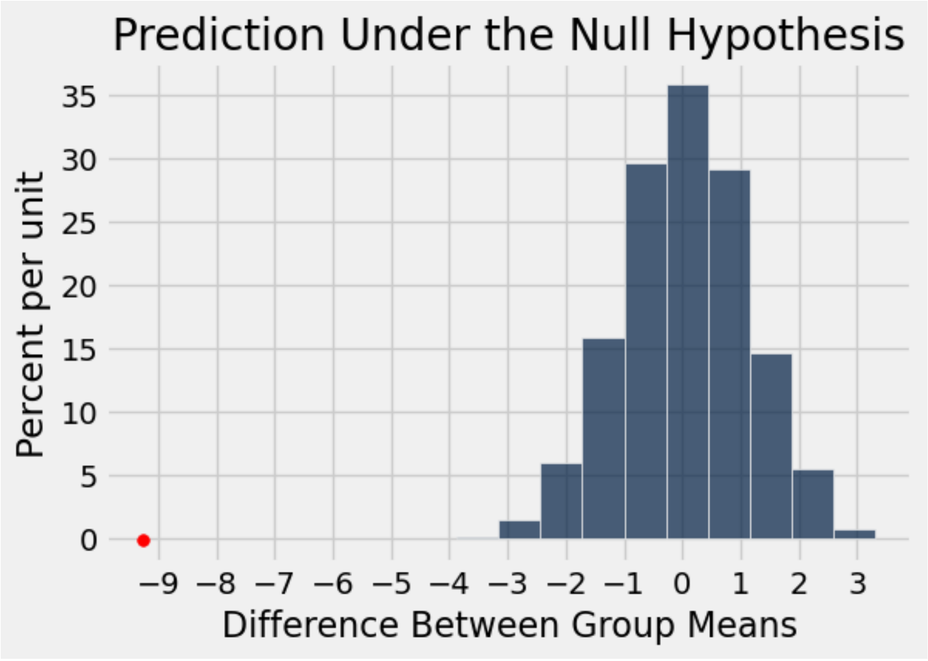
Process:

- Calculate the area of the tail (to the left/right of our observed value)

Hypothesis Testing

- Modeling expected outcomes under the null and comparing it to our observed outcome

Smoking vs Non-Smoking Mothers & Birthweight



p-value = 0

A/B Testing

Compare the difference between two groups

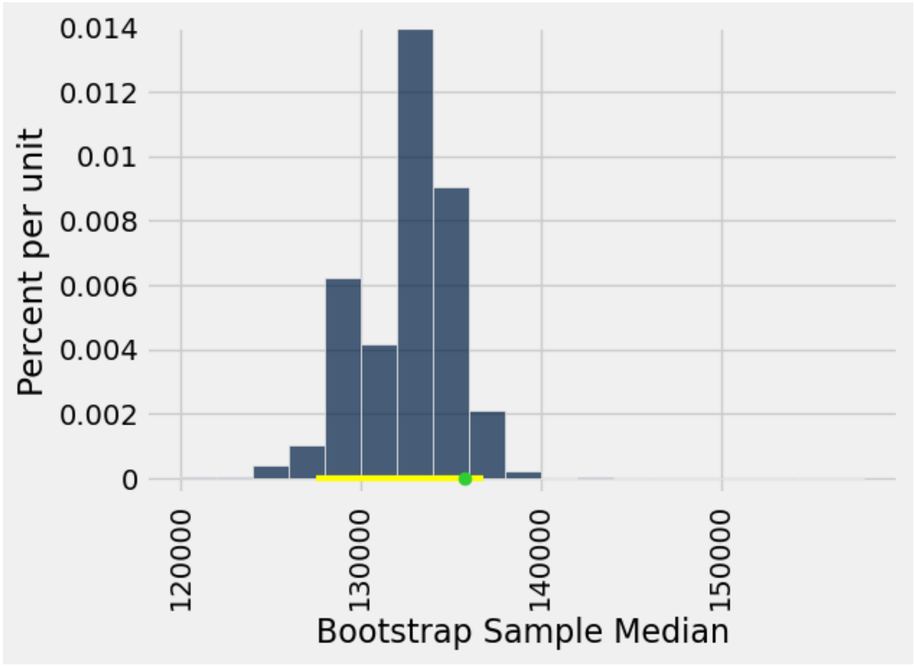
Process:

- Permutation test (shuffle labels)

Estimating a Parameter

- We want to estimate a population parameter from a sample statistic

Median Employee Salary



95% Confidence Interval: Median Salary between $125,745 and $140,318

Confidence Interval

Lets us estimate a range for what we think the parameter's value is

Process:

- Bootstrap

Next time

- Central Limit Theorem