Ch. 1: Significance: How Strong Is the Evidence?

Navigation

By Date

- January 16: start - end
- January 21: start - end
- January 23: start - end
- January 28: start - end
- January 30: Review

By Section

- Section 1.1
- Section 1.2
- Section 1.3
- Section 1.4
- Section 1.5

1.1: Introduction to Chance Models

Statistical Significance

Statistical significance - unlikely to have occurred by random chance

> Helper vs. Hinderer example, what is P(14/16 infants)? Is it rare? What is P(8/16 infants)?

Probability - Long run proportion of times an outcome from a random process occurs

> Let's Make a Deal example - result after 100s of simulations

Statistical Significance

General strategy:

1. Create a model for the process under random chance
2. Use the model to calculate the **probability** of the observation under random chance
3. If that probability is very small, conclude that the effect is likely not due to chance
   e.g. is **statistically significant**

Vocabulary

- **Statistic** - number summarizing the results from a *sample*
- **Parameter** - long-run numerical property of the *population*

Parameter -> **P**opulation

unknown value

Statistic -> **S**ample

known value

We use **statistics** to estimate **parameters**

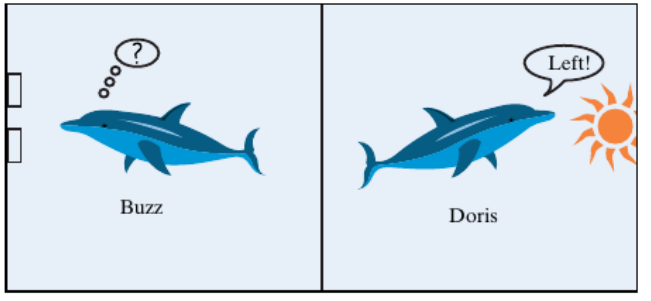
Example 1.1: Can Dolphins Communicate?

Training:

1. Shine headlight into tank
2. If light is steady, hit the right-side lever to get a fish
3. If light is blinking, hit the left-side lever to get a fish

Experiment:

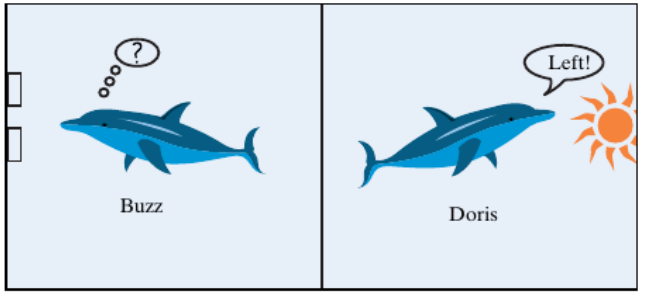
1. Show Doris the light
2. Give Buzz the levers
3. See if Doris could tell Buzz which lever to push



Example 1.1: Can Dolphins Communicate?

Training:

1. Shine headlight into tank
2. If light is steady, hit the right-side lever to get a fish
3. If light is blinking, hit the left-side lever to get a fish



Experiment:

1. Show Doris the light
2. Give Buzz the levers
3. See if Doris could tell Buzz which lever to push

Buzz hit the correct lever 15/16 times

Example 1.1: Can Dolphins Communicate?

1. Ask a research question:
   Can the dolphins communicate ideas?

2. Design a study and collect data:

   - Observational Units: Each of Buzz's attempts
   - Variable: Correct or incorrect button
   - Categorical or quantitative?: Categorical

3. Explore the data: We observed 15 correct responses out of 16 total trials

4. Draw inferences beyond the data:

   > What would we expect the **statistic** to be if Buzz was guessing?
   > Was the observed **statistic** significantly greater than that?

We need to build a chance model to find out!

Example 1.1: Can Dolphins Communicate?

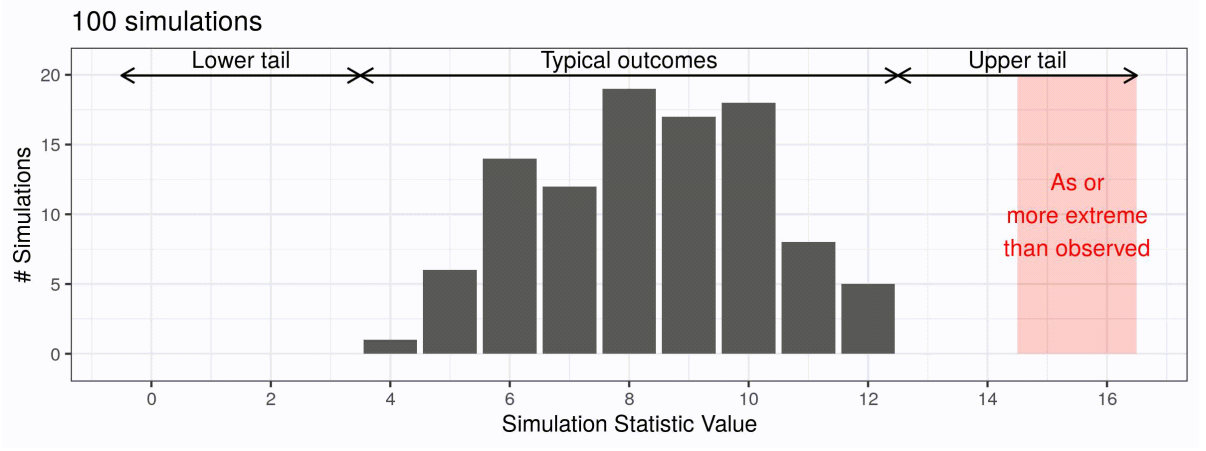
Real Study	Physical Simulation
Lever pull by Buzz	Coin Flip
Correct lever	Heads
Incorrect lever	Tails
Probability of correct button **when Buzz is guessing** (1/2)	Chance of heads (1/2)
16 attempts by Buzz	16 coin flips = One **Repetition** of the study

If Buzz is just guessing, what is a **typical** value we would see in one repetition of the study?

http://www.rossmanchance.com/ISIapplets.html

Determining Statistical Significance: 3 S Strategy

1. **Statistic** - numerical summary from sample
    - Helper/Hinderer: 14/16
    - Doris/Buzz: 15/16
2. **Simulate** - identify a "by chance alone" model for the scenario and repeatedly simulate values for the statistic under that model
3. **Strength of Evidence** - Is the observed statistic unusual?



Exploration 1.1: Can Dogs Understand Human Cues?

1. State the research question
2. Observational units
3. Variable
4. Sample size
5. Observed statistic
6. Is the statistic in the direction suggested by the researcher?
7. Is the observed result *possible*?
8. Is the observed result *probable*?
9. If he picked the cup at random, what is the probability (parameter) that he chooses correctly?
10. What would the probability of choosing the correct cup be if Harley understands the experimenter?
    (can be a range of values)

For next time

- Exploration P.3 due January 17 (Friday) at 6pm
- Ch. P HW due January 20 at 6pm
- Complete Exploration 1.1 and submit it on Canvas by January 20 (Monday) at 6pm
- Read Section 1.2 of the textbook
- Complete the first two questions of Exploration 1.2: Tasting water

1.2: Measuring the Strength of Evidence

Definitions

The **Null Hypothesis** H_0 is the "by random chance alone" explanation

The **Alternative Hypothesis** H_A is the explanation that consists of a not-random effect.

> This is usually the thing we want to show - our research hypothesis.

We want to disprove H_0, forcing us to conclude that H_A is a better (more probable) explanation.

Binary Variables

Binary variables have only two possible outcomes

- Success/Failure
- Heads/Tails
- Pass/Fail

π is used to represent the *parameter* for a long-run probability or proportion

n is the sample size or number of observational units

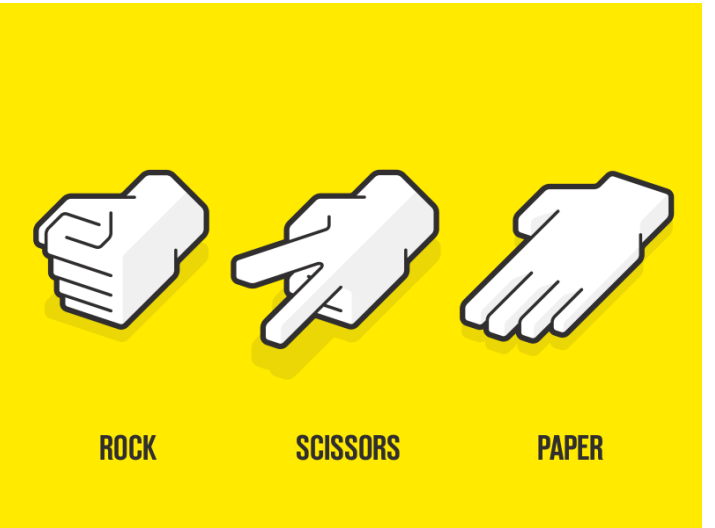
$\hat{p}$ represents our *sample statistic* (the observed value)

Rock Paper Scissors

- Two players
- Three options

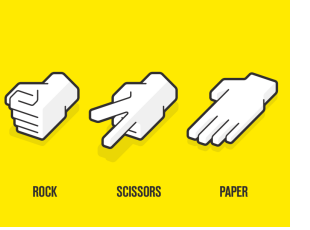
Are all options equally likely to be played?

Some evidence* suggests that novice players are less likely to choose scissors.



*See Eyler et al., 2009

Rock Paper Scissors



1. Research question: Are novice players are less likely to choose scissors?

2. Design a study and collect data:
   Play n rounds, recording 0 or 1, where 1 = scissors was played

   - Observational Unit: Each round of play
   - Variable: Outcome of a round (rock, paper, scissors) Categorical
   - Outcomes: Scissors or not scissors (rock, paper) Binary
   - Statistic: Let x=# scissors selected in n rounds; then $\hat{p} = \frac{x}{n}$

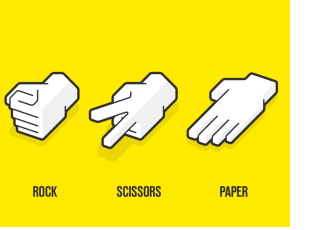
     > In our study, $n = 12, x = 2$

   - Parameter of interest: Probability of selecting scissors
     Options:
     - Equal preference for all 3 choices, so $\pi = 1/3$
     - $\pi_{\text{scissors}} < 1/3$, that is, scissors is chosen less frequently

       Note: You don't specify a particular value, just a direction from a particular value

3. Explore the data: 2 scissors in 12 trials; 2/12 = 1/6 ≈ 0.1667

Rock Paper Scissors



1. Research question

2. Design a study and collect data

   - Outcomes: Scissors or not scissors (rock, paper) Binary
   - Statistic: $\hat{p} = \frac{x}{n}$
   - Parameter of interest: Probability of selecting scissors
     Options:
     - Equal preference for all 3 choices, so $\pi = 1/3$
     - $\pi_{\text{scissors}} < 1/3$, that is, scissors is chosen less frequently

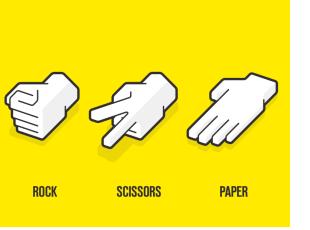
       Note: You don't specify a particular value, just a direction from a particular value

3. Explore the data

4. Draw inferences beyond the data:

Your parameter options represent the null and alternative hypotheses.
Which is which?

