

# CS 383 – Computational Text Analysis

## Lecture 19 Hypothesis Testing I

Adam Poliak

04/03/2023

# Announcements

Final Projects:

Originally 13 project ideations submitted

Now 12 submitted

Proposal: due this Friday

Today's lecture:

- <https://inferentialthinking.com/>
- Chapter 9.4 – 12 (inclusive)

Wednesday class on Zoom

# Midterm

- Allowed 1 page (double sided) cheatsheet
- Will post detailed topics covered on Midterm

# Machine Learning in a nutshell

In a ML model, what are we training?

- **Parameters!**

How do we train parameters in supervised learning?

*train parameters == figure out values for the parameters*

- Update weights by using them to make predictions and seeing **how far off our predictions** are
  - **Loss function!**

Algorithm to learn weights?

- **SGD**
- Others exist but not covering them

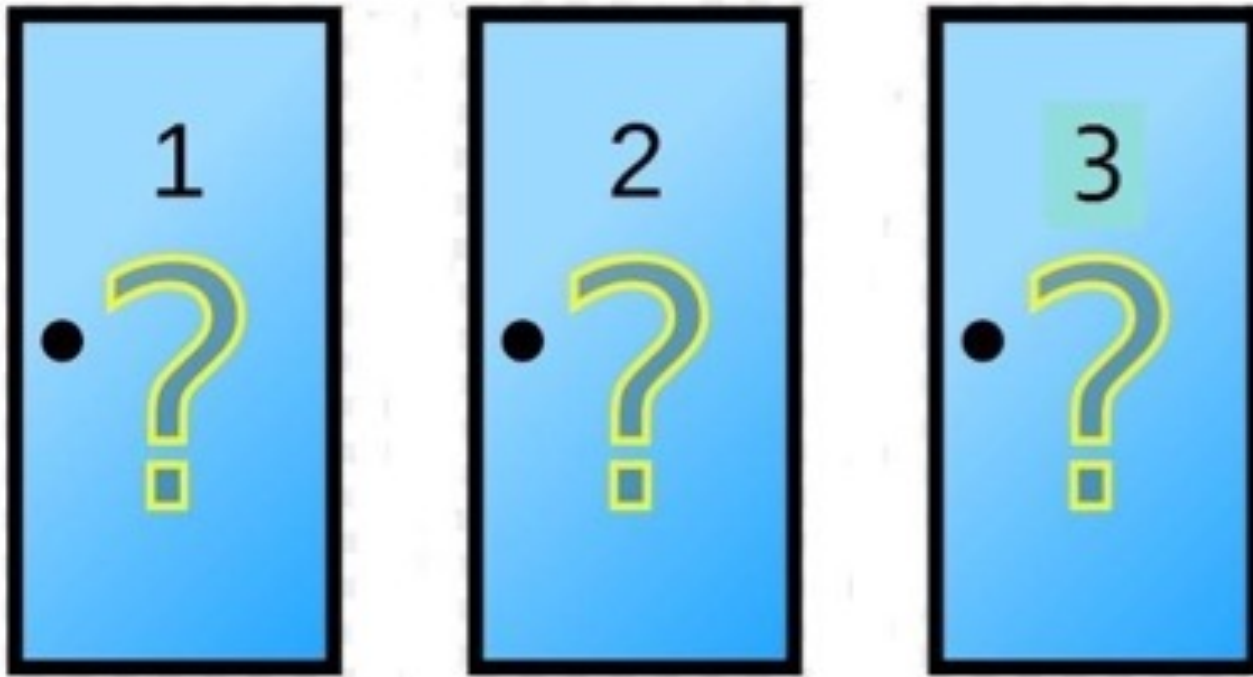
# Outline

- Monty Hall
- Distributions
  - Large Random Samples
- Statistic
- Assessing Models (overloading term here)
- Stat Sig
- A/B Testing (difference of means)

**LET'S  
MAKE  
A  
DEAL**

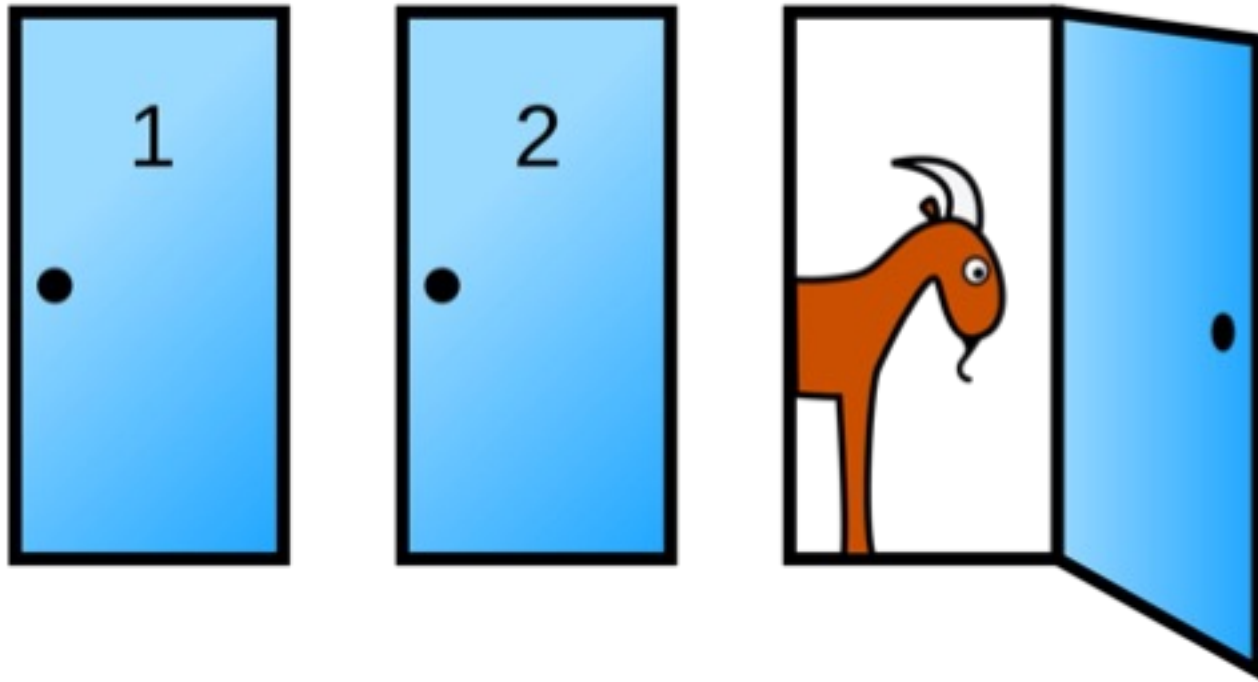


# Monty Hall Problem



<https://probabilityandstats.files.wordpress.com/2017/05/monty-hall-pic-1.jpg>

# Monty Hall Problem



[https://en.wikipedia.org/wiki/Monty\\_Hall\\_problem](https://en.wikipedia.org/wiki/Monty_Hall_problem)



# Monty Results

<b>Guess</b>	<b>Remaining</b>	
<b>car</b>	<b>first goat</b>	166
	<b>second goat</b>	178
<b>first goat</b>	<b>car</b>	320
<b>second goat</b>	<b>car</b>	336



# Distributions

# Probability Distribution

- Random quantity with various possible values
- “Probability Distribution”:
  - All the possible values of a quantity
  - The probability of each of the values
- Computing the probability distribution:
  - Math
  - Simulation .... often easier

# Empirical Distribution

- “Empirical” – based on observations
- Observations can be a repeated experiment
- “Empirical Distribution”:
  - All observed values
  - The proportion of times each value appears





