Inference Review

Hypothesis Testing

Prof. Maria Tackett

01.15.20

[Click for PDF of slides]

Announcements

- Complete surveys and consent form by Wed at 11:59p
- Reading for next Wednesday
- Labs start tomorrow!
- No class Monday - Martin Luther King, Jr. Holiday
- Find more info about statistics related events on Sakai

STA 210

Today's Agenda

- Calculating & interpreting hypothesis tests
- Drawing conclusions using hypothesis tests and confidence intervals

Sesame Street

- *Sesame Street* is a television series designed to teach children ages 3-5 basic education skills such as reading (e.g. the alphabet) and math (e.g. counting)
- Today we are going to analyze data from an [study conducted by the Educational Testing Service] in the early 1970s to test the effectiveness of the program.

Sesame Street study

- Children from 6 locations around the United States (including Durham!) participated in the 26-week study. The children were split into two groups (`treatment`):
  - **Group 1**: Those who were encouraged to watch the show (assume watched regularly)
  - **Group 2**: Those who didn't get encouragement to watch the show (assume didn't watch regularly)
- 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 (`change`)

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.*

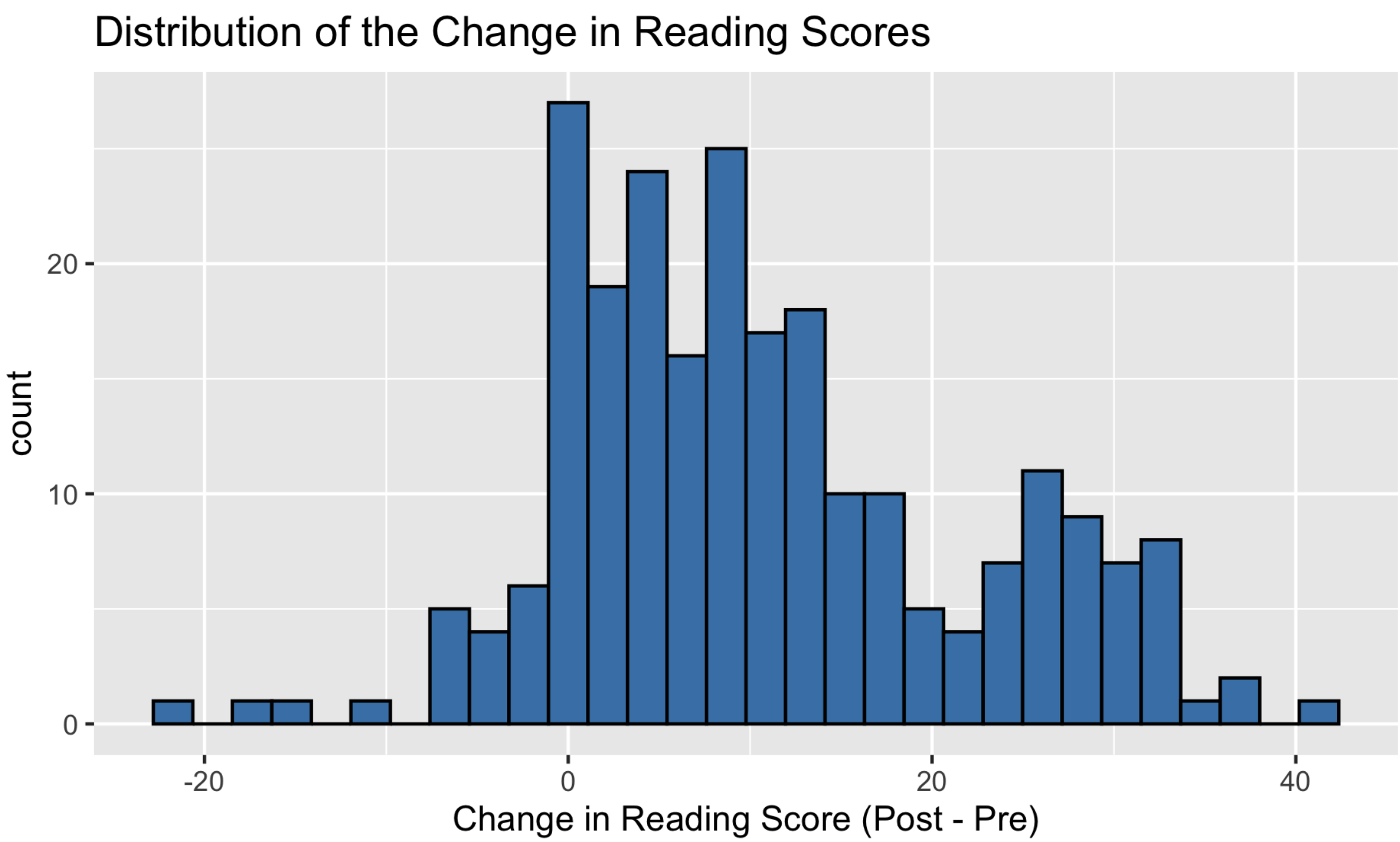
Let's look at the data

`sesame_street.csv` is available in the datasets repo on GitHub.