Rock Paper Scissors



1. Research question

2. Design a study and collect data

   - Outcomes: Scissors or not scissors (rock, paper) Binary
   - Statistic: $\hat{p} = \frac{x}{n}$
   - Parameter of interest: Probability of selecting scissors
     Options:
     - Equal preference for all 3 choices, so $\pi = 1/3$
     - $\pi_{\text{scissors}} < 1/3$, that is, scissors is chosen less frequently

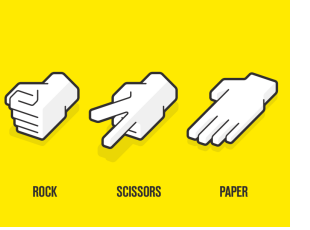
       Note: You don't specify a particular value, just a direction from a particular value

3. Explore the data

4. Draw inferences beyond the data:

$$H_0: \ \pi = 1/3$$
$$H_A: \ \pi < 1/3$$

Rock Paper Scissors



- Is $\hat{p}_{scissors}$ less than the probability specified in the null hypothesis?

  Yes, the sample proportion of 1/6 is in the direction of the alternative hypothesis

- Is it possible that $\hat{p}$ could turn out this small even if H_0 were true?

  Yes, it is possible that it could be this small

P-values

The **p-value** is the probability of obtaining a value of the statistic at least as extreme as the observed statistic *when the null hypothesis is true*

We estimate a p-value by finding the

proportion of simulated statistics

- using the random chance model
- assuming H_0

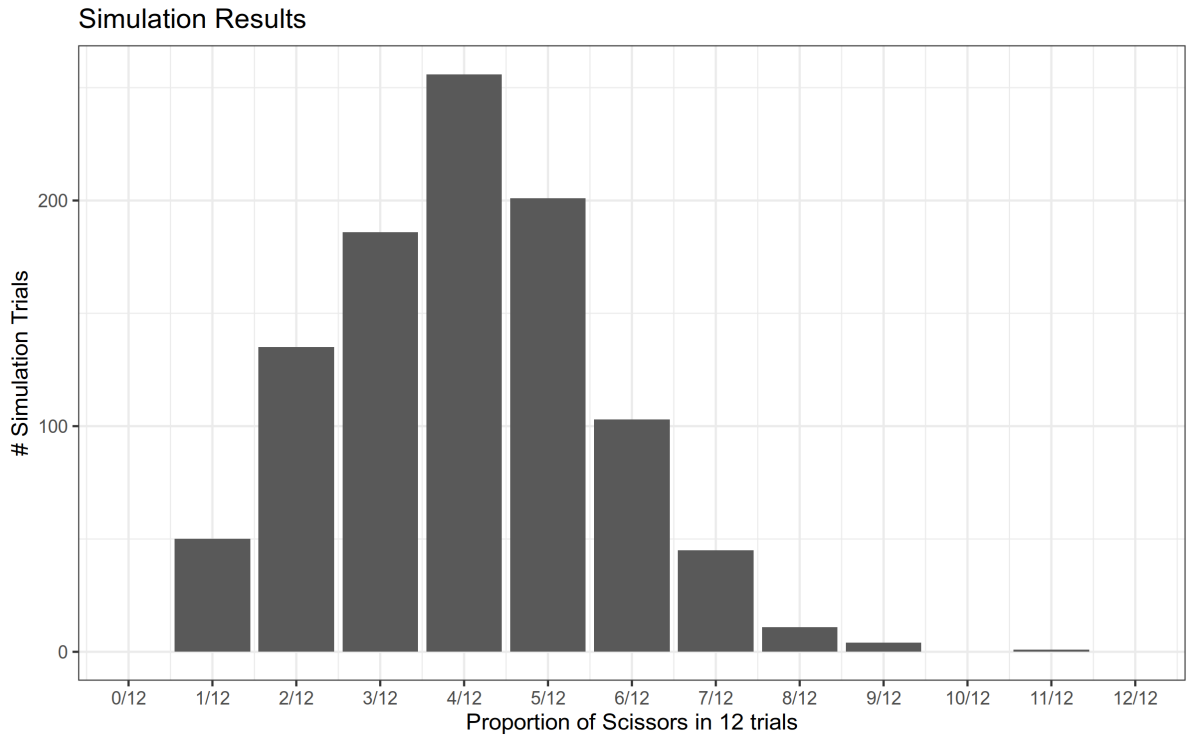
that are at least as extreme

- (in the direction of H_A)

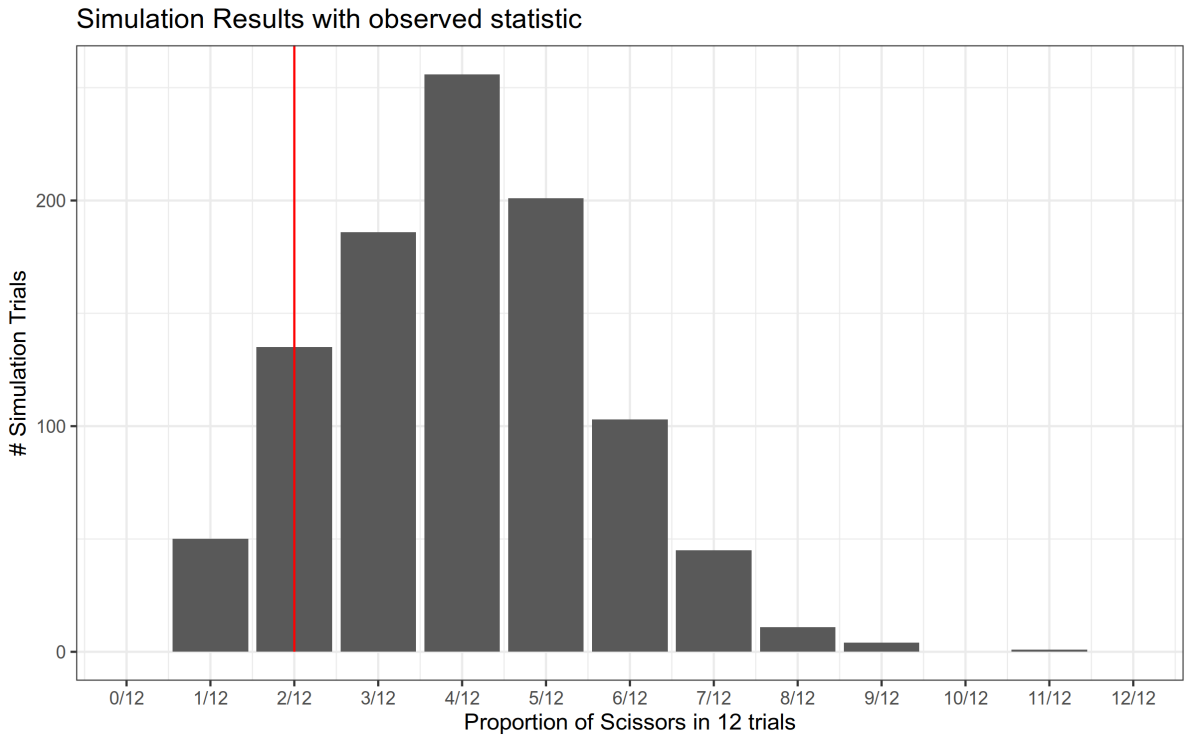
as the value of the observed statistic from the research study.

Note that the **p-value** is not the same p as the sample proportion, $\hat{p}$.
p-values can be computed using any sample statistic, not just proportions.

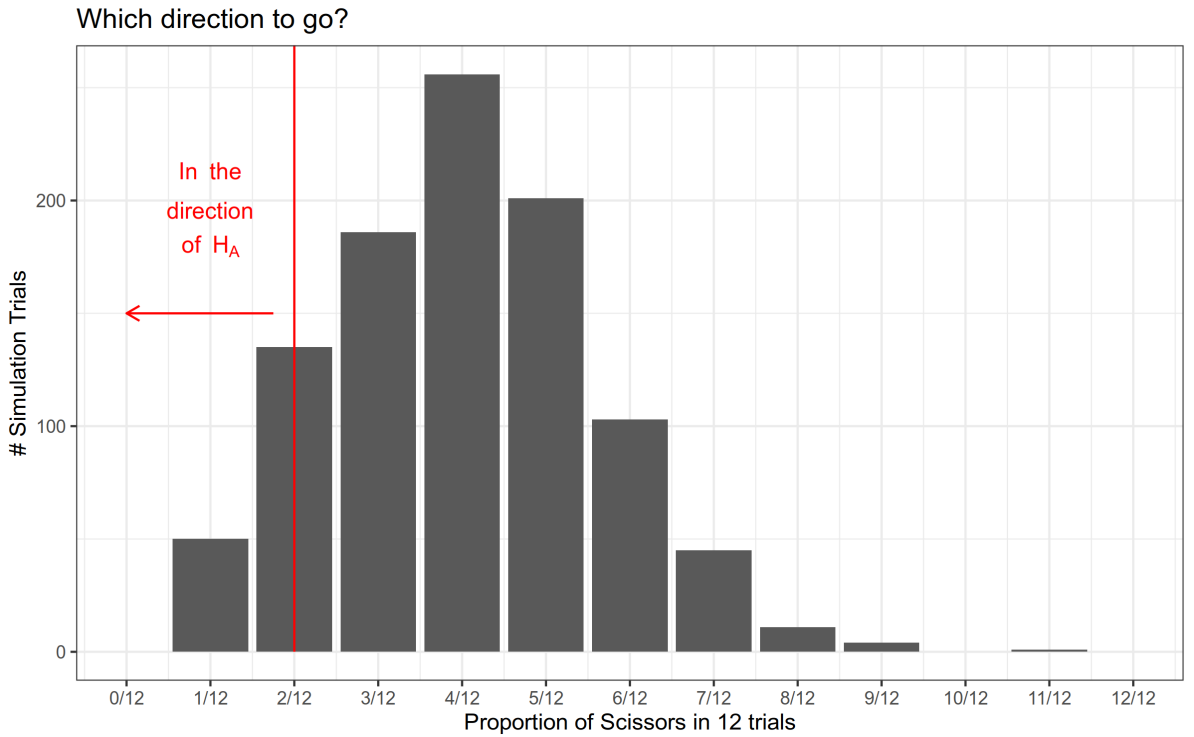
P-values


Simulation Results
Simulation Trials
200
100
0
0/12
1/12
2/12
3/12
4/12
5/12
6/12
7/12
8/12
9/12
10/12
11/12
12/12
Proportion of Scissors in 12 trials

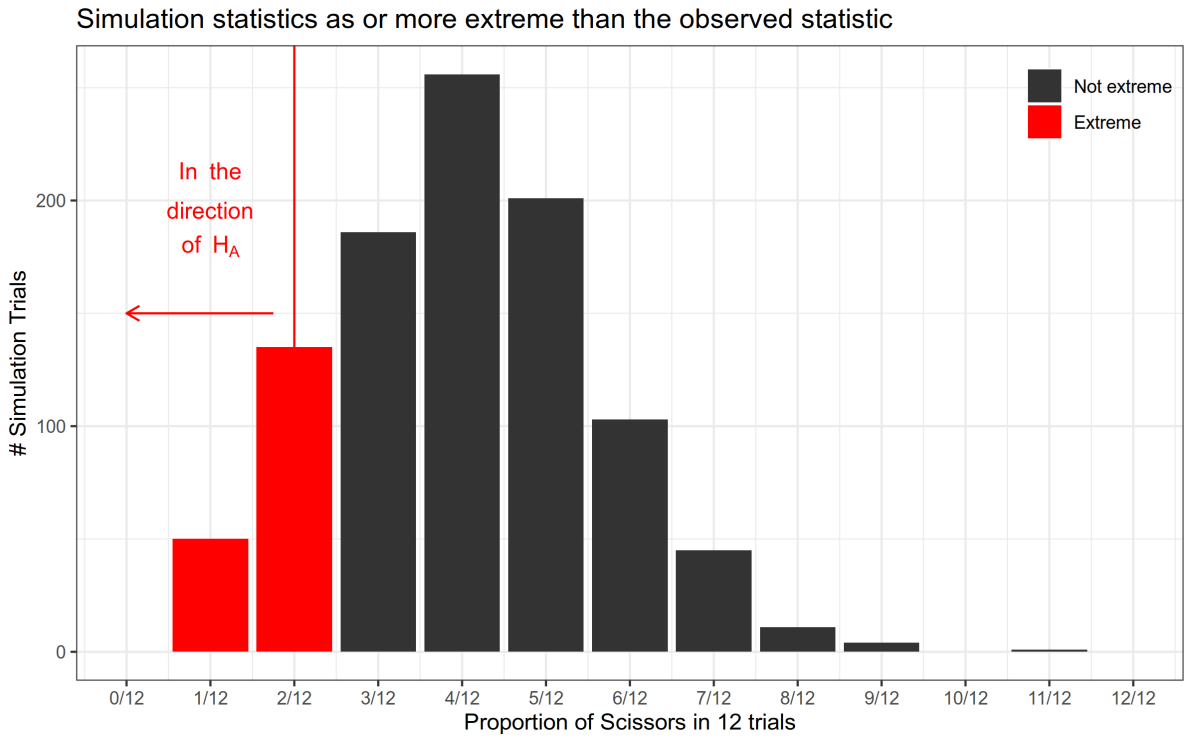
P-values


Simulation Results with observed statistic
Simulation Trials
200
100
0
0/12
1/12
2/12
3/12
4/12
5/12
6/12
7/12
8/12
9/12
10/12
11/12
12/12
Proportion of Scissors in 12 trials

