Inference Review

Hypothesis Testing

Dr. Maria Tackett

08.28.19

Click for PDF of slides

Announcements

- Get to Know You survey - due Monday at 11:59p
- Labs start tomorrow. Make sure you...
  - Are a member of the `sta210-fa19` organization on GitHub
  - Can access the STA 210 course workspace on RStudio Cloud
- All regular office hours start Monday. See the course homepage for the office hour schedule.
- Duke Libraries Rfun - Intro to R Workshop: Data Transformations, Data Structures, and the Tidyverse
  - September 12 1p - 3p
  - To register: https://duke.libcal.com/event/5497129

Any questions from last class?

Today's Agenda

- Review hypothesis testing

Sesame Street

- *Sesame Street* is a television series designed to teach children ages 3-5 skills such as reading and math.
- The show originally had a particular focus on reaching economically disadvantaged children. In the early 1970s, the Educational Testing Service (ETS) conducted a study to determine the show's effectiveness in helping this group of children develop the skills needed to be successful in school.

Sesame Street

- A study was conducted to test whether the show was effective in helping children improve their reading and math skills. The 240 children who participated in the study were split into two groups:
  - **Group 1**: Those who were encouraged to watch the show (assumed watched regularly)
  - **Group 2**: Those who didn't get encouragement to watch the show
- Each child was given a test before and after the study to measure their knowledge of basic math, reading, etc.
- We will focus on the change in reading (identifying letters) scores

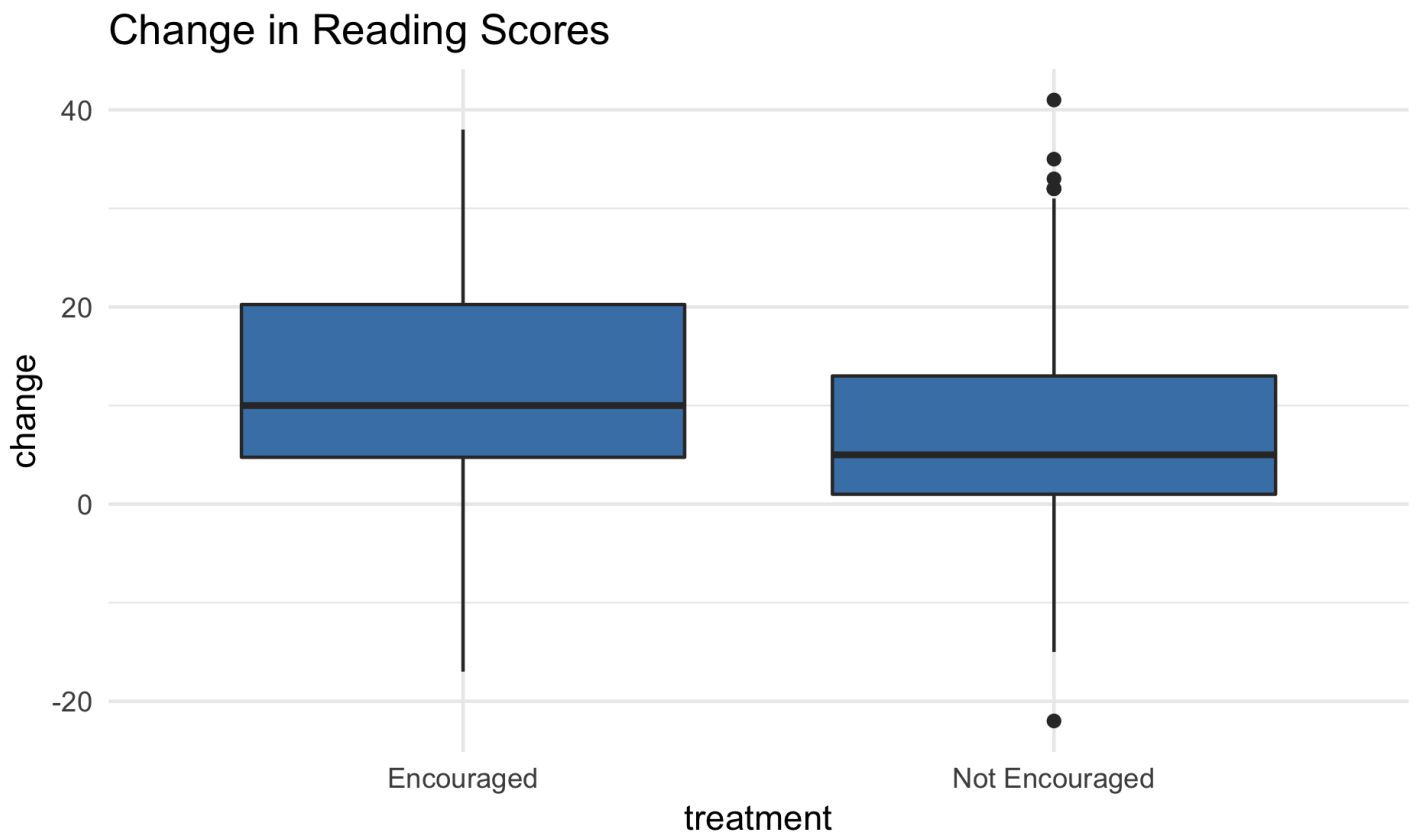
Sesame Street Data - Full Description Original Study: *Ann Bogatz, Gerry & Ball, Samuel. (1971). The Second Year of Sesame Street: A Continuing Evaluation. Volume 1. vols. 1 & 2.*