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

```
## # A tibble: 10 x 4
##    treatment      prelet postlet change
##    <chr>           <dbl>   <dbl>  <dbl>
##  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
```

Exploratory Data Analysis - Univariate



STA 210

Exploratory Data Analysis - Univariate

- Calculate summary statistics for change

```
sesame_street %>%
  summarise(n = n(), min = min(change), median = median(change), n
            IQR = IQR(change),
            mean = mean(change), std_dev = sd(change))
```

```
## # A tibble: 1 x 7
##       n   min median   max   IQR  mean std_dev
##   <int> <dbl>  <dbl> <dbl> <dbl> <dbl>   <dbl>
## 1   240   -22      9    41    15  10.8    11.2
```

95% CI for mean change in reading score

The 95% confidence interval for the mean change in reading score is

$$[9.384, 12.224]$$

- Interpret the interval at http://bit.ly/sta210-sp20-CI-2
- Use **NetId@duke.edu** for your email address.
- You are welcome (and encouraged!) to discuss these questions with 1 - 2 people around you, but **each person** must submit a response.

03:00

STA 210

Confidence Interval Recap

Hypothesis Tests

Question we want to answer

- Let's focus on the children who were encouraged to watch *Sesame Street*
- In general, those children watched the show regularly, so let's see if the show impacted their reading skills

> The mean change in reading scores after 26 weeks for children ages 3 - 5 is 11.
>
> Is there evidence that mean change in reading score for children encouraged to watch *Sesame Street* is "significantly" different from the mean change in reading score for all children?

Let's look at the data


```
## # A tibble: 1 x 3
##       n  mean std_dev
##   <int> <dbl>   <dbl>
## 1   152  12.5    10.7
```

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

STA 210

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 occurred

- 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.

STA 210

Identify parameter & hypotheses

- **Null hypothesis, H_0**: This is the baseline hypothesis, i.e. the "there is nothing going on" hypothesis.
    - The mean change in reading score for children encouraged to watch the show is 11 (same as the mean for all children)

- **Alternative hypothesis, H_a**: This is typically what you want to show, i.e. the "there is something going on" hypothesis
    - The mean change in reading score for children encouraged to watch the show not 11 (different from the mean for all children)

$$H_0 : \mu = 11$$
$$H_a : \mu \neq 11$$

Distribution $\bar{x}$ under H_0

- We want to draw conclusions about μ, so we'll use our best guess $\bar{x}$
- Recall from the Central Limit Theorem, when certain conditions are met (they are!), we know that

$$\bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

- We conduct a hypothesis test under the assumption that H_0 is true, so for this test