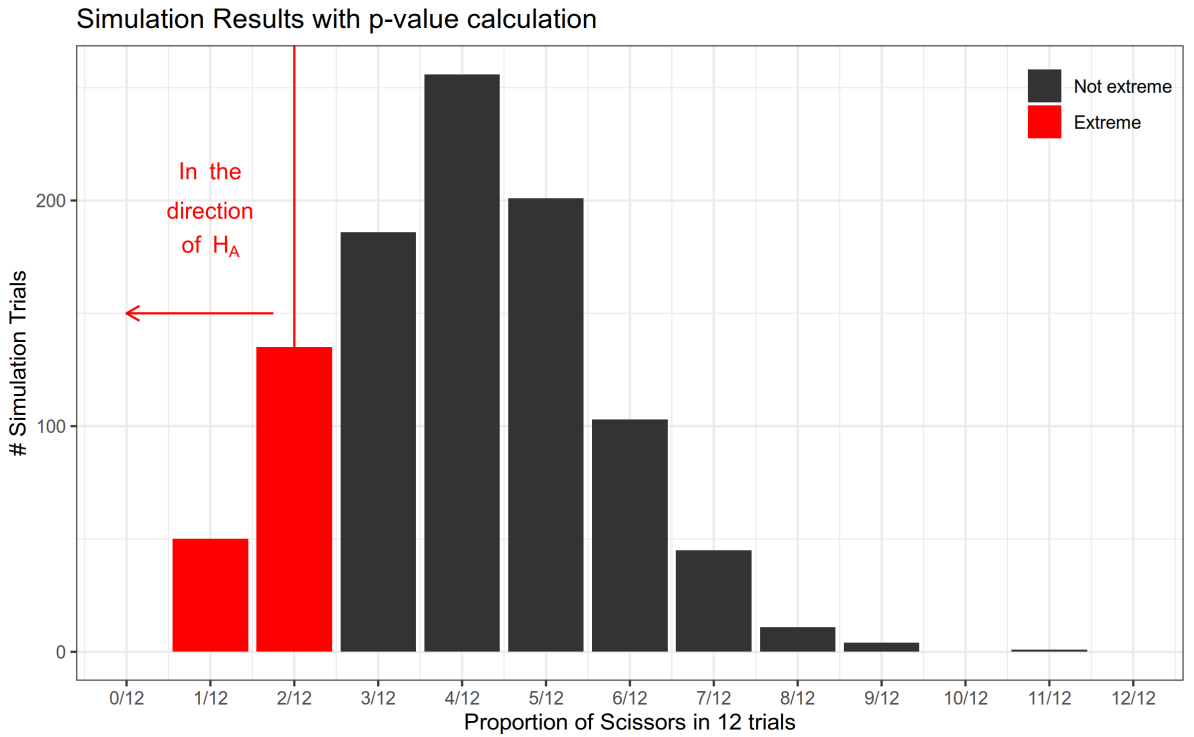
P-values


Which direction to go?
In the direction of H_A
Simulation Trials
200
100
0
0/12
1/12
2/12
3/12
4/12
5/12
6/12
7/12
8/12
9/12
10/12
11/12
12/12
Proportion of Scissors in 12 trials

P-values


Simulation statistics as or more extreme than the observed statistic
Not extreme
Extreme
In the direction of H_A
Simulation Trials
200
100
0
0/12
1/12
2/12
3/12
4/12
5/12
6/12
7/12
8/12
9/12
10/12
11/12
12/12
Proportion of Scissors in 12 trials

P-values



$$\underbrace{P(\hat{p} \leq 2/12)}_{\text{under random chance}} \approx P(\underbrace{p^* \leq 2/12}_{\text{simulation}}) = \frac{193}{1000} = .193$$

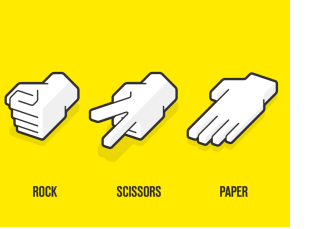
Guidelines for Interpreting p-values

p-value range	Interpretation
$0.10 < p$	Not much evidence against H_0 (H_0 is plausible)
$0.05 < p \leq 0.10$	Moderate evidence against H_0
$0.01 < p \leq 0.05$	Strong evidence against H_0
$p < 0.01$	Very strong evidence against H_0

If you have done the modeling correctly, you'll still be wrong $(100 \times p)$% of the time when H_0 is actually true

The smaller the p-value, the more evidence against H_0

Rock Paper Scissors



3 S Strategy:

1. Statistic: $2/12 \approx 0.1667$

2. Simulation: "coin flip" with p(heads) = 1/3

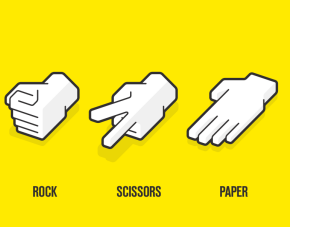
   Actual coins don't work, but we could use dice - rolls of 1 or 2 represent scissors... or we could let the computer do it for us using the **One Proportion** applet

3. Strength of evidence:

   193 of 1000 simulation samples showed a value of 0, 1, or 2 scissors (out of 12 games)

   ~19% of the time, we would see a value as or more extreme than 2 under the random chance model

Rock Paper Scissors



$$\underbrace{P(\hat{p} \leq 2/12)}_{\text{under random chance}} \approx P(\underbrace{p^* \leq 2/12}_{\text{simulation}}) = \frac{193}{1000} = .193$$

Is this an unlikely result if H_0 is true?

No, our sample statistic is not in the tail of the distribution. There is an almost 20% chance of obtaining a result as or more extreme as this by random chance.

What would you expect to happen to the p-value if we had seen only 1 scissors in the 12 rounds of the original study?

There would be fewer simulation statistics as or more extreme than the observed statistic, so the p-value would be smaller.

How much smaller?

$$P(\hat{p} \leq 1/12) \approx P(p^* \leq 1/12) = \frac{58}{1000} = 0.058$$

If we had originally observed only 1/12 scissors, we would have moderate evidence against the null hypothesis.

Exploration 1.2: Tasting Water

People spend a lot of money on bottled water. But do they really prefer bottled water to ordinary tap water?

Researchers at Longwood University investigated this question by presenting people who came to a booth at a local festival with four cups of water.

Three cups contained different brands of bottled water, and one cup was filled with tap water.

Each participant (person) was asked which of the four cups of water they most preferred.

Researchers kept track of how many people chose tap water in order to see whether tap water was chosen significantly less often than would be expected by random chance.

Exploration 1.2: Tasting Water

Step 1: Ask a research question

What is the question the researcher hoped to answer?

Exploration 1.2: Tasting Water

Step 2: Design a study and collect data

Identify the observational units in this study

In groups, continue with Exploration 1.2. Answer questions 1-24.

Hand in one submission per group on Canvas by Jan 24 at 6pm.

Review

Review

The p-value is the _____ of obtaining a value for the statistic which is as or more _____ as the observed statistic when _____ is true

Review

If the p-value is ≤ 0.05, we have _____ evidence _____ the null hypothesis (suggesting the phenomenon is NOT random)

- We are _____ likely to see our observed result if the process happens by random chance.
- **Written conclusion**: With a p-value of ____, I _____ the null hypothesis and conclude the alternative (in the context of the problem)

Review

If the p-value is > 0.05, we have _____ evidence against the null (suggesting the phenomenon could plausibly occur due to chance).

- **Written conclusion**: With a p-value of _____, I _____ the null hypothesis. We do not have enough evidence to conclude that the alternative hypothesis is more plausible than the null hypothesis that the effect occurred by chance alone.
- This does not mean we know the _____!

Review

- P-value calculation
    - Want to calculate statistics as or more extreme than our observed statistic, in the direction of the alternative hypothesis
    - Compare our statistic (from the data) to the randomly simulated statistics
- When solving questions
    - Calculate the p-value with the applet
    - Is the p-value less than or greater than 0.05?
    - Should I reject or fail to reject?
    - Make conclusion
    - Interpret in the context of the problem

1.3: Alternative Measure of Strength of Evidence

Standardized Statistics

Example 1.3 - Heart Transplant Operations

- Sudden spike in heart transplant mortality rates at St. George's Hospital in London
- Of the last 10 transplants, 80% resulted in deaths within 30 days of the transplant
    - Over 5x the national average mortality rate of 15% for the procedure
- **Research question:**

- **Observational Units**:
- **Variable**:
- **Parameter**:

- **Statistic**:

Example 1.3 - Heart Transplant Operations

- Binary variables are often coded as 0 = failure, 1 = success, where the choice of 0/1 is arbitrary or based on the desired statistic (survival rate vs. mortality rate)
- In medical and epidemiological studies, 1 or success may be used to denote "the patient died" because X-day mortality rate may be more informative

Example 1.3 - Heart Transplant Operations

- **Null hypothesis**

- **Alternative hypothesis**

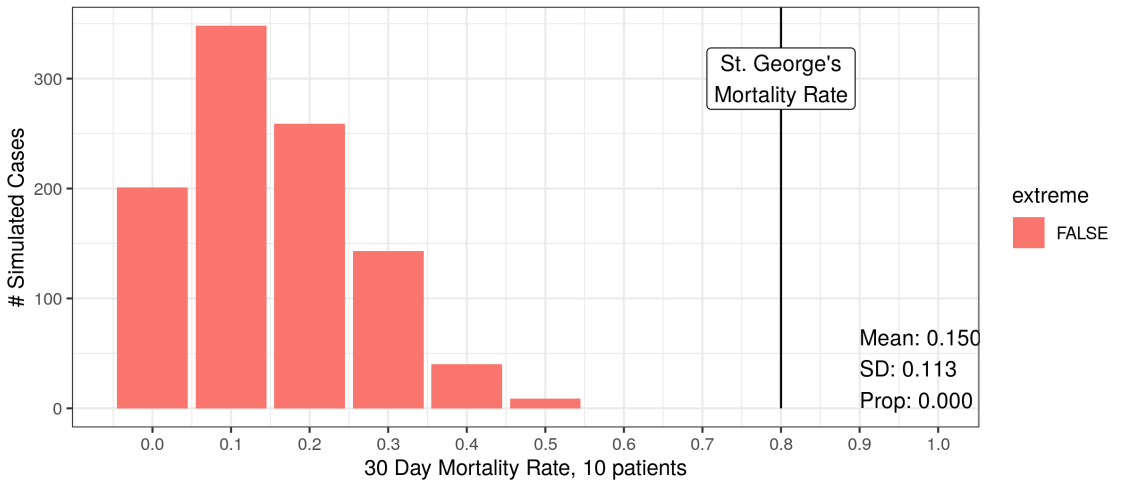
Example 1.3 - Heart Transplant Operations