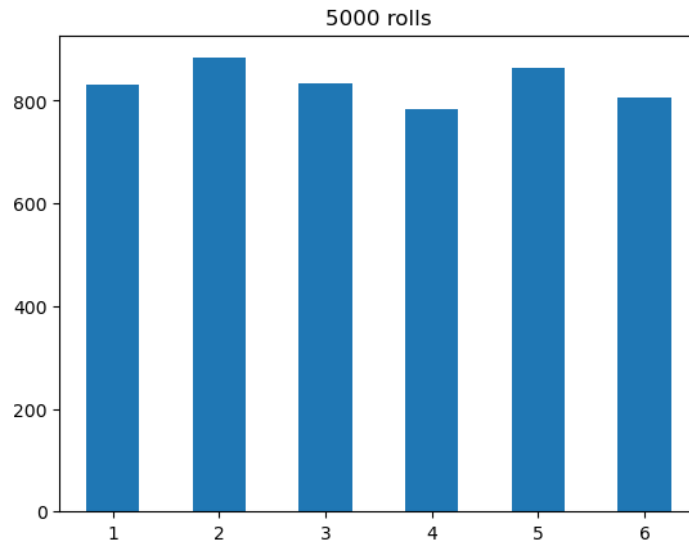
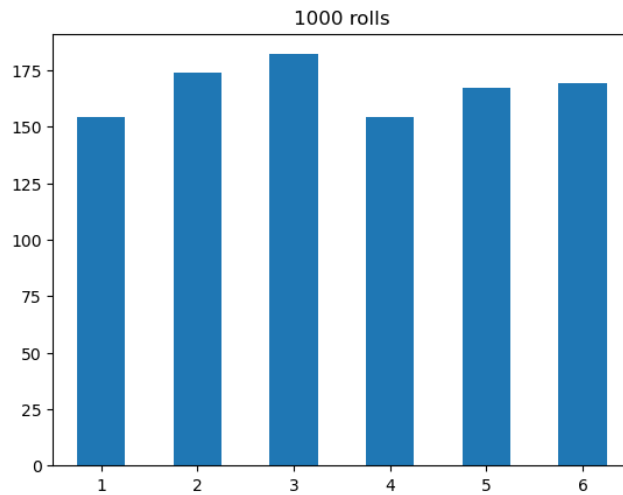
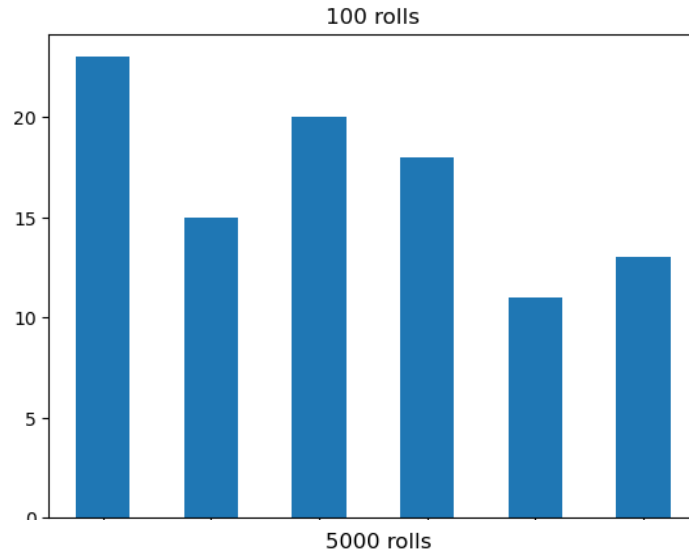
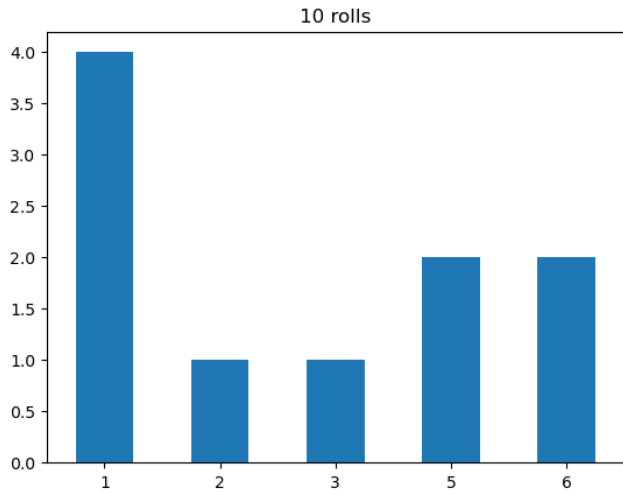
# Large Random Samples

# Law of Averages / Law of Large Numbers

If a chance experiment is repeated many times, independently and under the same conditions, then the proportion of times that event occurs gets closer to the theoretical probability of the event

As you increase the number of rolls of a die, the proportion of times you see the face with 5 dots gets closer to  $1/6$

# Law of Large Numbers



# Empirical Distribution of a Sample

If the sample size is large,  
then the empirical distribution of a uniform random  
sample  
resembles the distribution of the population,  
with high probability





# A Statistic

# Inference

- **Statistical Inference:**


- Making conclusions based on data in random samples

- **Example:**

- Use the data to guess the value of an unknown number



fixed



Depends on the  
random sample

- Create an **estimate** of an unknown quantity

# Terminology

- **Parameter**
  - Numerical quantity associated with the population
- **Statistic**
  - A number calculated from the sample
- A statistic can be used as an **estimator** of a parameter

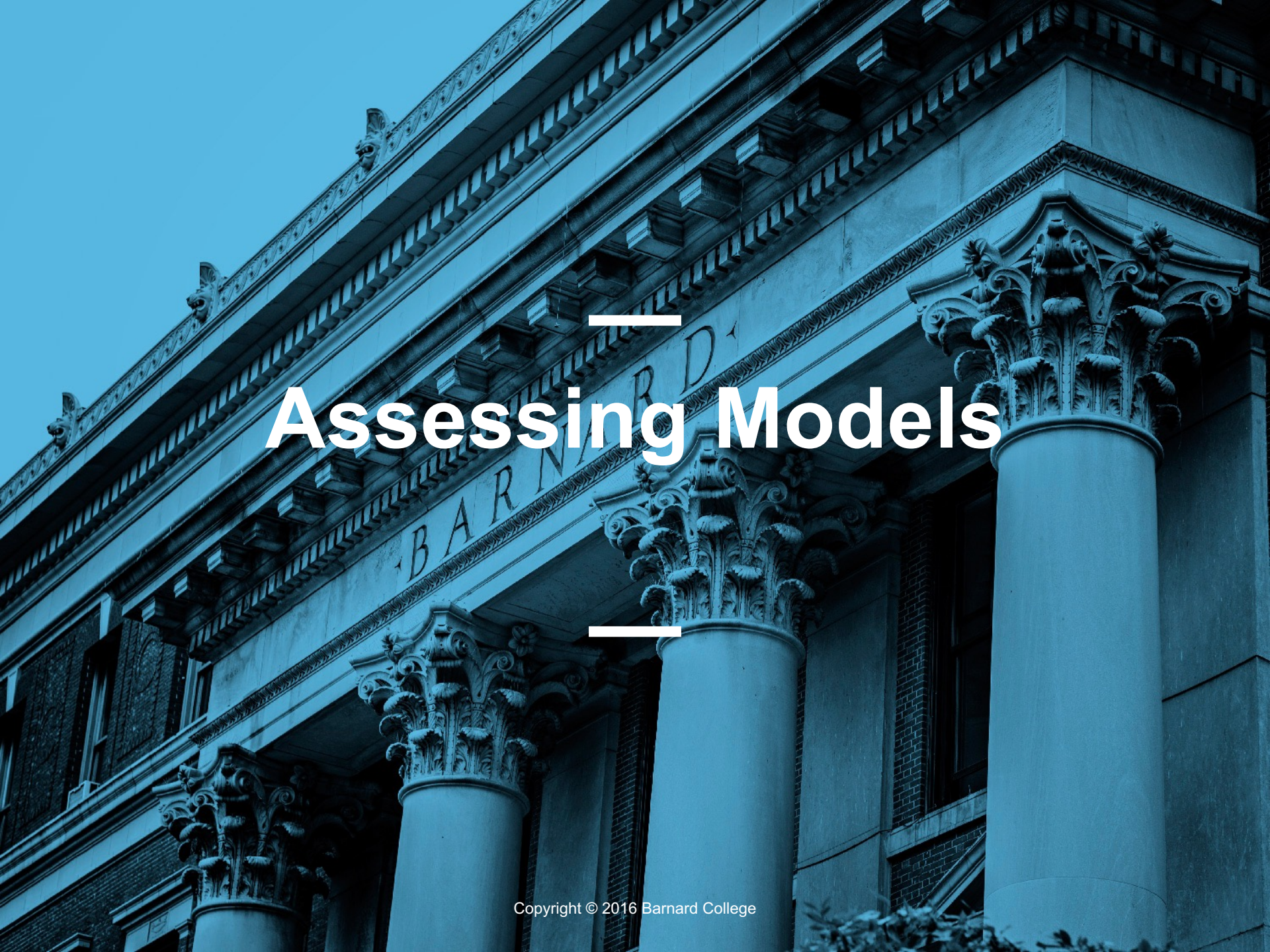


# Probability distribution of a statistic

- Values of a statistic vary because random samples vary
- “Sampling distribution” or “probability distribution” of the statistic:
  - All possible values of a statistic
  - and all corresponding probabilities
- Can be hard to calculate:
  - Either have to do math
  - Or generate all possible samples and calculate the statistic based on the each sample

# Empirical Distribution of a Statistic

- Based on simulated values of a statistic
- Consists of all observed values of the statistic,
- and the proportion of times each value appeared
  
- Good approximation to the probability distribution of a statistic
  - If the number of repetitions in the simulation is large



---

# Assessing Models

---

# Models

- A model is a set of assumptions about the data
- In data science, many models involve assumptions about processes that involve randomness:
  - “Change models”
- **Key question:** does the model fit the data?