Let's look at the data

```
sesame_street %>%
  slice(1:10)
```

```
##          treatment prelet postlet change
## 1      Encouraged     23      30      7
## 2      Encouraged     26      37     11
## 3  Not Encouraged     14      46     32
## 4  Not Encouraged     11      14      3
## 5  Not Encouraged     47      63     16
## 6  Not Encouraged     26      36     10
## 7  Not Encouraged     12      45     33
## 8      Encouraged     48      47     -1
## 9      Encouraged     44      50      6
## 10     Encouraged     38      52     14
```

```
ggplot(data = sesame_street, mapping = aes(y = change, x = treatme
  geom_boxplot(fill = "steelblue") +
  labs(title = "Change in Reading Scores") +
  theme_minimal()
```



```
sesame_street %>%
  group_by(treatment) %>%
  summarise(n = n(), mean = mean(change), sd = sd(change))
```

```
## # A tibble: 2 x 4
##   treatment           n  mean    sd
##   <chr>           <int> <dbl> <dbl>
## 1 Encouraged        152 12.5   10.7
## 2 Not Encouraged     88  7.88  11.4
```

Based on this, do you think there is enough evidence to conclude that *Sesame Street* is effective in helping children learn reading skills? Why or why not?

What is statistical inference?

- **Statistical inference** is the process of using sample data to make conclusions about the underlying population from which the sample was taken
- Types of inference: testing and estimation
  - **Confidence Intervals:** Estimate the parameter of interest
  - **Hypothesis Tests:** Test a specified claim or hypothesis
- Today, we will focus on hypothesis testing

Hypothesis Tests

- Question we want to answer:

Are the data consistent or inconsistent with the specified hypothesis?

- To answer that question, we will determine

Given the collected data, is there evidence against a specified hypothesis about the corresponding parameter?

Is the averagee improvement in reading scores significantly greater for children who regularly watch *Sesame Street* than the change for those who don't?

Hypotheses

- Null hypothesis, H_0: This is the baseline hypothesis, i.e. the "there is nothing going on" hypothesis.
    - There is no difference in the reading development between children who watch *Sesame Street* regularly and those who don't.

- Alternative hypothesis, H_a: This is typically what you want to show, i.e. the "there is something going on" hypothesis
    - Children who regularly watch *Sesame Street* have significantly greater improvement in reading scores compared to those who don't

Outline of a Hypothesis Test

- Identify the parameter of interest.
- Identify a null hypothesis, H_0, that represents the baseline
- Set an alternative hypothesis, H_a, that represents the research question, i.e. what you're testing
- Conduct a hypothesis test under the assumption that the null hypothesis is true and calculate a p-value
    - The p-value is the probability of getting the observed outcome or a more extreme outcome given the null hypothesis is true

Outline of a Hypothesis Test

- Assess the p-value. A small p-value means...

  a. The assumed (null) hypothesis is incorrect

  b. The assumed (null) hypothesis is correct and a rare event has occured

- State a conclusion about the hypothesis based on the assessment of the p-value

  - Since event (b) is by definition rare, we will conclude a "small" p-value indicates that there is sufficient evidence to claim that the assumed hypothesis is false.