In your groups, use the online applet to conduct a relevant simulation, and apply the 3 S's strategy to the problem:

- Statistic
- Simulate
- Strength

Example 1.3 - Heart Transplant Operations

- Statistic: 0.80 (8/10)
- Simulate: __ trials per simulation, with success probability ____



Simulated 30 day mortality rate

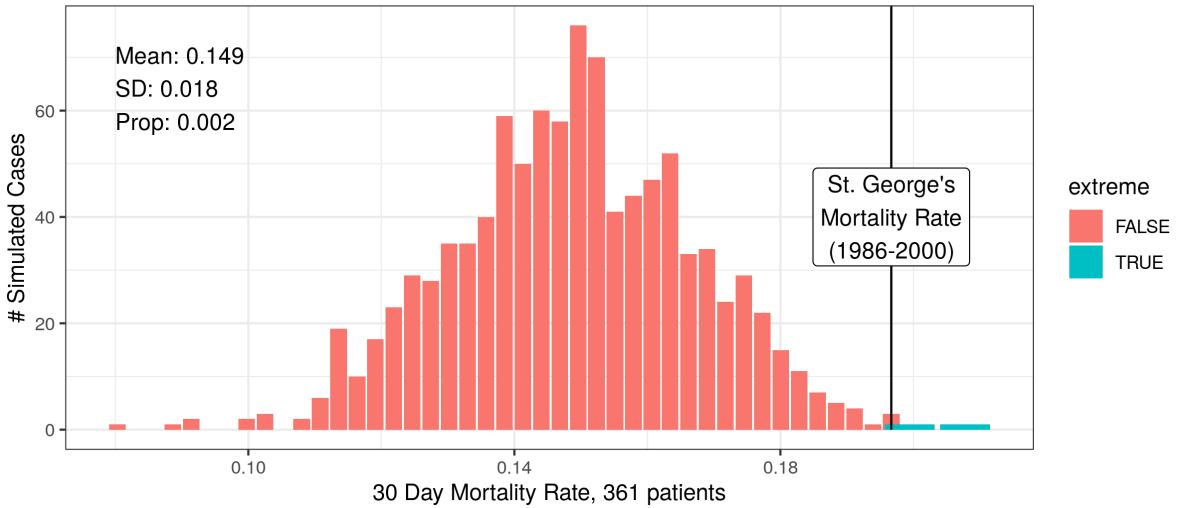
- Strength: An as or more extreme result than our statistic, assuming the national mortality rate of 15%, occurs with probability < 0.001

  We only conducted 1000 simulations, so we can't say the p-value is equal to 0. We can say it is less than 1/1000. Remember that this approach is an *approximation*

Example 1.3 - Heart Transplant

- Could it be that people are only paying attention because of a run of bad luck?
- Has something changed in the hospital recently?
- Has St. George's historically had a higher mortality rate than the national average?
    - Previous 361 heart transplants (dating back to 1986) - 71 of the patients died within 30 days of the transplant

Example 1.3



Simulated 30 day mortality rate

- Our simulated p-value is $P(x > \frac{71}{361}) = 0.01$

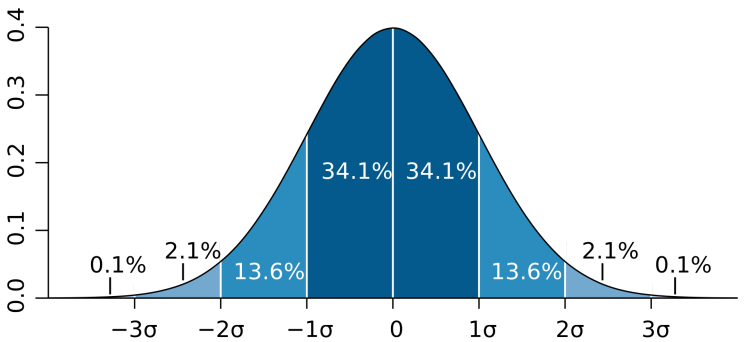
An Alternative to the P-value: Standardized Value of a Statistic

Standardized statistic - a measure of how far an observed statistic is from the mean of the distribution

- Commonly denoted by z
- $z = \frac{\text{Statistic} - \text{Mean of null distribution}}{\text{standard deviation of null distribution}}$

An Alternative to the P-value: Standardized Value of a Statistic

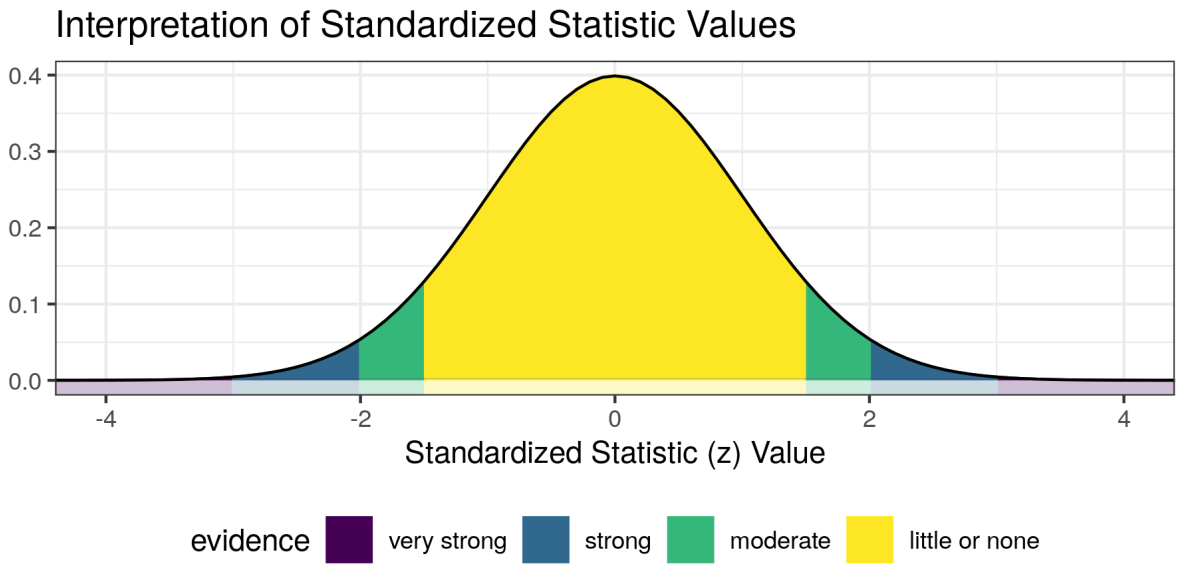
Standard Deviation - a measure of the distance a "typical" value is away from the mean



Observations more than 2 or 3 standard deviations from the mean are "in the tail of the distribution" or extreme

Guidelines for Interpreting Standardized Statistics

z-value range	Interpretation
$\lvert z\rvert < 1.5$	**little or no** evidence against H_0
$1.5 \leq \lvert z\rvert < 2$	**moderate** evidence against H_0
$2 \leq \lvert z\rvert < 3$	**strong** evidence against H_0
$3 \leq \lvert z\rvert$	**very strong** evidence against H_0



An Alternative to the P-value: Standardized Value of a Statistic

Standardized statistic - a measure of how far an observed statistic is from the mean of the distribution

- Commonly denoted by z
- $z = \frac{\text{Statistic} - \text{Mean of null distribution}}{\text{standard deviation of null distribution}}$ (Equation 1.5)

Indicates how many standard deviations from the observed value to the hypothesized process mean

Ex: St. George

$n = 10$

- **Statistic**:
- **Mean of H_0**:
- **Std. Dev of H_0**:
- **z**

$n = 361$

- **Statistic**:
- **Mean of H_0**:
- **Std. Dev of H_0**:
- **z**

Ex: St. George

$$n = 10$$

- **Statistic**: 0.8
- **Mean of H_0**: 0.15
- **Std. Dev of H_0**: 0.1134587

$$z = \frac{0.8 - 0.15}{0.113} \approx 5.729$$

$$n = 361$$

- **Statistic**: $\frac{71}{361} = 0.1966759$
- **Mean of H_0**: 0.15
- **Std. Dev of H_0**: 0.0183464

$$z = \frac{0.197 - 0.15}{0.018} \approx 2.544$$

Standardized Statistic

- Applet will calculate z as well
- Normally use _____ as a single cut-off
  - z can be _____
  - The _____ the standardized statistic is from zero, the stronger the evidence against the null model

Summary - Standardized Statistics

Standardized statistic far from 0 $\rightarrow$ _____ p-value $\rightarrow$ _____

Standardized statistic close to 0 $\rightarrow$ _____ p-value $\rightarrow$ _____

Standardized Statistic

- Can calculate both p-value and standardized statistic
- Quizzes/Exams may have either one, or both
- P-value and standardized statistic should always give the same conclusion
- Standardized Statistic takes into account sample size and spread of the data - helpful to compare multiple samples with different sample sizes

What you need to find Standardized Statistic

- Observed statistic (from the sample)
- Mean of null distribution
- Standard deviation of null distribution

  May be given to you in the question, from the applet, or from equation 1.5

$$z = \frac{\text{Statistic} - \text{Mean of null distribution}}{\text{standard deviation of null distribution}} \quad \text{(Equation 1.5)}$$

1.4: What Impacts Strength of Evidence?

Bob vs. Tim

Is the man on the left named Tim or Bob?

Link to full study

What Impacts the Strength of Evidence?

- Bob/Tim: 85% of people say Tim is on the left
  - Compared to .5 as the null (chance) value
- What happens as the statistic gets farther away from .5?
  - **p-value** gets _____
  - **standardized statistic** gets _____
- Closer to 0.5?
  - **p-value** gets _____
  - **standardized statistic** gets _____

What Impacts the Strength of Evidence?

1: Difference between the statistic and the null hypothesis

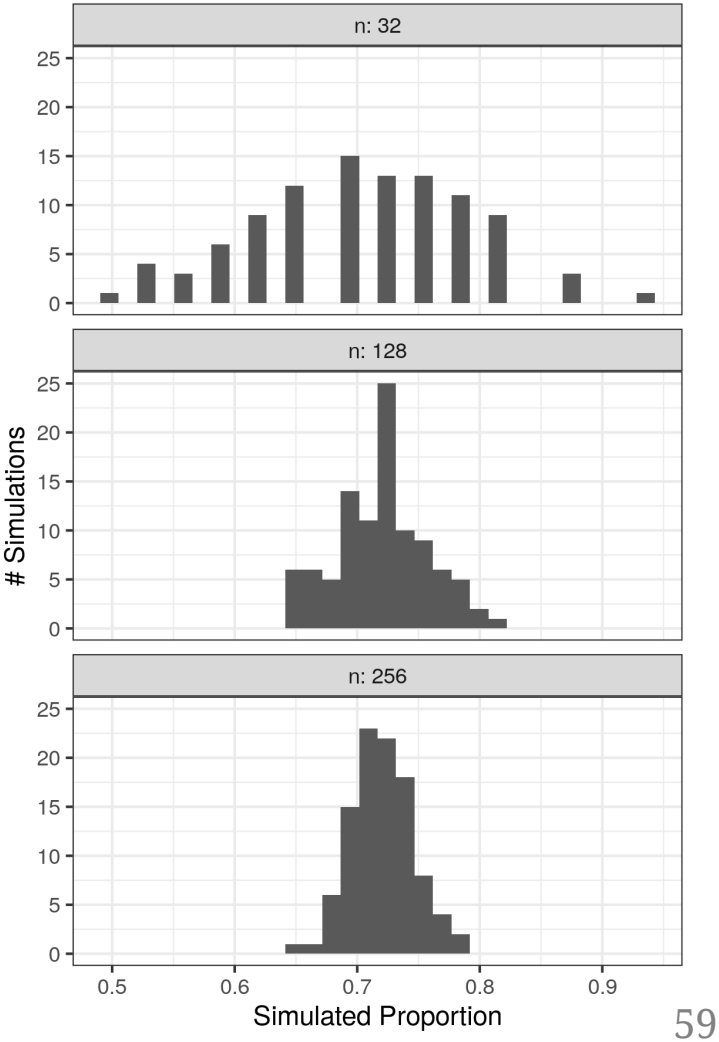
- Farther apart: Random chance/null is less plausible
    - p-value gets smaller
    - standardized statistic gets larger
- Closer together: Random chance becomes more plausible
    - p-value gets larger
    - standardized statistic gets smaller

What Impacts the Strength of Evidence?