# Approach to Assessing Models

- If we can simulate data according to the assumptions of the model, we can learn what the model predicts
- We can compare the model's predictions to the observed data
- If the data and the model's predictions are not consistent, that is evidence against the model





# Jury Selection

# Swain vs. Alabama, 1965

- Talladega County, Alabama
- Robert Swain, black man convicted of crime
- Appeal: one factor was all white-jury
- Only men 21 years or older were allowed to serve
- 26% of this population were black
- Swain's jury panel consisted of 100 men
- 8 men on the panel were black

# Supreme Court Ruling [in English]

- About disparities between the percentages in the eligible population and the jury panel, the Supreme Court wrote:
  - “... the overall percentage disparity has been small and reflects no studied attempt to include or exclude a specified number of Negroes”
- Supreme Court denied Robert Swain’s appeal



# Supreme Court Ruling [in Data]

- **Paraphrase:** 8/100 is less than 26%, but not different enough to show Black men were systematically excluded
- **Question:** is 8/100 a realistic outcome if the jury panel selection process were truly unbiased?

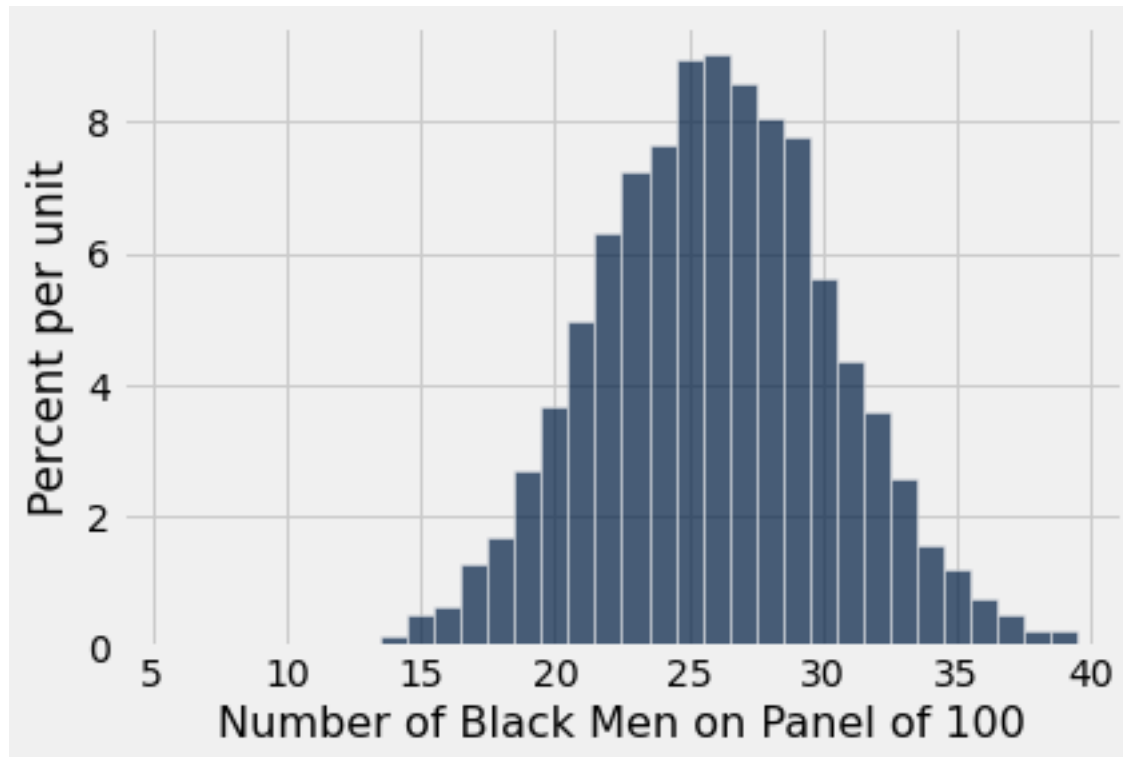
# Steps in Assessing a Model

- 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***

# 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

# Simulated Values



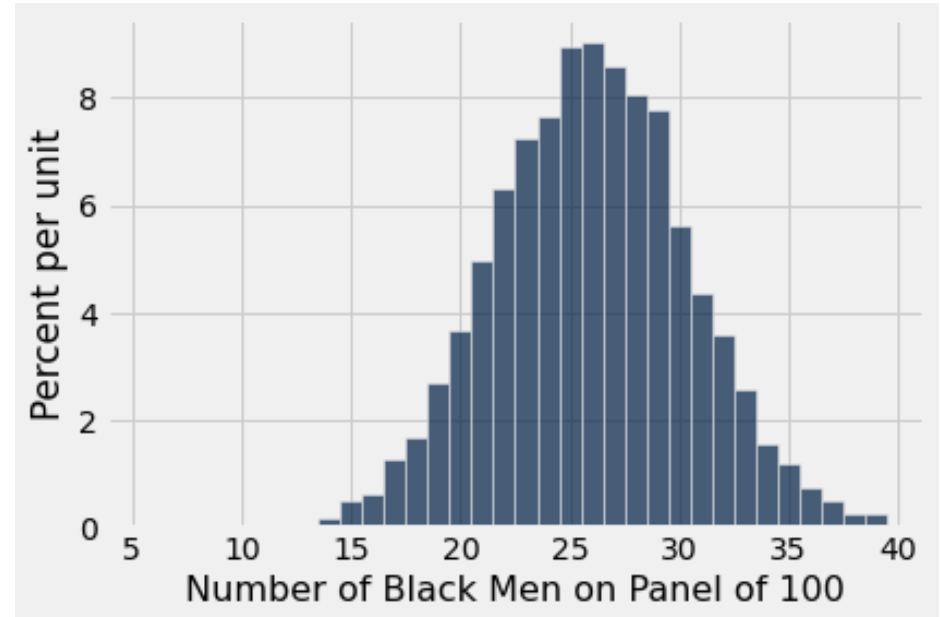
# Steps in Assessing a Model

- 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*



# Assessing model

Simulated Empirical  
Distribution



Observed statistic: 8%

Are they consistent?

No, evidence against random model

# Model and Alternative

- **Model:** The people on the jury panels were selected at random from the eligible population
- **Alternative viewpoint:** No, they weren't

# Steps in Assessing a Model

- Choose a statistic to measure the “discrepancy” between model and data
- Simulate the statistic under the model’s assumptions
- Compare the data to the model’s predictions:
  - Draw a histogram of simulated values of the statistic
  - Compute the observed statistic from the real sample
- If the observed statistic is far from the histogram, that is evidence against the model



—

# Terminology

—

# Testing Hypotheses

- A test chooses between two views of how data are generated
- What are these views called?
  - Answer: **hypotheses**
- The test picks the hypothesis that is better supported by the observed data
- What is the method for choosing the hypotheses?
  - Simulate data under one of the hypotheses
  - Compare the simulation results and the observed data
  - Pick one of the hypotheses based on whether the simulated results and observed data are consistent

# 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

# Test Statistic

- The statistic that we choose to simulate, to decide between the two hypotheses

Questions before choosing the statistic:

- What values of the statistic will make us lean towards the null hypothesis?
- What values will make us lean towards the alternative?
  - Preferably, the answer should be just a “high” or just a “low” value
  - Try to avoid “both high and low”

# 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



# Conclusion of the Test

Resolve choice between null and alternative hypotheses

- Compare the **observed test statistic** and its empirical distribution under the null hypothesis
- If the observed value is not **consistent** with the empirical distribution
  - The test favors the alternative
  - “data is more consistent with the alternative”