    - In other words, the data are not consistent with the assumed hypothesis

- When the p-value is "not small", we will conclude that there is not sufficient evidence to claim the assumed hypothesis is false.

Sesame Street Example

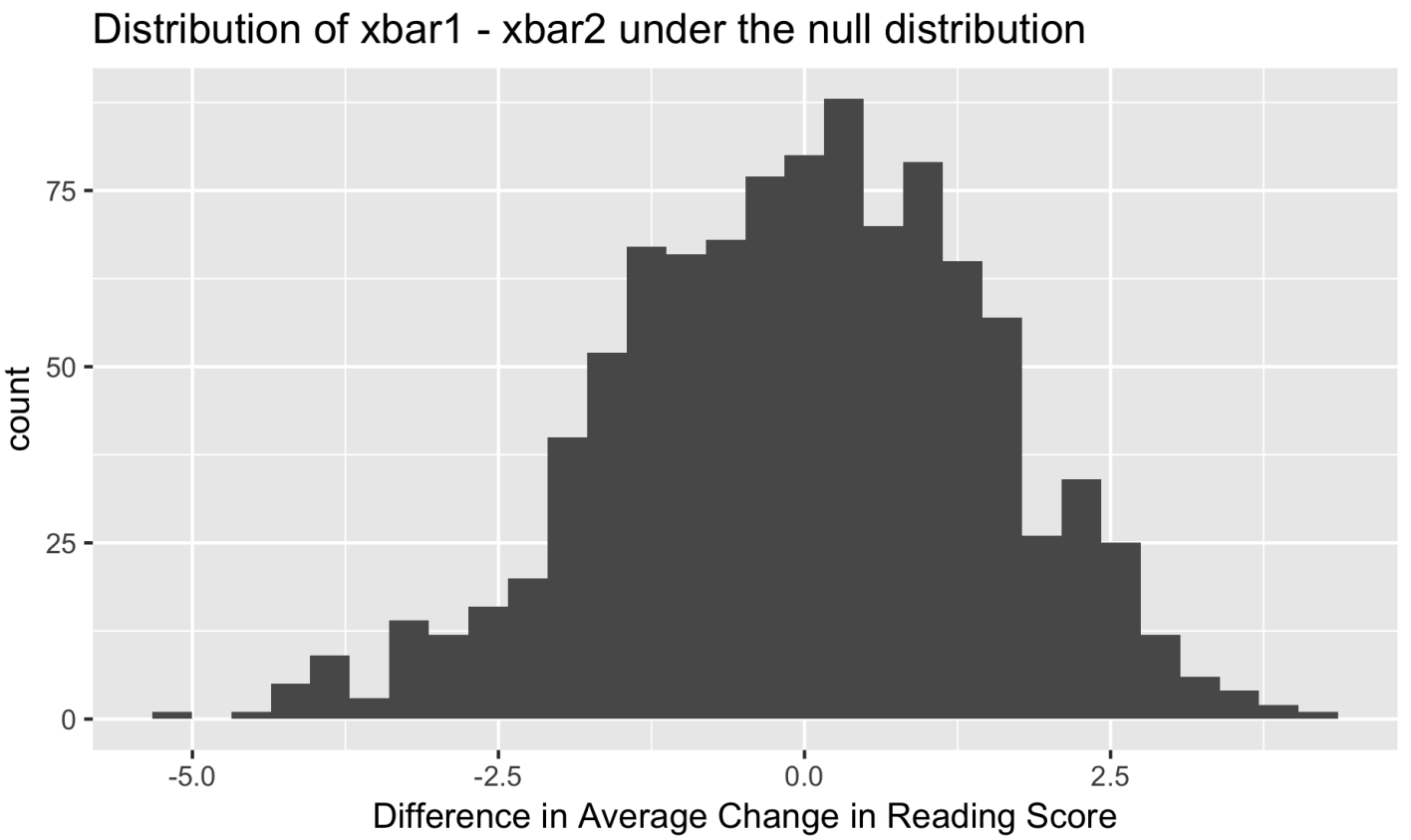
- **Question:** Is the averagee improvement in reading scores significantly greater for children who regularly watch *Sesame Street* than the change for those who don't?
- **Baseline:** There is no difference between the two groups
- **Claim:** The average improvement in reading scores for children who watch *Sesame Street* is greater than the average improvement of children who don't. We will use encouragement as a proxy for watching the show.
- **Hypotheses:**