2: Sample size

- What happens as sample size increases (holding the statistic constant)?
- Depends on standard deviation, variability, and the spread of the data

  Assume the statistic is 0.6. What happens to the strength of evidence against the null as the sample size increases?

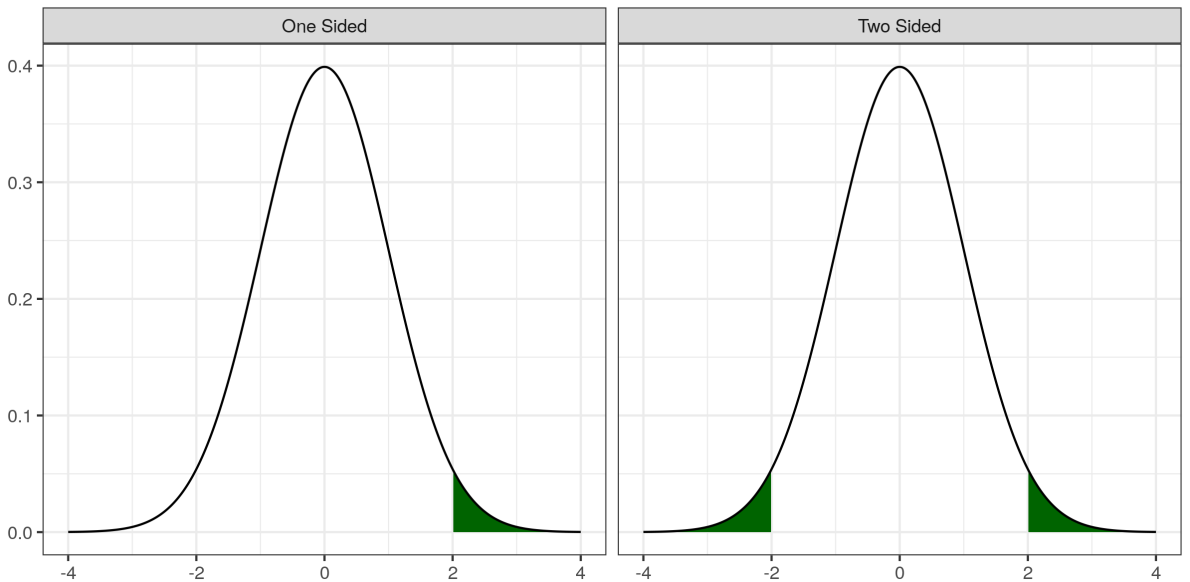


What Impacts the Strength of Evidence?

3: One-sided vs. Two-sided alternatives

Two-sided alternative:

- don't care about the direction of the effect - only care about the magnitude difference from the null



- Using a two-sided hypothesis causes p-values to _____

Scenarios Worksheet

- Write the hypotheses in symbols for all five scenarios
- Choose 1 to complete a simulation on. Find the p-value, standardized statistic, and draw a conclusion.

This is practice, so ask questions!

Due dates

- Exploration 1.2 is due at 6pm on Friday, Jan 24
- WileyPlus Chapter 1 Homework is due Feb 3 at 6pm
- Quiz over Ch. P and Ch. 1 is due Feb 5 at 11:59 pm

Review

Review

- Standardized Statistic
- What impacts strength of evidence? - Distance between hypothesis and statistic
    - Sample size
    - One sided vs. Two sided tests

Standardized Statistic

- We can't use raw values to determine significance

Example: Determining if coins are fair

- Flip a coin twice and get heads both times
  - $H_0 : \pi = 0.5$
  - $\hat{p} = 1$
  - $|\hat{p} - \pi_0| = 0.5$
- Flip a coin 100 times and get heads 80 times
  - $H_0 : \pi = 0.5$
  - $\hat{p} = .8$
  - $|\hat{p} - \pi_0| = 0.3$
- Standardization allows us to add some context (the mean and spread of the null distribution) to our estimate

Standardized Statistic

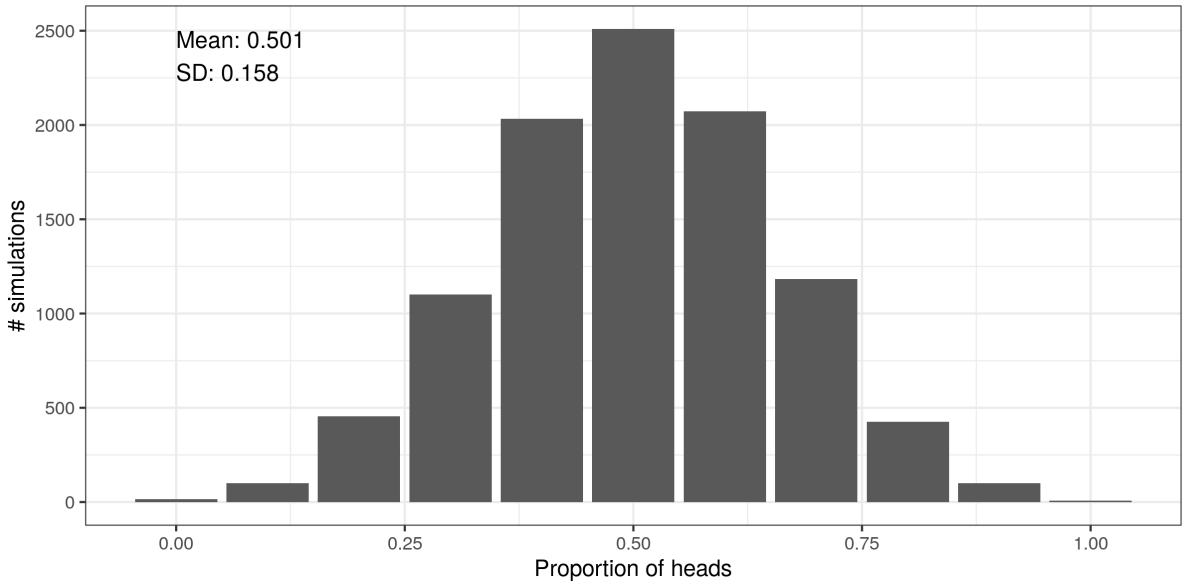
- Length of I-80 traveling across NE: 455.3 miles
- Distance from LA to Hawaii: 2491 miles

How many times could you fit the state of Nebraska in between LA and Hawaii?

Roughly 5.47 times

Standardized Statistic Practice

- Flip a coin 10 times and get heads 8 times
  - What value is H_0 centered at?
  - How many standard deviations could fit between the null hypothesis value and the observed value?
  - Is this value significant?



What Impacts Strength of Evidence?

- Strength of evidence - difference between the observed statistic and the null hypothesis

Which would give us more evidence against H_0?

- Flipping a coin 100 times and getting 55 heads
- Flipping a coin 100 times and getting 70 heads

What Impacts Strength of Evidence?

- Strength of evidence - difference between the observed statistic and the null hypothesis

Which would give us more evidence against H_0?

- Flipping a coin 5 times and getting 4 heads
- Flipping a coin 1000 times and getting 800 heads

What Impacts Strength of Evidence?

One Sided vs. Two Sided

- One Sided
  - Testing whether we get heads more often than tails
  - $H_0 : \pi = 0.5$, $H_A : \pi > 0.5$
- Two Sided
  - Testing whether a coin is fair
  - $H_0 : \pi = 0.5$, $H_A : \pi \neq 0.5$

One Sided vs. Two Sided Examples

Decide whether the test should be one sided or two sided and determine what H_A should be

- My favorite outside temperature is 75° F. Has the average temperature this year been significantly different from that?
- In 2009, a dog brought to a shelter had about a 55% chance of being adopted or returned to its owner. In 2017, the chance of adoption or return to owner was almost 70%. Has the probability of a good outcome for shelter pups increased significantly?
- A government official claims that the dropout rate for local schools is 25%. Last year, 190 out of 603 students dropped out. Is there enough evidence to reject the government official's claim?

One Sided vs. Two Sided Examples

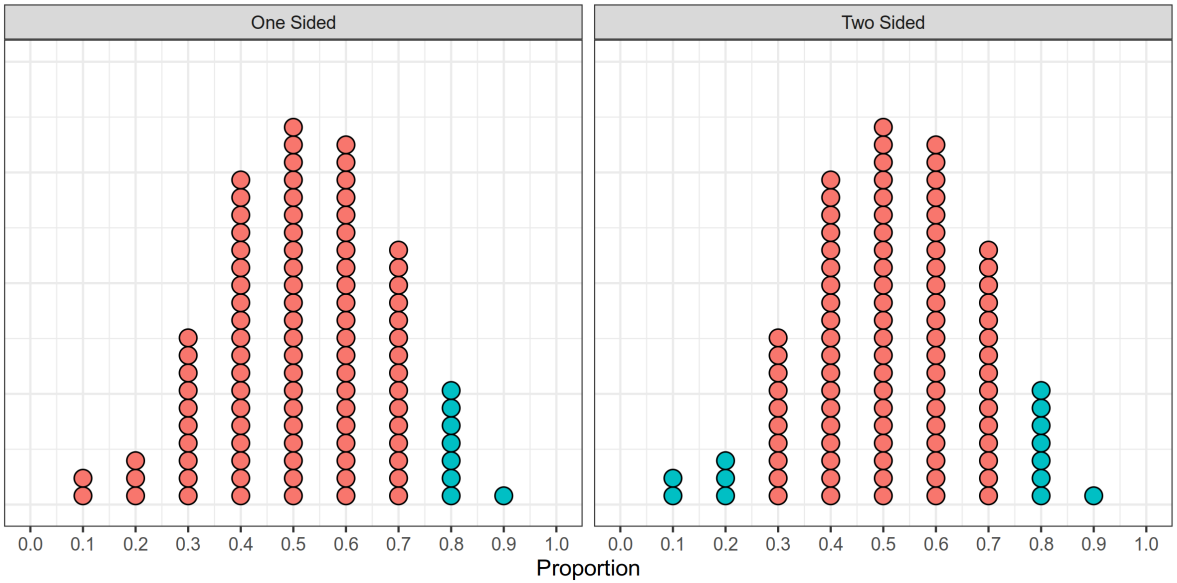
Decide whether the test should be one sided or two sided and determine what H_A should be

- My favorite outside temperature is 75° F. Has the average temperature this year been significantly **different** from that?
  Different -> $H_A : T \neq 75$ -> two-sided test

- In 2009, a dog brought to a shelter had about a 55% chance of being adopted or returned to its owner. In 2017, the chance of adoption or return to owner was almost 70%. Has the probability of a good outcome for shelter pups **increased** significantly?
  increased -> $H_A : \pi > .55$ -> one-sided test

- A government official claims that the dropout rate for local schools **is** 25%. Last year, 190 out of 603 students dropped out. Is there enough evidence to reject the government official's claim?
  is (equal to) -> $H_A : \pi \neq .25$ -> two sided test