Whether a value is consistent with a distribution:

- A visualization may be sufficient
- If not, there are conventions about “consistency”

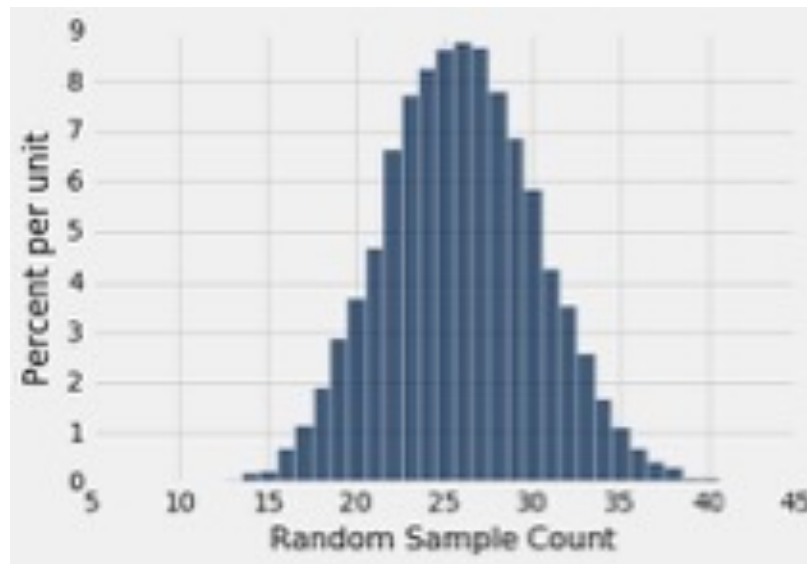


# Statistical Significance



# Tail Areas

## Alabama Jury

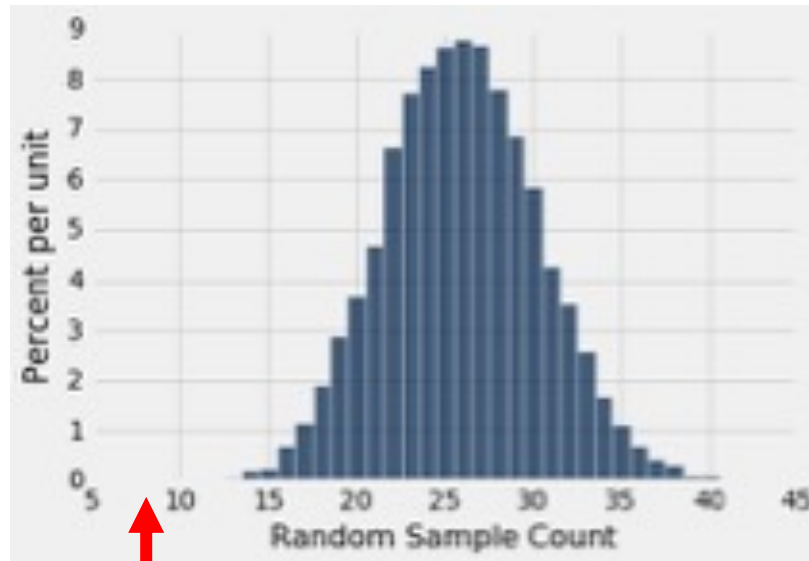


# Conventions About Inconsistency

- **“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”

# Tail Areas

## Alabama Jury



Observed Number (8)

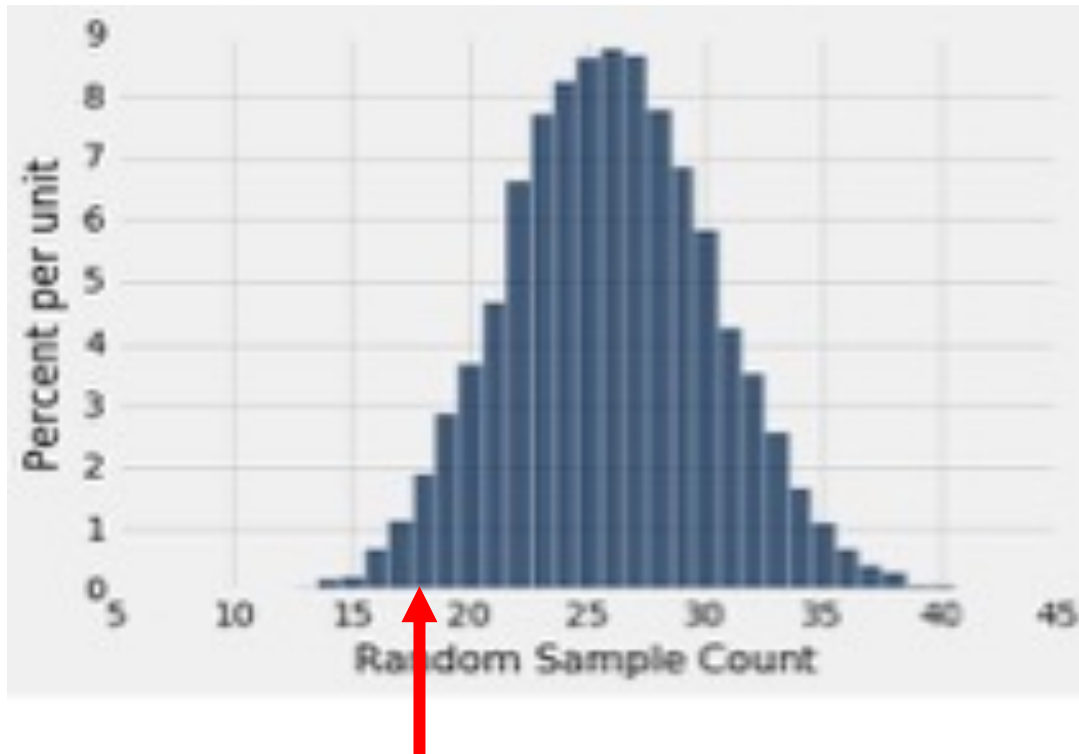
# Definition of the P-value

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

# Not so clear example



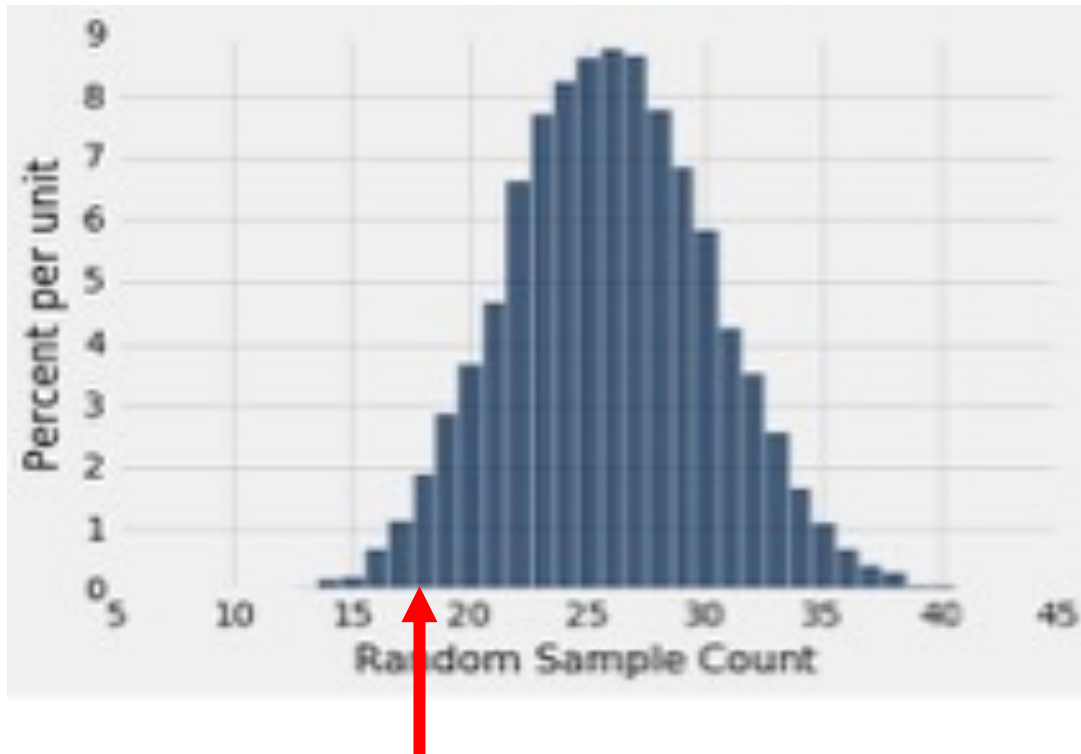
Observed Number (18)