$$H_0 : \mu_e - \mu_{ne} = 0$$
$$H_a : \mu_e - \mu_{ne} > 0$$

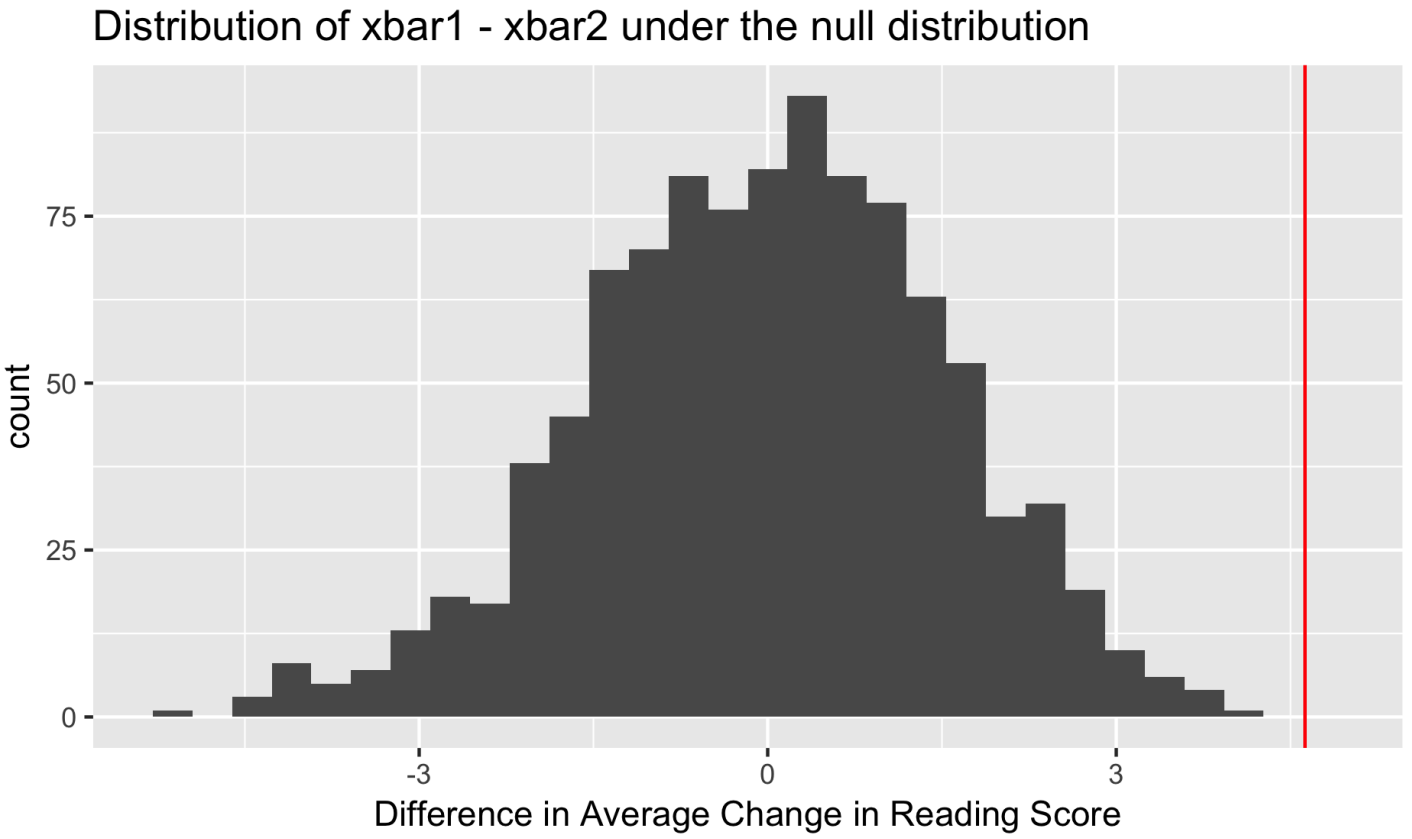
Distribution of sample statistic under H_0