One-Sided vs. Two-Sided tests

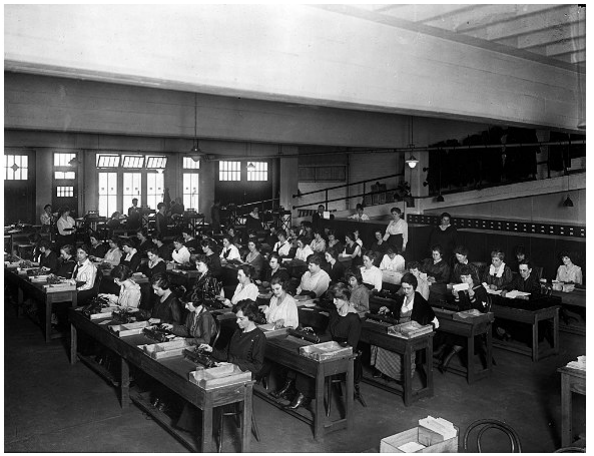


- There are 100 dots. To find the p-value, we count the dots which are as or more extreme than what we observed (0.8)
- For a 1-sided test of whether we observed more heads (0.8, 0.9, and 1.0), $p = 0.08$
- For a 2-sided test of whether the coin is fair (0.8, 0.9, 1.0, and 0.2, 0.1, 0.0), $p = 0.13$

1.5: Inference for a Single Proportion: Theory-Based Approach

Simulations vs. Theory

- Simulation provides an easy method for comparing a statistic to a model representing random chance, but it requires a lot of computer power



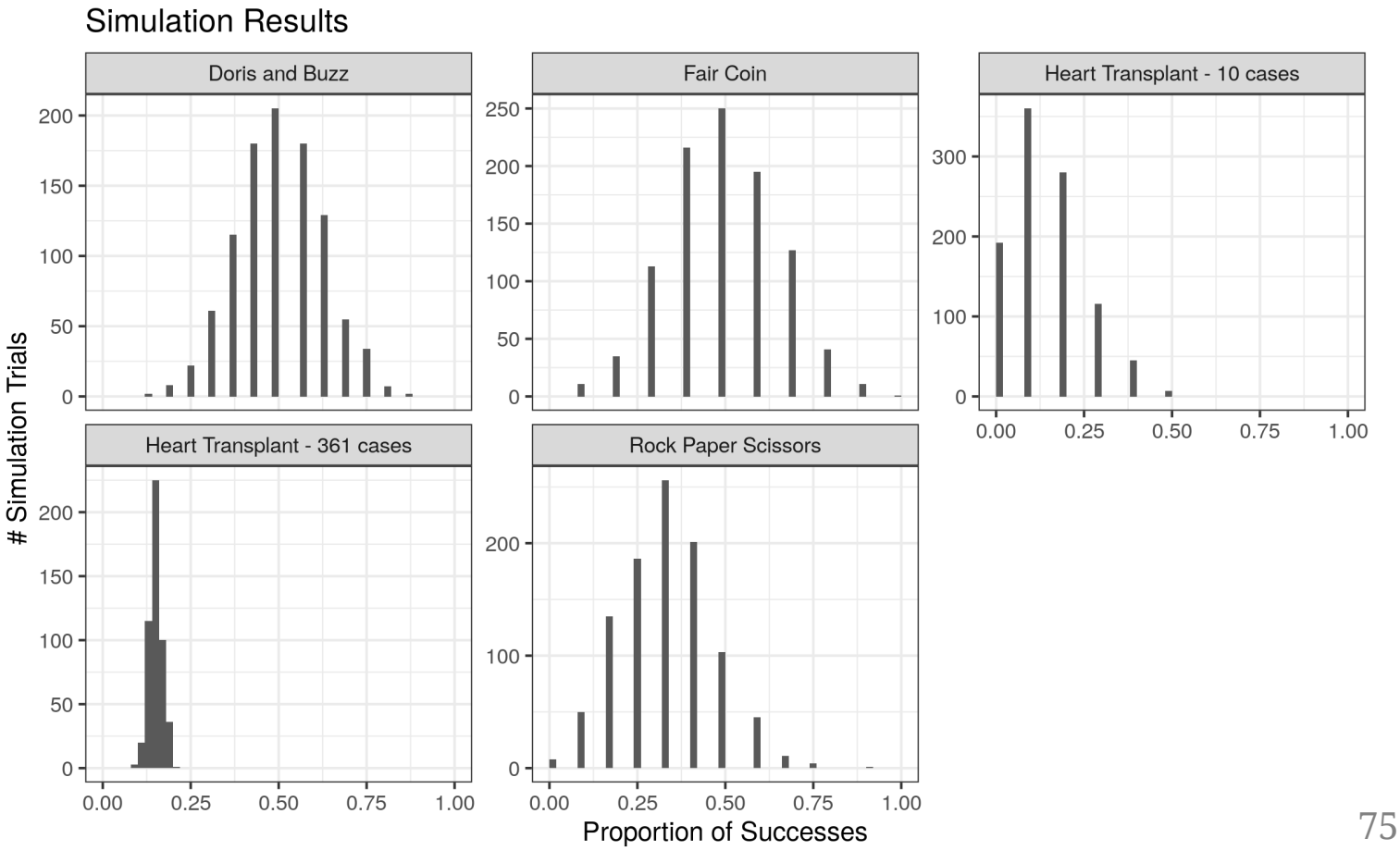
A room full of "computers" during WWII

- Historically, statisticians didn't have enough computing power for simulation, so they developed theory-based approaches to find p-values and standardized statistics

If you'd like to read more about human computers, this book is very interesting.

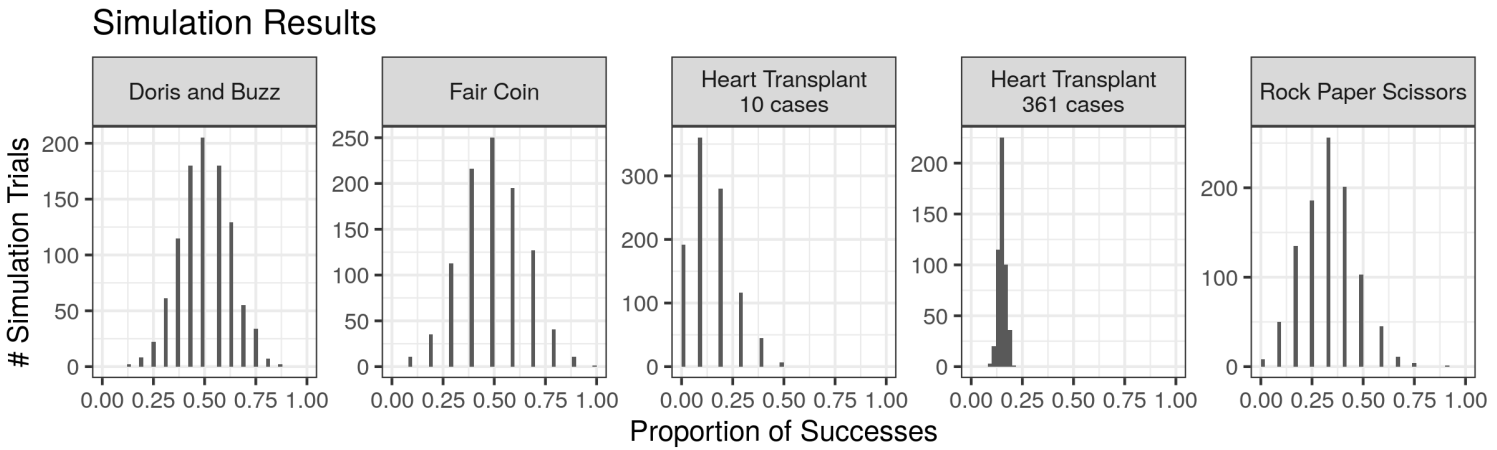
Distributions

What do the following pictures have in common? What makes the exceptions different?



Theoretical Null Distributions

Most of the null distributions we've seen so far have been fairly symmetric and "bell shaped"



The not symmetric distribution has p = 0.15 and only 10 cases. In the other examples with small n, p is 0.333 or 0.5.

Theoretical Null Distribution

When n is "big enough"*, we can use theoretical approximations instead of simulations.

Using a theoretical distribution, we can calculate p-values and standardized statistics, if we have

1. the mean of the distribution
2. the theoretical standard deviation of the distribution
3. a good idea of whether the assumptions hold

2 (and sometimes 1) will require formulas and calculations.

*More on this in a couple of slides

Theory Based approach

Based on the **Central Limit Theorem** (see Wikipedia for an illustration)

> If the sample size n is large enough, the distribution of the *sample proportion* will be approximately bell shaped, centered at the long-run probability π with the standard deviation of the null distribution.

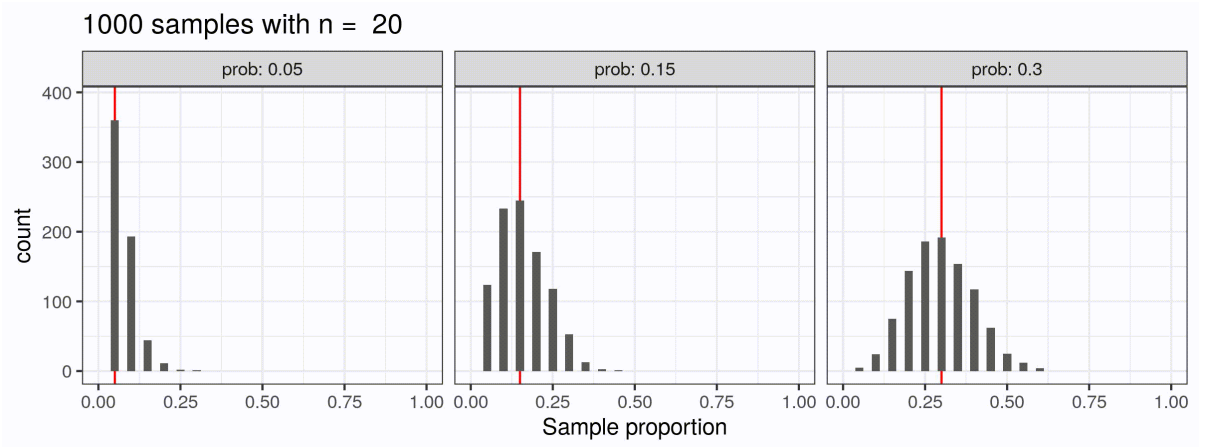
The standard deviation of the null distribution, σ, will also be used in the standardized statistic calculation.

We need a formula for σ because we are no longer doing simulation.

How large is large enough?

- Our book: if the number of successes and failures are at least 10
- Others: $n > 30, n > 50, ...$
- Depends on the discipline (and likely values for p within the discipline)

In the animation below, when does each distribution appear to be centered around the red line?



Example

St George: $\pi = 0.15, \hat{p} = 0.197, n = 361$

New SD? Standardized Statistic?

Simulation approach: SD = 0.018 and standardized statistic = 2.544

Simulation results could be different for everyone.

Normal Distribution

- Replaces the simulated null distribution
- Fits curve to what we'd expect to see in the data given random chance
- Used to find p-values

> Technically, we'd need calculus to find p-values, but the applet will do it for us.

- We can see if it's a good approximation or not by comparing to the simulation results

> The prediction is valid if $n\pi \geq 10$ and $n(1-\pi) \geq 10$
> (**Validity condition** for 1-proportion problems)

For proportions, the normal distribution has a mean of π and a standard deviation of $\sqrt{\pi(1-\pi)/n}$

Example 1.5: Halloween Treats

Do children show any preference for candy vs. toys?

Measure the proportion of children choosing candy, compare to $\pi = 0.5$

- H_0 :
- H_A :

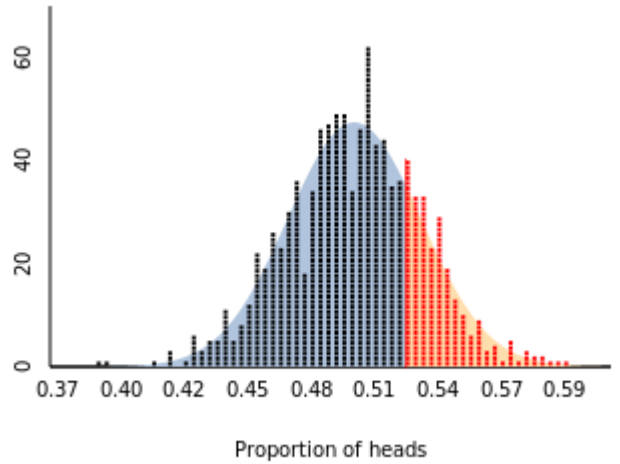
Data:

- Sample Size $n = 283$
- Sample Statistic $\hat{p} = 148/283 = 0.523$

Use the app to find the standard deviation of the null distribution. How does it compare to the theoretical standard deviation?