$$\bar{x} \sim N\left(11, \frac{\sigma}{\sqrt{n}}\right)$$

STA 210

Distribution $\bar{x}$ under H_0

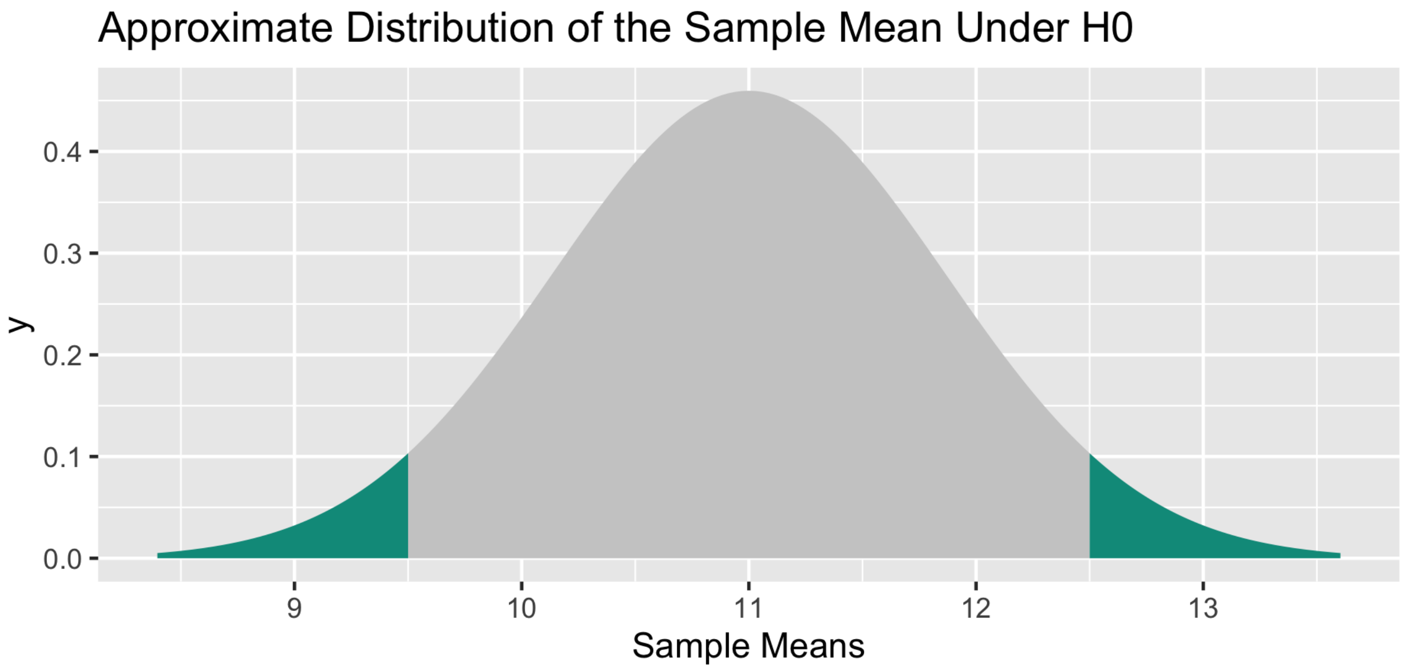
- We don't know σ, so we can use the standard error $s/\sqrt{n}$ to approximate $\sigma/\sqrt{n}$.
- Thus, putting it all together, we know

$$\bar{x} \approx N\left(11, \frac{10.7}{\sqrt{152}}\right)$$

Given $\bar{x} \approx N\left(11, \frac{10.7}{\sqrt{152}}\right)$, what is the probability of observing a mean change in score at least 1.5 points away from the center (11) in a random sample of 152 children ages 3 - 5?

STA 210

Visualize



The shaded area represents the (approximate) probability of obtaining a sample mean at least as far away from the center as the one we observed given the true mean change is 11.

Test Statistic

- Let's quantify how "unusual" our observed sample mean is given $H_0 : \mu = 11$ is true

- We'll begin by calculating how "far away" the observed mean is from the center of the distribution under H_0

- The **test statistic** is the number of standard errors the observed value is from the hypothesized value. The general form of the test statistic is

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

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{12.5 - 11}{\frac{10.7}{\sqrt{152}}} \approx 1.728$$

where the test statistic follows the t distribution with $n - 1$ df

STA 210

Motivating the p-value

- We got a test statistic of 1.728. In other words…

Given $\bar{x} \approx N\left(11, \frac{10.7}{\sqrt{152}}\right)$, what is the probability of observing a mean change in score at least 1.5 points away from the center (11) in a random sample of 152 children ages 3 - 5?