- We conduct hypothesis testing under the assumption that the null hypothesis is true, $H_0 : \mu_e - \mu_{ne} = 0$
- We use sample statistics $\bar{x}_e$ and $\bar{x}_{ne}$ to help us understand the parameters.
- If the null hypothesis is true, we could randomly relabel the observations as Group 1 and Group 2 and recalculate the sample statistic $\bar{x}_e - \bar{x}_{ne}$
- Let's do this relabeling about 1000 times and see what the distribution of $\bar{x}_e - \bar{x}_{ne}$ looks like

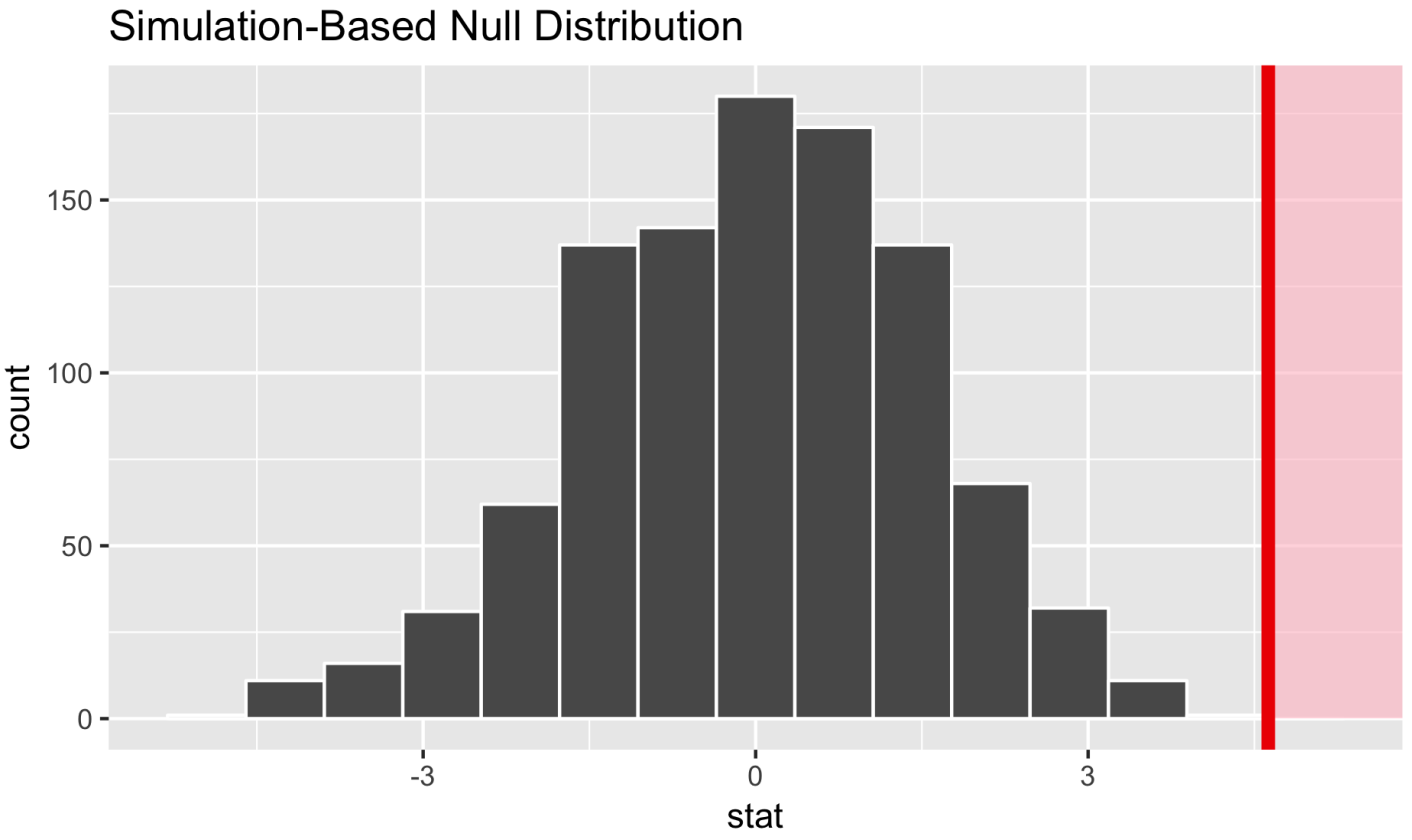
Simulated distribution of sample statistic under H_0