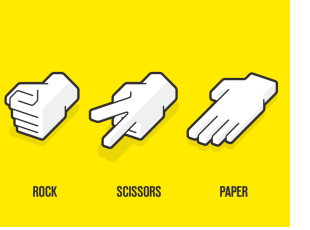
Example 1.5: Halloween Treats

Does the distribution look like one where a bell shaped curve will approximate it?



Quantity	Theory	Simulation
SD		
Z-score		
Decision		

Example: Rock Paper Scissors



$n = 12$, with 10 successes and 2 failures

- Are the validity conditions met? Why or why not?
- Normal distribution predictions:
    - bell shaped, approximately normal
    - null centered at .333
    - standard deviation will be:
- What happens with more/less skew?
- What happens with a smaller/larger sample size?

Summary: Chapter 1.5

- If you want to use a theory based approach:
  - need a sample size large enough to say H_0 follows a normal distribution
  - Applet: Normal approximation button (calculates p-value for you)
  - Can always calculate standardized statistic by hand
- Theory based approach doesn't always work
  - if the sample size is small, null distribution will not be approximately bell shaped
  - if the distribution is not bell shaped, the normal approximation to the p-value and the standardized statistic are not accurate
- Simulation ALWAYS works

Reminders

- Ch 1 HW due on Canvas Feb 3 at 6pm
- Quiz over Ch P and Ch 1 is due Feb 5. Please work independently on the quizzes.
- Exam 1 is Feb 13 during class and covers Ch P, Ch 1, and Ch 2.
- Read Section 2.1 for Feb. 4

Chapter 1.5 Review

You should know:

- That the Central Limit Theorem describes how the distribution of a sample statistic behaves if n is large enough
- How to calculate the standardized statistic for a proportion
- The validity conditions for a one-proportion theory based test
- How to use the app to get a theory-based p-value for a proportion (this can't be easily tested)

In Ch. 2, we will discuss cases where theory-based inference is the only reasonable option.

For now, it's enough to know that it's another method to calculate p-values and standardized statistics

Chapter 1 Review

Scientists suspect dyslexia is much more common than recorded diagnoses would indicate; some argue that up to 15% of the population may show signs of dyslexia. More conservative estimates based off of recorded diagnoses put the prevalence of dyslexia around 5%.

In order to examine the prevalence of dyslexia, researchers would like to test a sample of the adult population to see if there is evidence that the rate of dyslexia in the population is higher than 5%

- State the null and alternative hypotheses

Chapter 1 Review

Scientists suspect dyslexia is much more common than recorded diagnoses would indicate; some argue that up to 15% of the population may show signs of dyslexia. More conservative estimates based off of recorded diagnoses put the prevalence of dyslexia around 5%.

In order to examine the prevalence of dyslexia, researchers would like to test a sample of the adult population to see if there is evidence that the rate of dyslexia in the population is **higher** than **5%**

- State the null and alternative hypotheses

$$H_0 : \pi = 0.05$$
$$H_A : \pi > 0.05$$

Chapter 1 Review

Sample 250 adults and find that 23 likely have dyslexia.

- State the validity conditions for a one-proportion theory-based hypothesis test. Are they met?
- Calculate the population standard deviation, σ
- Calculate the standardized statistic, z
- Conduct a test of your hypothesis using theory-based methods. Use the theory-based inference app to get a p-value for your test.
- What is your conclusion?

Chapter 1 Review

Sample 250 adults and find that 23 likely have dyslexia.

- State the validity conditions for a one-proportion theory-based hypothesis test. Are they met?
  - Validity conditions: 10 successes and 10 failures under H_0
  - If $\pi = 0.05$, then we would expect $n(\pi) = 250(0.05) = 12.5$ people with dyslexia and $n(1-\pi) = 237.5$ people without dyslexia.
  - 12.5 > 10 and 237.5 > 10 so the validity conditions are met

Chapter 1 Review

Sample 250 adults and find that 23 likely have dyslexia.

- Calculate the population standard deviation, σ

$$\begin{aligned}\sigma &= \sqrt{\frac{\pi(1-\pi)}{n}} \\ &= \sqrt{\frac{0.05(0.95)}{250}} \\ &= \sqrt{0.0475/250} \\ &\approx 0.01378\end{aligned}$$

Chapter 1 Review

Sample 250 adults and find that 23 likely have dyslexia.

- Calculate the standardized statistic, z

$$z = \frac{\hat{p} - \pi}{\sigma}$$

$$= \frac{\frac{23}{250} - 0.05}{0.01378}$$

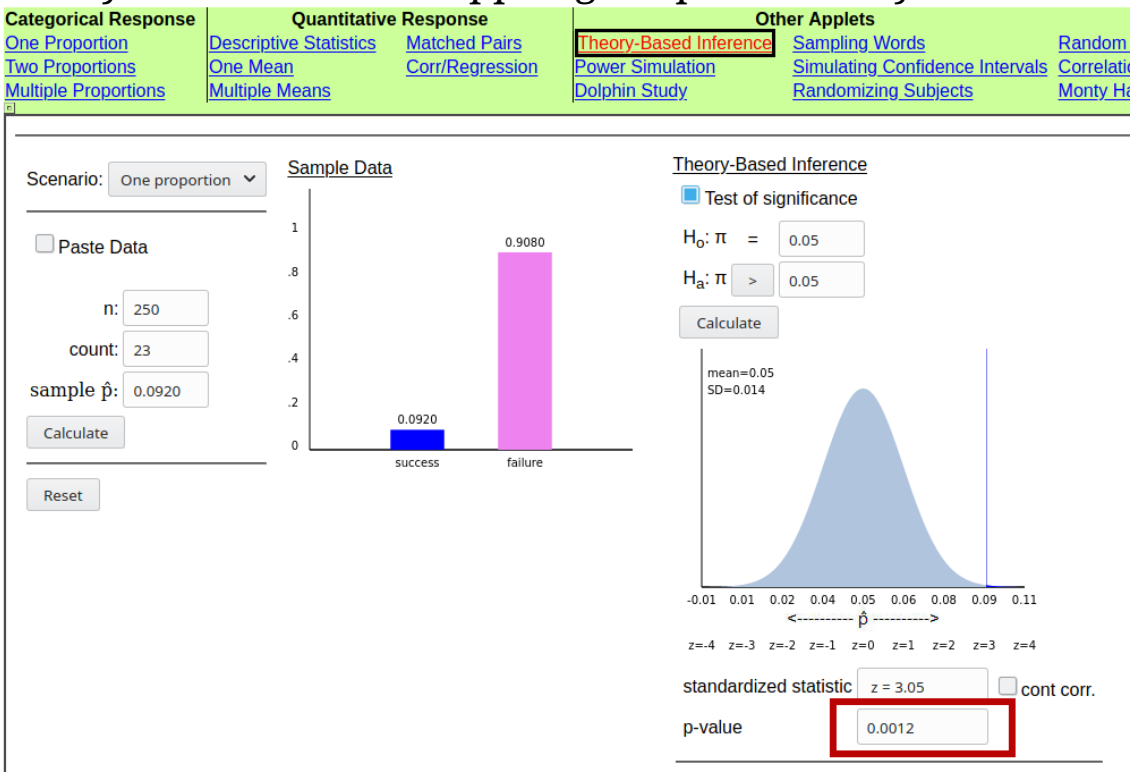
$$= \frac{0.092 - 0.05}{0.01378}$$

$$= \frac{0.042}{0.01378} \approx 3.05$$

Chapter 1 Review

Sample 250 adults and find that 23 likely have dyslexia.

- Conduct a test of your hypothesis using theory-based methods. Use the theory-based inference app to get a p-value for your test.



Chapter 1 Review

Sample 250 adults and find that 23 likely have dyslexia.

- What is your conclusion?

We have very strong evidence against the null hypothesis that the rate of dyslexia in the population is 5%. With $p \approx 0.0012$, we reject the null hypothesis and conclude that there is evidence that more than 5% of the population has dyslexia.

On the test:

We have very strong evidence against the null hypothesis that the rate of dyslexia in the population is 5%. Since $z > 3$, we know that our observed result would occur in the tail of the distribution under H_0. Thus, we reject the null hypothesis and conclude that there is evidence that more than 5% of the population has dyslexia.

Chapter 1 Review

You conduct a second sample of 120 adults, and this time, you find that 14 individuals have dyslexia.

- Does this second sample contradict the results of the first sample? Why or why not?
- Use the one proportion applet to conduct 1000 simulations of this experiment. Calculate or use the app to get the following:
    - Sample proportion
    - Population standard deviation
    - Standardized statistic
    - p-value

Chapter 1 Review

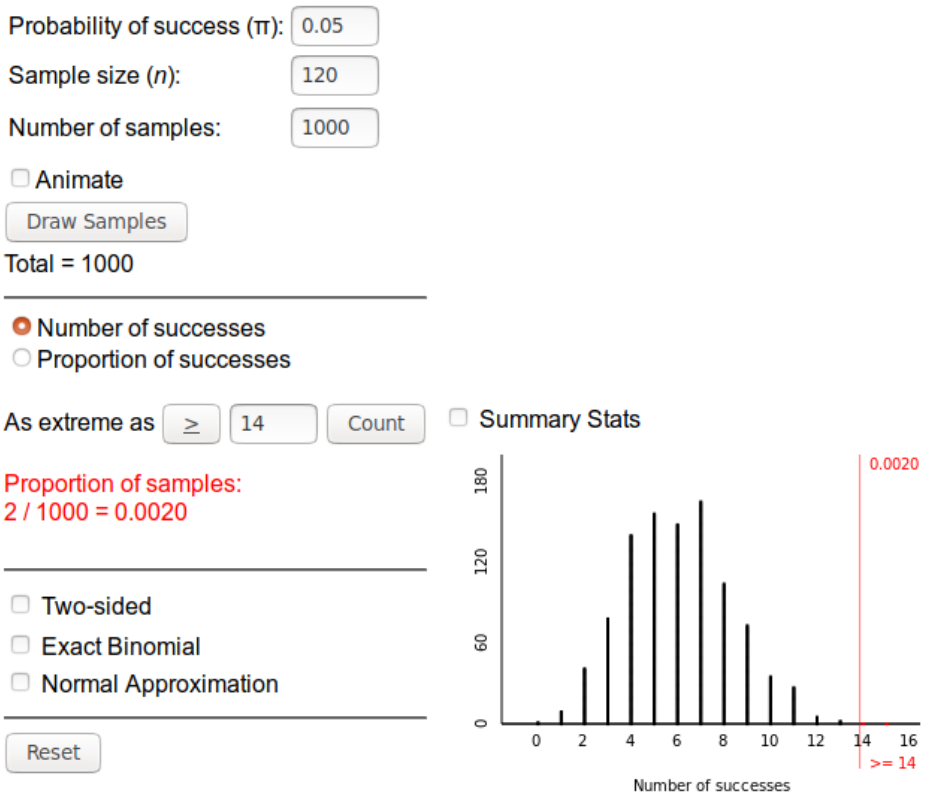
You conduct a second sample of 120 adults, and this time, you find that 14 individuals have dyslexia.

- Does this second sample contradict the results of the first sample? Why or why not?

No, this sample does not contradict the results of the first sample. We would expect that different random samples would contain different individuals and thus lead to a different result.