# Conventions About Inconsistency

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



# Not so clear example



Observed Number (18)

# 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”

# Observed significance level (*aka* P-value)

Formal name: **observed significance level**

The  $P$ -value is the chance,

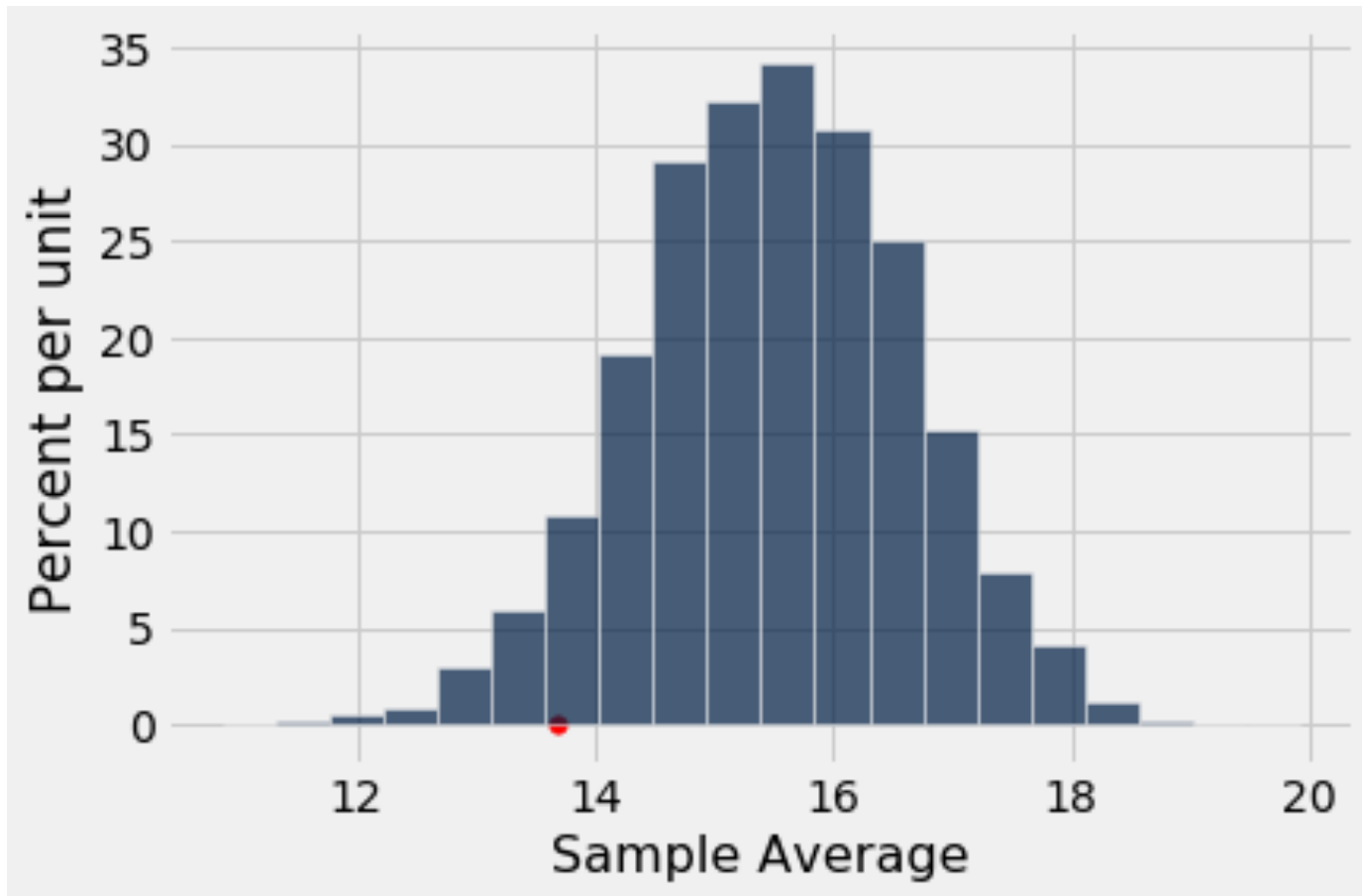
- Under the null hypothesis,
- That the test statistic
- Is equal to the value that was observed in the data
- Or is even further in the direction of the tail

# Assessing a Model

- Choose a statistic to measure the “discrepancy” between model and data
  - Average score per 27 students
- Simulate the statistic under the model’s assumptions
- Compare the data to the model’s predictions:
  - Draw a histogram of simulated values of the statistic
  - Compute the observed statistic from the real sample

# Histogram of simulated values & observed statistic

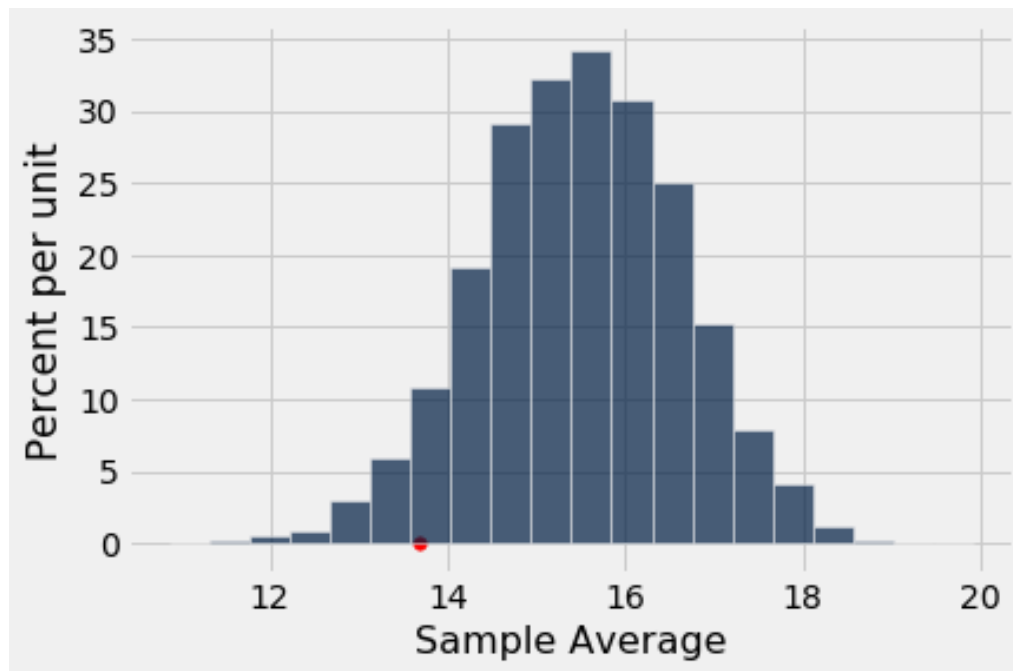
Is the observed statistic consistent with the histogram?