In our actual data, we observed a difference of **12.5 - 7.875 = 4.625**



What proportion of the 1000 observations yielded values of $\bar{x}_e - \bar{x}_{ne}$ as or more extreme as the value we observed in the data?



What is your conclusion about the effectiveness of *Sesame Street* in helping students improve their reading skills?

Inference Using the Central Limit Theorem

Sampling Distributions

- Sampling distribution: the distribution of sample statistics of random samples of size n taken with replacement from a population
- If we take repeated samples from a population, each sample will yield a slightly different value of the sample statistic
- We can measure the variability in these sample statistics by the standard error

In practice...

We can't directly know what the sampling distributions looks like, because we only draw a single sample.

- The whole point of statistical inference is to deal with this issue: observe only one sample, try to make inference about the entire population
- One approach is to simulate the distributions as we did earlier, but we will now rely on theoretical results.
- The Central Limit Theorem is a theoretical result that tells us what the sampling distribution should look like (for certain sample statistics)
- The Central Limit Theorem provides the foundation we need to conduct inference on parameters such as the mean, proportion, difference in means, and regression coefficients (coming soon!)

Central Limit Theorem

If certain conditions are met, the sampling distribution of the sample statistic will..

- be approximately Normal
- have a mean equal to the unknown population parameter
- have a standard error proportional to the inverse of the square root of the sample size.

Central Limit Theorem

One Sample:

- Single mean: $\bar{x} \sim N\left(mean = \mu, sd = \frac{\sigma}{\sqrt{n}}\right)$
- Single proportion: $\hat{p} \sim N\left(mean = p, sd = \sqrt{\frac{p(1-p)}{n}}\right)$

Two Sample:

- Difference between two means:
  $(\bar{x}_1 - \bar{x}_2) \sim N\left(mean = (\mu_1 - \mu_2), sd = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$
- Difference between two proportions:
  $(\hat{p}_1 - \hat{p}_2) \sim N\left(mean = (p_1 - p_2), sd = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$

Conditions required for the CLT

- **Independence:** The sampled observations must be independent. This is difficult to check, but the following are useful guidelines:
  - the sample must be random
  - if sampling without replacement, sample size must be less than 10% of the population size
- **Sample size / distribution:**
  - numerical data: The more skewed the sample (and hence the population) distribution, the larger samples we need. Usually n > 30 is considered a large enough sample for population distributions that are not extremely skewed.
  - categorical data: At least 10 successes and 10 failures.
- If comparing two populations, the groups must be independent of each other, and all conditions should be checked for both groups.

Standard Error

The **standard error** is the *standard deviation* of the *sampling distribution*, calculated using sample statistics (since we don't know the population parameters like σ or p).

- Single mean: $SE = \frac{s}{\sqrt{n}}$
- Difference between two means: $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
- Single proportion: $SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
- Difference between two proportions: $SE = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

How are standard error and sample size associated? What does that say about how the spread of the sampling distribution changes as n increases?