Chapter 1 Review

- Use the one proportion applet to conduct 1000 simulations of this experiment.



Chapter 1 Review

Calculate or use the app to get the following:

- Sample proportion: $\hat{p} = 14/120 = .1167$
- Population standard deviation:

$$\sigma = \sqrt{\frac{\pi_0(1-\pi_0)}{n}} = \sqrt{\frac{.05(.95)}{120}} \approx 0.0199$$

- Standardized statistic

$$z = \frac{\hat{p} - \pi_0}{\sigma} = \frac{.1167 - .05}{.0199} = 3.35$$

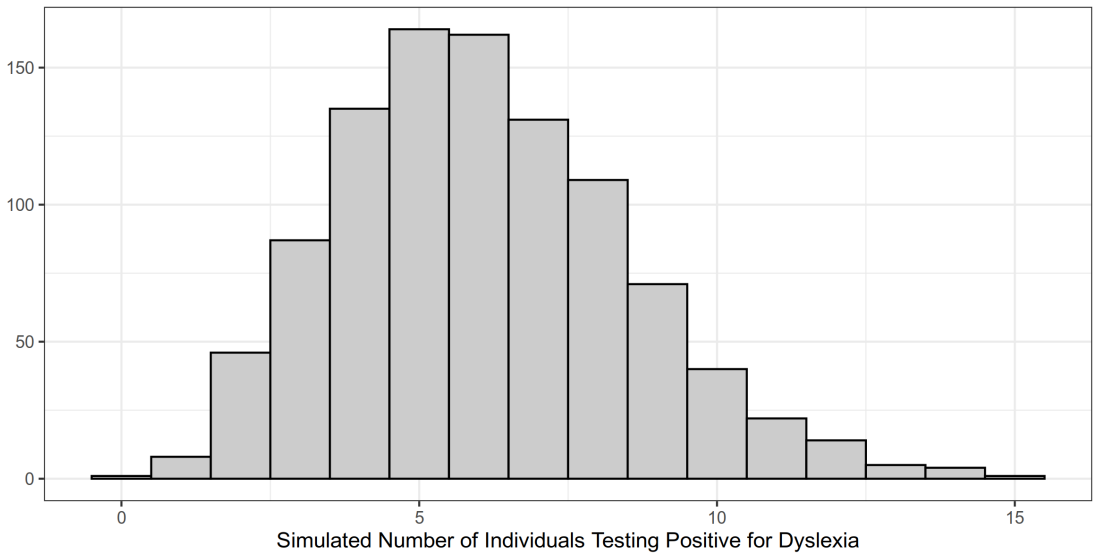
- p-value: (From the app)

$$p \approx 0.0020$$

Chapter 1 Review

You conduct a second sample of 120 adults, and this time, you find that 14 individuals have dyslexia.

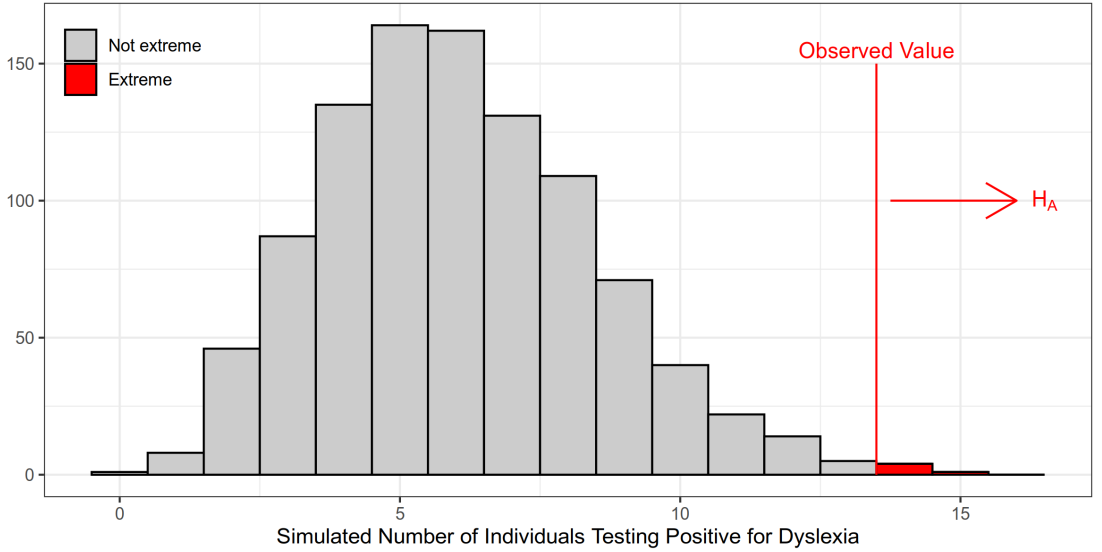
- Using the histogram below, draw lines indicating the cutoff(s) for extreme values, draw one or more arrows corresponding to the direction of H_A, and shade in the values considered "extreme".



Chapter 1 Review

You conduct a second sample of 120 adults, and this time, you find that 14 individuals have dyslexia.

- Using the histogram below, draw lines indicating the cutoff(s) for extreme values, draw one or more arrows corresponding to the direction of H_A, and shade in the values considered "extreme".



Note: This is a different simulated dataset than the one from the app.

(This is how I would test whether you understand how simulation works)

Chapter 1 Review

You conduct a second sample of 120 adults, and this time, you find that 14 individuals have dyslexia.

- How much evidence does the standardized statistic provide? What is your conclusion based on the standardized statistic?
- Based on your simulated p-value, what is your conclusion?

Chapter 1 Review

You conduct a second sample of 120 adults, and this time, you find that 14 individuals have dyslexia.

- How much evidence does the standardized statistic provide? What is your conclusion based on the standardized statistic?

The standardized statistic is **3.35**. This indicates that the sample value is in the tails of the distribution of values expected under H_0, which means it is **not very likely to occur if H_0 is true**. It is more likely that H_A is true, so we **reject H_0 and conclude that there is evidence that more than 5% of the population has dyslexia.**

Chapter 1 Review

You conduct a second sample of 120 adults, and this time, you find that 14 individuals have dyslexia.

- Based on your simulated p-value, what is your conclusion?

Under the null hypothesis, sampling 120 individuals and finding that 14 or more have dyslexia occurs with $\mathbf{p \approx 0.002}$, which is very unlikely. Thus, having observed that 14/120 adults have dyslexia in or sample, **we have very strong evidence against H_0, and conclude that it is more likely that H_A is true, that is, that more than 5% of the population has dyslexia**

Chapter 1 Review - The Adventures of Edison

Your reward for getting this far is a puppy picture and puppy-focused problems!

Chapter 1 Review - The Adventures of Edison

Edison is a very talented Jack Russell Terrier pup who is competing to be in a demonstration at the Puppy Bowl. To qualify, Edison must catch more than 75% of the frisbees thrown for him (over the long run). In one practice, he catches 16 of 20 frisbees thrown.

- Parameter:
- Success:
- H_0:
- H_A:
- $\hat{p} =$
- Can use theory?

Chapter 1 Review - The Adventures of Edison

Edison is a very talented Jack Russell Terrier pup who is competing to be in a demonstration at the Puppy Bowl. To qualify, Edison must catch more than 75% of the frisbees thrown for him (over the long run). In one practice, he catches 16 of 20 frisbees thrown.

- Parameter: *The long run proportion of frisbees caught*
- Success: *Catching a frisbee*
- $H_0: \pi = 0.75$
- $H_A: \pi > 0.75$
- $\hat{p} = 16/20 = .8$
- Can use theory? *No.* $n\pi = 15 > 10$ *but* $n(1-\pi) = 5$*, so we do not have 10 successes and 10 failures.*

Chapter 1 Review - The Adventures of Edison

Edison has been watching squirrels and would like to know if, when being chased, squirrels turn right and left equally often, so that he can catch them more easily when he escapes. He monitors the squirrels that come into his yard for a week, and notes that in 30 observations, squirrels turned left 20 times and right 10 times.

- Parameter:
- Success:
- H_0:
- H_A:
- $\hat{p} =$
- Can use theory?

Chapter 1 Review - The Adventures of Edison

Edison has been watching squirrels and would like to know if, when being chased, squirrels turn right and left equally often, so that he can catch them more easily when he escapes. He monitors the squirrels that come into his yard for a week, and notes that in 30 observations, squirrels turned left 20 times and right 10 times.

- Parameter: *The long run proportion of times squirrels turn left*
- Success: *Turning left*
- $H_0: \pi = 0.5$
- $H_A: \pi \neq 0.5$
- $\hat{p} = 20/30 \approx .6667$
- Can use theory? *Yes. Under H_0 we would have 15 successes and 15 failures.*

Chapter 1 Review - The Adventures of Edison

Edison might also be called "destroyer of toys". His owner would like to know whether, in the long run, he destroys the squeakers in more than 65% of the toys she buys for him. Over the course of a year, she carefully records the fate of each toy, finding that of the 24 squeaky toys she bought, only 8 still make noise.

- Parameter:
- Success:
- H_0:
- H_A:
- $\hat{p} =$
- Can use theory?

Chapter 1 Review - The Adventures of Edison

Edison might also be called "destroyer of toys". His owner would like to know whether, in the long run, he destroys the squeakers in more than 65% of the toys she buys for him. Over the course of a year, she carefully records the fate of each toy, finding that of the 24 squeaky toys she bought, only 8 still make noise.

- Parameter: *The long-run death rate of squeakers in Edison's toys*
- Success: *Edison destroys a squeaker in a toy*
- $H_0: \pi = 0.65$
- $H_A: \pi > 0.65$
- $\hat{p} = 16/24 = .6667$
- Can use theory? *No. Under H_0, we would expect 15.6 successes and 8.4 failures. We need 10 of each to use theory based techniques.*

Chapter 1 Review - The Adventures of Edison

Edison is perfecting his puppy dog eyes. He has practiced very hard for the past month, and would like to know if giving his people "the look" results in treats more than 50% of the time. Over the course of two weeks, Edison uses "the look" 56 times, and gets a total of 30 treats.

- Parameter:
- Success:
- H_0:
- H_A:
- $\hat{p} =$
- Can use theory?

Chapter 1 Review - The Adventures of Edison

Edison is perfecting his puppy dog eyes. He has practiced very hard for the past month, and would like to know if giving his people "the look" results in treats more than 50% of the time. Over the course of two weeks, Edison uses "the look" 56 times, and gets a total of 30 treats.

- Parameter: *The long run rate of getting treats after using puppy dog eyes*
- Success: *Using puppy dog eyes and getting a treat*
- $H_0: \pi = 0.5$
- $H_A: \pi > 0.5$
- $\hat{p} = 30/56 \approx 0.5457$
- Can use theory? *Yes. Under H_0 we would expect 28 successes and 28 failures.*

Chapter 1 Review - The Adventures of Edison

Edison has staring contests with the cat next door, who he has named Satan. He would like to know whether he has an even chance of winning the staring contests with Satan, over the long run. As a statistically inclined dog, he collects data over the course of two months, finding that of 184 staring contests, he won 80 times.

- Parameter:
- Success:
- H_0:
- H_A:
- $\hat{p} =$
- Can use theory?

Chapter 1 Review - The Adventures of Edison

Edison has staring contests with the cat next door, who he has named Satan. He would like to know whether he has an even chance of winning the staring contests with Satan, over the long run. As a statistically inclined dog, he collects data over the course of two months, finding that of 184 staring contests, he won 80 times.

- Parameter: *The long-run proportion of times Edison wins a staring contest with Satan*
- Success: *Winning one staring contest*
- $H_0: \pi = 0.5$
- $H_A: \pi \neq 0.5$
- $\hat{p} = 80/184 \approx .4348$
- Can use theory? *Yes, we would expect 92 successes and 92 failures under H_0, so we have at least 10 of each and can use theory based techniques*

Chapter 1 Review