# Compute the p-value

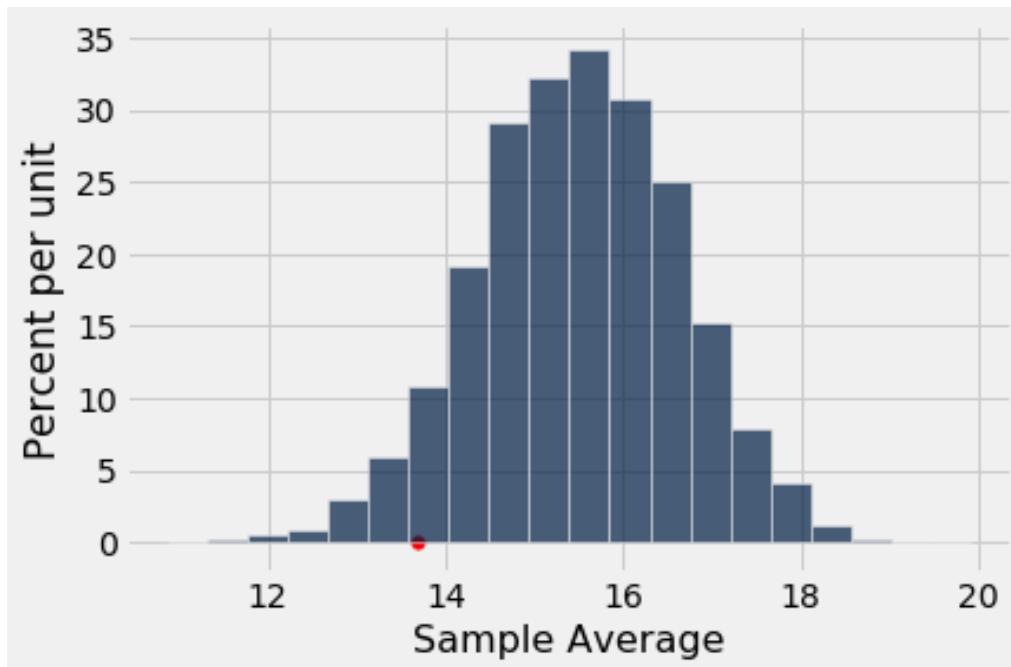
The  $P$ -value is the chance,

- 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



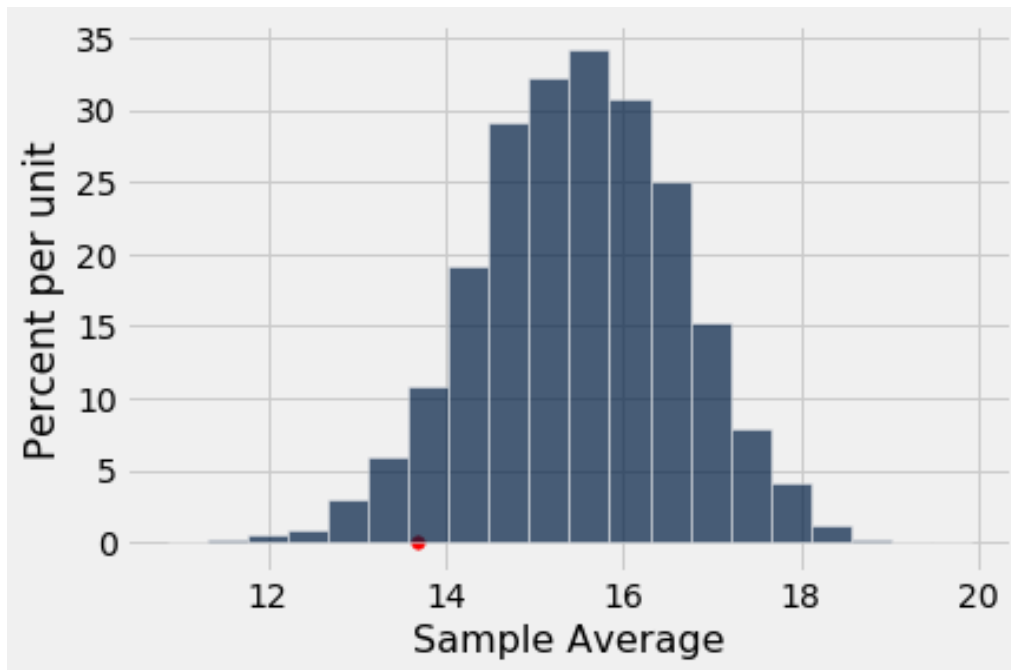
# Compute the p-value

$$\text{Probability (A)} = \frac{\text{number of outcomes that make A happen}}{\text{total number of outcomes}}$$



# Compute the p-value

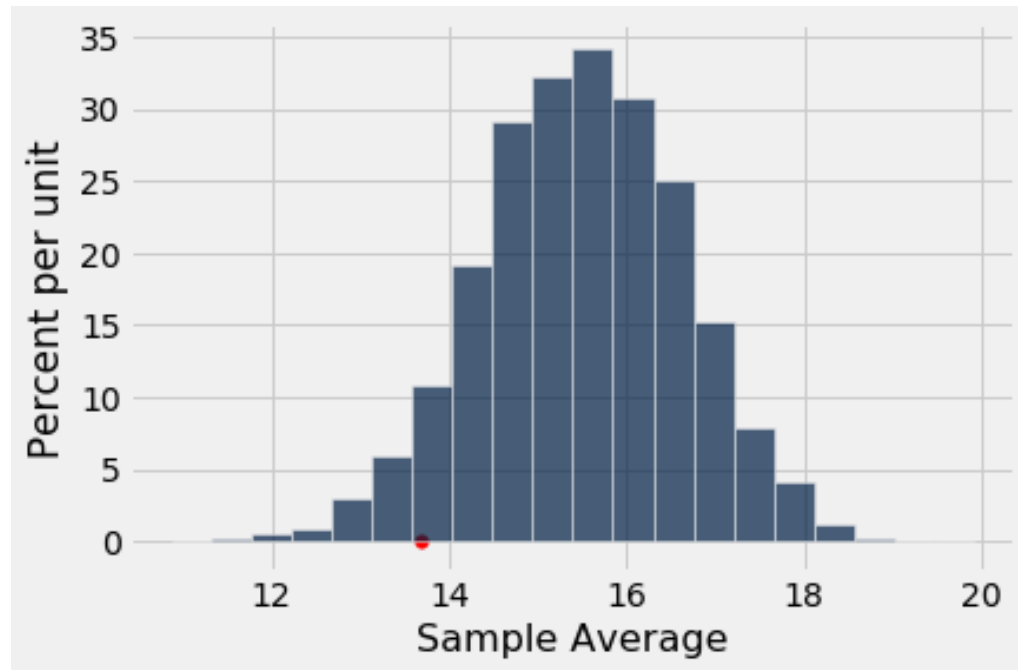
A = the sampled statistic was less than or equal to the observed statistic





# Compute the p-value

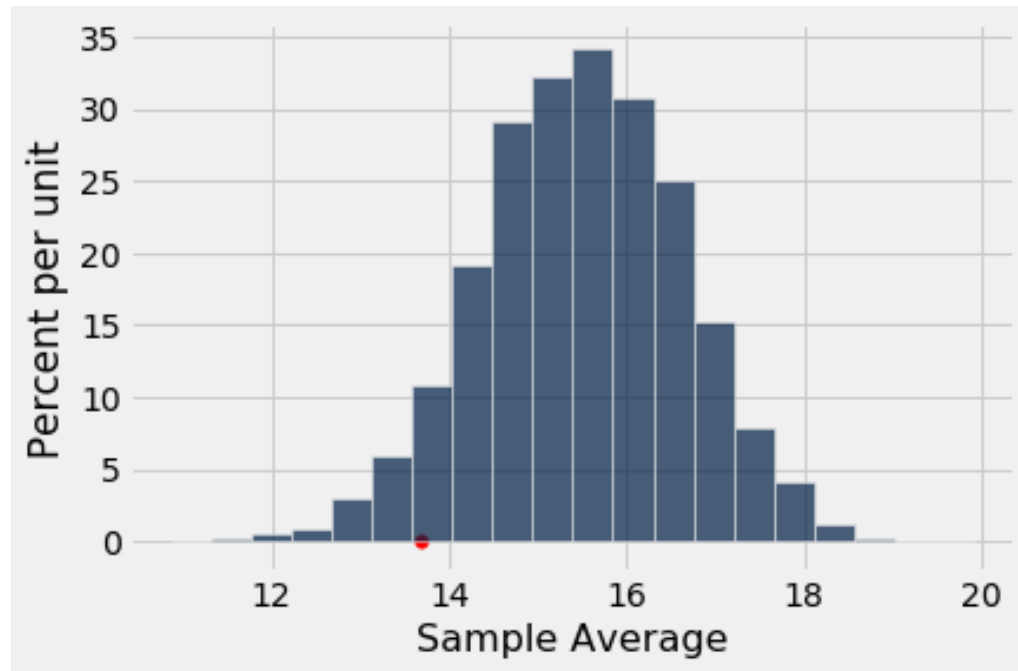
$P(A)$  = (the number of times the sampled statistic was less than the observed statistic) divided by the number of samples