Hypothesis Tests Using the CLT

- State the null and alternative hypotheses
- Calculate the standard error of the sample statistic of interest (sample mean, sample proportion, difference between sample means, etc.)
- Calculate the **test statistic**, the number of standard errors the observed value is from the hypothesized null value. The appropriate test statistic is...

$$\frac{\text{observed value} - \text{hypothesized value}}{SE}$$

  - *z* for proportions
  - *t* for means, along with appropriate degrees of freedom

Hypothesis Tests Using the CLT

- Use the test statistic to calculate the p-value, the probability of the observed outcome or a more extreme outcome given that the null hypothesis is true
  - Standard Normal distribution for proportions
  - *t* distribution with appropriate degrees of freedom for means (and eventually regression coefficients)

Sesame Street Example using the CLT

- Sample Statistic, $\bar{x}_e - \bar{x}_{ne}$:

```
## [1] 4.625
```

- Standard Error, $\sqrt{s_e^2/n_e + s_{ne}^2/n_{ne}}$:

```
## [1] 1.49096
```

- Test Statistic, $\frac{(\bar{x}_e - \bar{x}_{ne} - 0)}{\sqrt{s_e^2/n_e + s_{ne}^2/n_{ne}}}$:

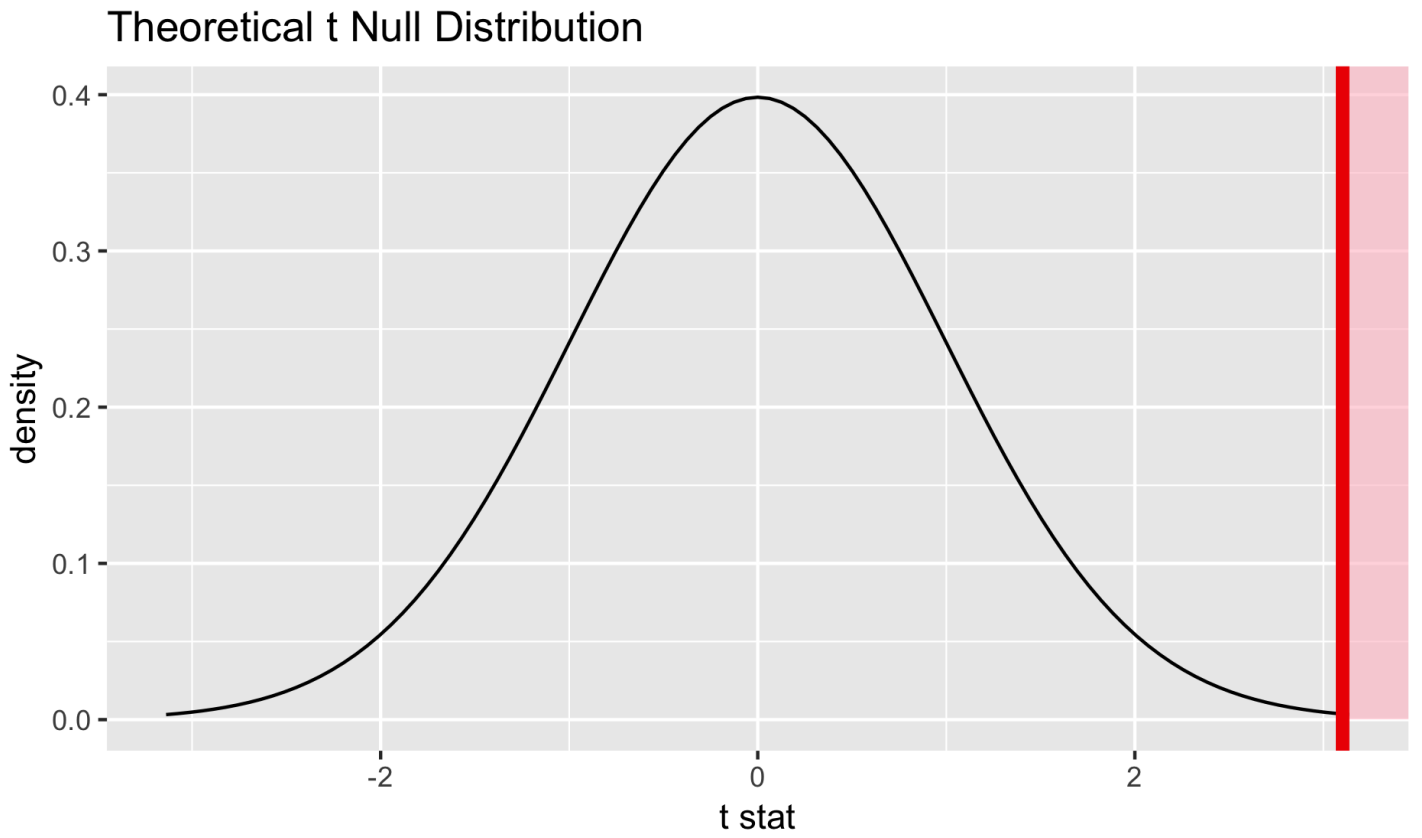
```
## [1] 3.102028
```

- P-value:

```
## [1] 0.001121931
```

Theoretical distribution under H_0

t distribution with approximately 173.59 degrees of freedom.



Understanding the Hypothesis Test

Calculating the p-value

- **p-value:** probability of getting a test statistic as extreme or more extreme than the calculated test statistic, assuming the null hypothesis is true
- When the alternative has a $>$, the p-value is calculated using the area to the right of the test statistic
- When the alternative has a $<$, the p-value is calculated using the area to the left of the test statistic
- When the alternative has $\neq$, the p-value is calculated as the area to the left of $-|\text{test statistic}|$ and to the right of $|\text{test statistic}|$

Interpreting the p-value

What the p-value is NOT:

- It is *not* the probability the null hypothesis is true
  - The null hypothesis is either true or not true
- (1 - *p-value*) is *not* the probability that the alternative hypothesis is true
  - The alternative hypothesis is either true or not true

Correct interpretation of the p-value:

The probability of getting a test statistic as extreme or more extreme than the calculated test statistic, *assuming the null hypothesis is true.*

Interpreting the p-value

Magnitude of p-value	Interpretation
p-value < 0.01	strong evidence against H_0