# Compute the p-value

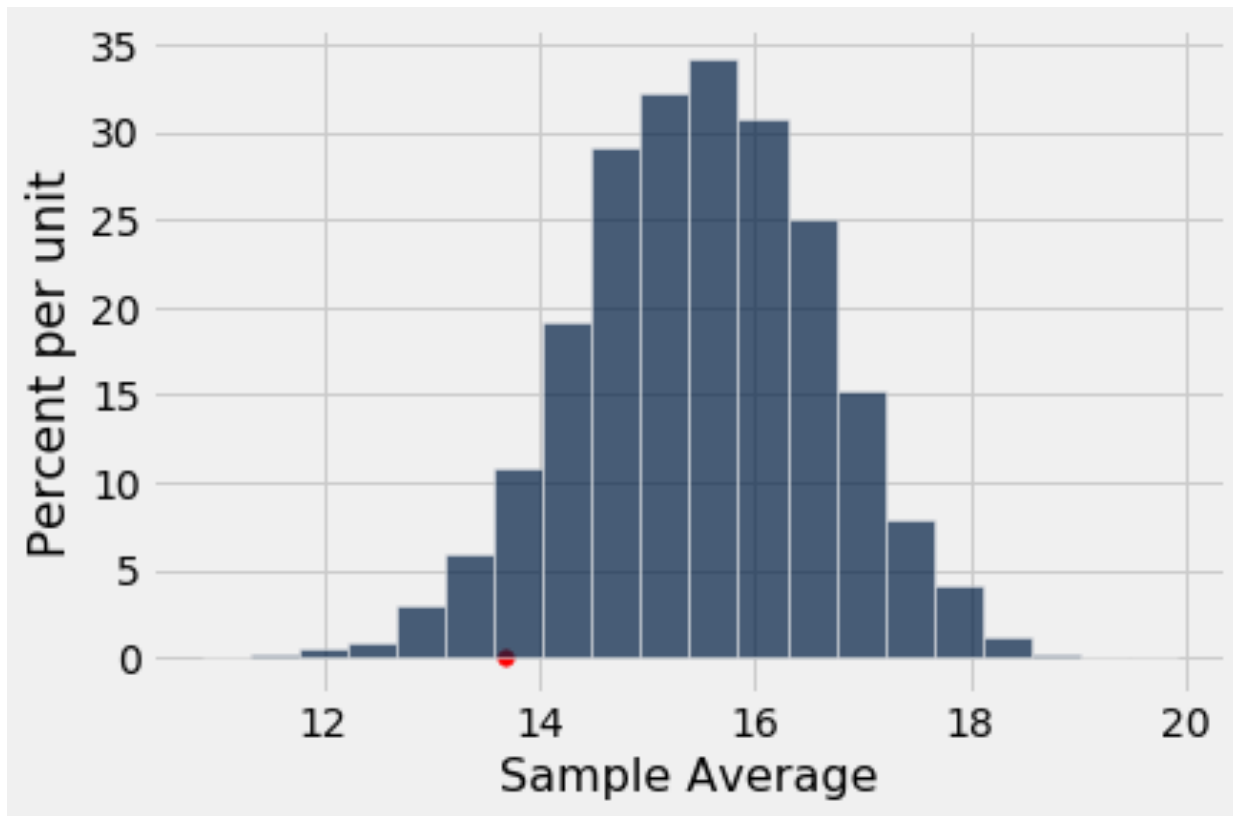
$P(A) =$

$$\frac{\text{sum}(\text{sample averages} \leq \text{observed averages})}{50K}$$

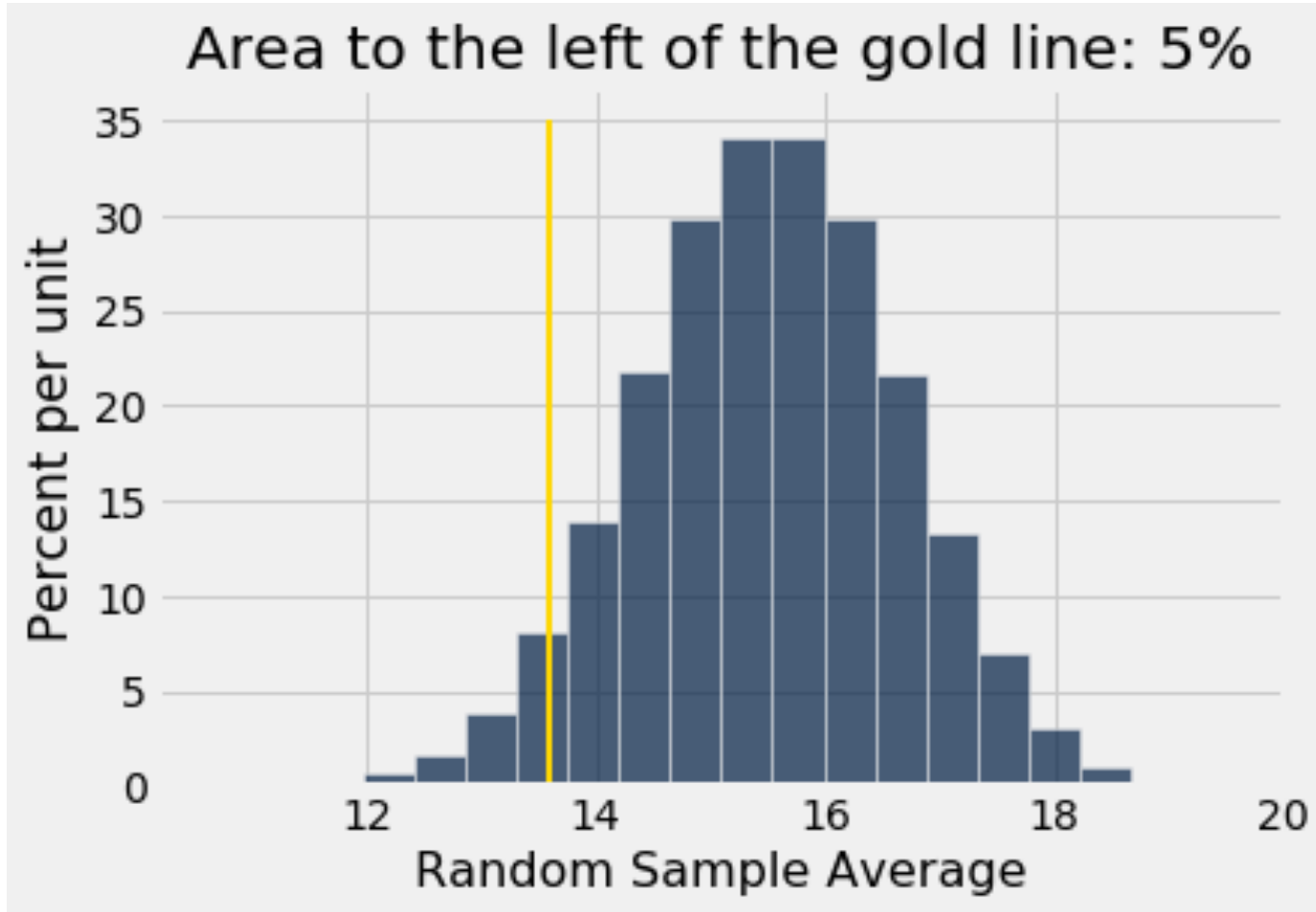


# Compute the p-value

$$P(A) = 0.05682 \approx 5\%$$



# Compute the p-value







# Comparing Two Samples A/B Testing

# 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**.



# 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?

# 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

# 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

# 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

# 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

# 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



# Randomized Assignment & Shuffling

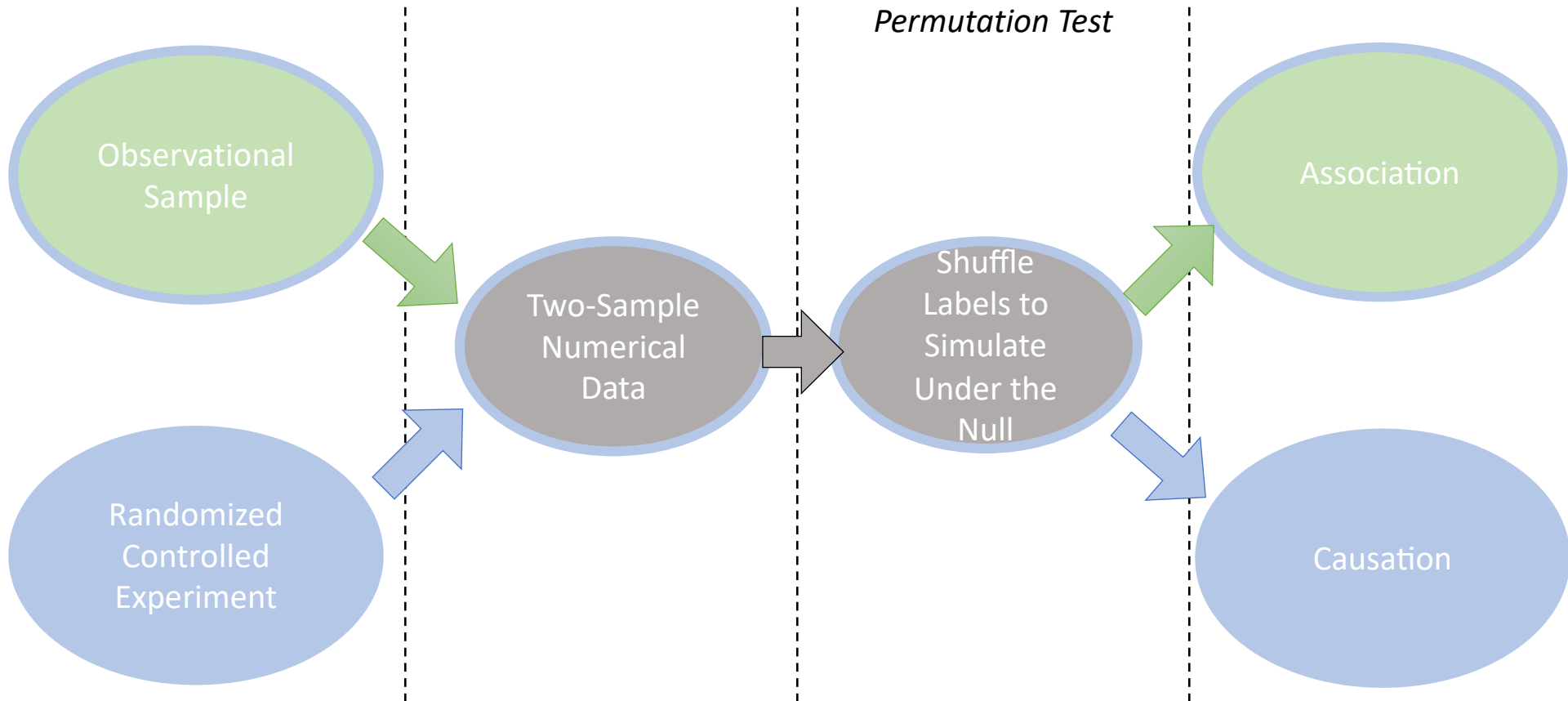
**Data Generation**

**Sample Data**

**Hypothesis Testing**

**Conclusions**

*Difference of Means  
Permutation Test*





# Estimation



# 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



# — Estimation Variability —



# 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:

$$\text{Estimate} = \text{Parameter} + \text{Error}$$

- 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

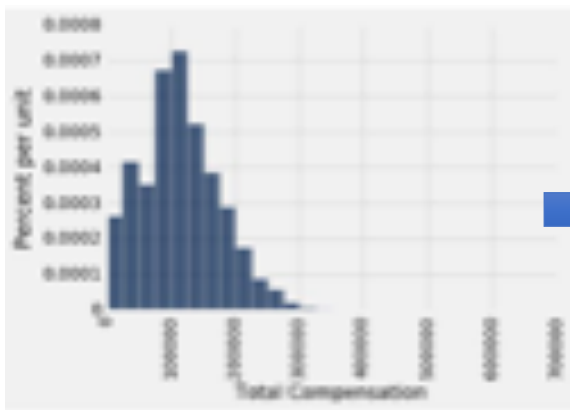