0.01 < p-value < 0.05	moderate evidence against H_0
0.05 < p-value < 0.1	weak evidence against H_0
p-value > 0.1	effectively no evidence against H_0

Note: These are general guidelines. The strength of evidence depends on the context of the problem.

Statistical Significance

- A threshold can be used to decide whether or not to reject H_0.
- This threshold is called the **significance level** and is usually denoted by α
- When H_0 is rejected, we use the term **statistically significant** to describe the outcome of the test.
- *Example*: When $\alpha = 0.05$, results are statistically significance when the p-value is < 0.05

Statistical Significance

- Do not rely strictly on the significance level to make a conclusion!
- Suppose the significance level is 0.05
  - If the p-value is 0.05001, we do not reject H_0
  - If the p-value is 0.04999, we do reject H_0
- 0.05001 and 0.04999 are practically the same, yet they led to different conclusions.
- Always state the p-value when reporting results and assess it's magnitude in the context of your problem.

Results that Aren't Statistically Significant

- An outcome of failing to reject H_0 is *not* a failed study/experiment
- Obtaining an outcome of "no significant effect" or "no significant difference" is still valid
- It is often just as important to learn that the H_0 can't be refuted

Statistical Inference

- We concluded that there is a statistically significant difference in the average reading improvement between children who watched *Sesame Street* regularly compared to those who didn't
- From our sample data, the estimated difference is $\bar{x}_e - \bar{x}_{ne}$ = 12.5 - 7.875 = 4.625
- Though this may be a good estimate, we saw earlier that there is variability in statistics calculated from sample data.
- Next class, we will discuss how to account for that variability in our estimate of the parameter.

Before Next Class

- Fill out the **Getting To Know You Survey on Sakai** - due 9/2 at 11:59p
- Accept invite to join `sta210-fa19` organization on GitHub
- **New to R or need a refresher?**
  - Duke Libraries Rfun - Intro to R Workshop: Data Transformations, Data Structures, and the Tidyverse
    - September 12 1p - 3p
    - To register: https://duke.libcal.com/event/5497129
  - *Work with Data* primer on RStudio Cloud: https://rstudio.cloud/learn/primers/2
  - "Data Visualization" in *R for Data Science*: https://r4ds.had.co.nz/data-visualisation.html
- **More on statistical inference**
  - OpenIntro Statistics Chapter 5: Inference for numerical data