# The Bootstrap

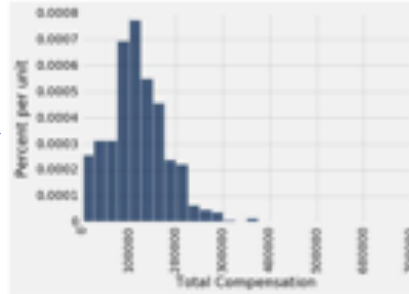
- A technique for simulating repeated random sampling
- All that we have is the original sample
  - ... which is large and random
  - Therefore, it probably resembles the population
- So we sample at random from the original sample!

# How the Bootstrap works

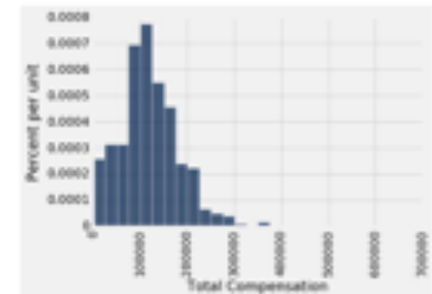
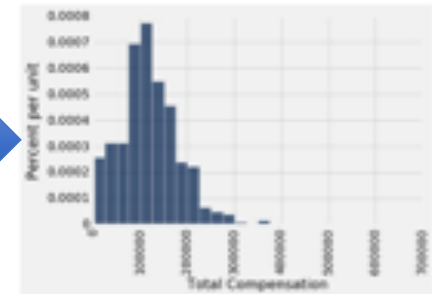
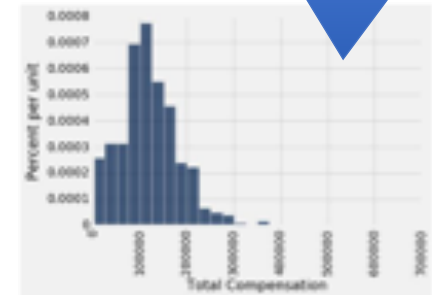
Population



Sample

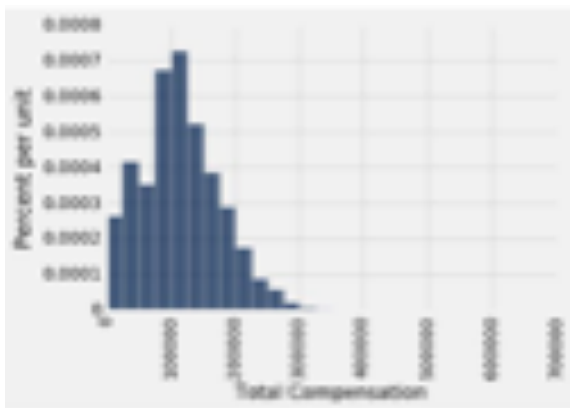


Resamples



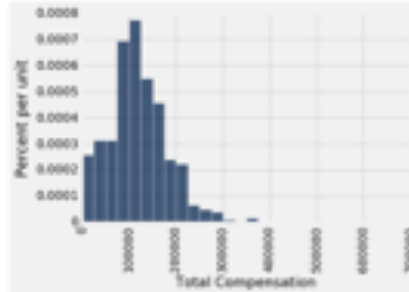
# Why the Bootstrap works

Population



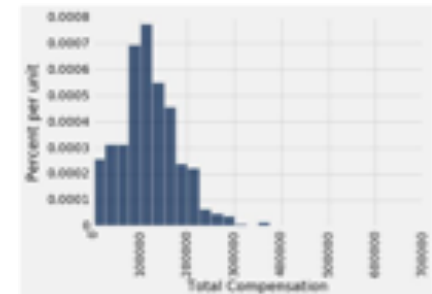
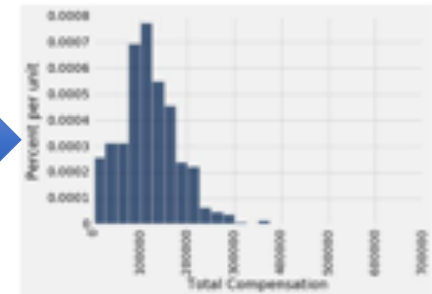
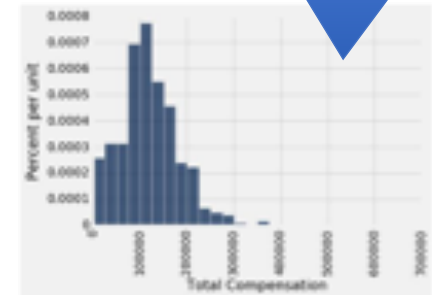
What we wish we could get

Sample



What we actually can get

Resamples





# 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

# The Bootstrap Principle

- The bootstrap principle:
  - **Bootstrap-world** sampling  $\approx$  **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

# 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



# — 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 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



# Confidence Intervals & Hypothesis Tests



# 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