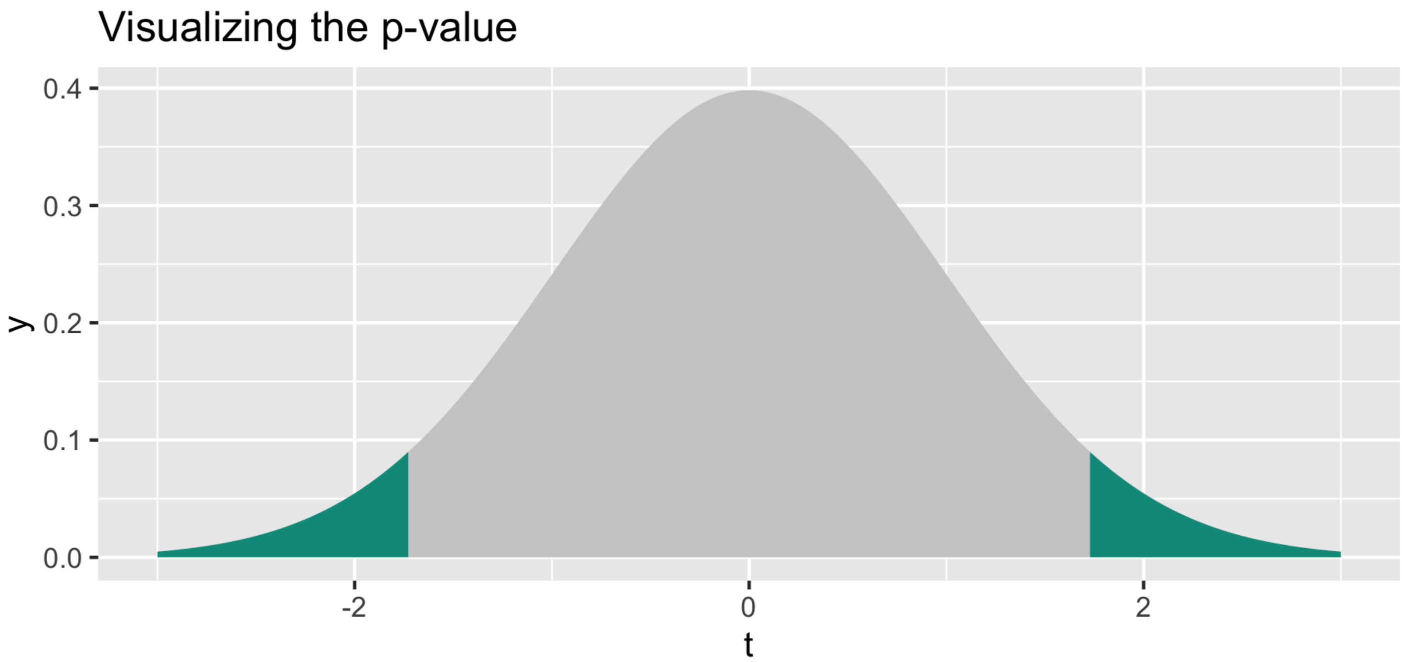
$$\Downarrow$$

Given the t distribution with 151 degrees of freedom, what is the probability of observing a test statistic with magnitude 1.728 or larger?

STA 210

p-value

Given the t distribution with 151 degrees of freedom, what is the probability of observing a test statistic with magnitude 1.728 or larger?




```r
(p_value <- 2 * pt(-1.728, 151))
```

```
## [1] 0.08603259
```

General guide for 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.

Drawing the conclusion: Part 1

- A threshold can be used to decide whether or not to reject H_0 in favor of the alternative H_a
- This threshold is called the significance level and is usually denoted by α
- If the p-value is less than α, then we conclude there is sufficient evidence against $H-$ and we **reject the null hypothesis**
- Otherwise, we conclude that there isn't sufficient evidence against H_0 and **fail to reject the null hypothesis**

STA 210

Don't just rely on p-values

- Do not rely strictly on the p-value and significance level to make a conclusion!
- Suppose the significance level is 0.05
  - If the p-value is 0.05001, we fail to reject H_0
  - If the p-value is 0.04999, we reject H_0
- 0.05001 and 0.04999 are practically the same, yet they led to different conclusions.

> Use confidence intervals and other statistical summaries to provide more context about the results.

t-test for *Sesame Street* data

```r
enc <- sesame_street %>%
  filter(treatment == "Encouraged")

t.test(enc$change, mu = 11, conf.level = 0.9,
       direction = "two.sided")
```

```
##
##      One Sample t-test
##
## data:  enc$change
## t = 1.7226, df = 151, p-value = 0.08701
## alternative hypothesis: true mean is not equal to 11
## 90 percent confidence interval:
##  11.05883 13.94117
## sample estimates:
## mean of x
##      12.5
```

In-class exercise

- Answer the questions: http://bit.ly/sta210-sp20-ht
- Use **NetId@duke.edu** for your email address.
- You are welcome (and encouraged!) to discuss these questions with 1 - 2 people around you, but **each person** must submit a response.

04:00

Conclusion

p-value: 0.087

90% confidence interval: [11.059, 13.941]

- Using a significance level of 0.1, what is your conclusion from the test?
- Suppose you are advising a group of educators about whether they should spend additional time and money to encourage children to watch *Sesame Street*. Based on these results, would you advise the educators to spend the resources? Why or why not?

STA 210

Inference for difference in means $\mu_1 - \mu_2$

By the Central Limit Throem, when the conditions are met,

$$(\bar{x}_1 - \bar{x}_2) \sim N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

- We don't know σ_1 and σ_2 in practice, so we use the **standard error** in all calculations.

$$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Inference for difference in means $\mu_1 - \mu_2$

Confidence Interval to estimate $\mu_1 - \mu_2$

$$(\bar{x}_1 - \bar{x}_2) \pm t^*_{df} \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

t^* follows a t distribution with degrees of freedom computed in R.

STA 210

Inference for difference in means $\mu_1 - \mu_2$

Hypothesis Test: Is there a difference in the means between Group 1 and Group 2?

- Null hypothesis: $H_0 : \mu_1 - \mu_2 = 0$
- Test statistic:

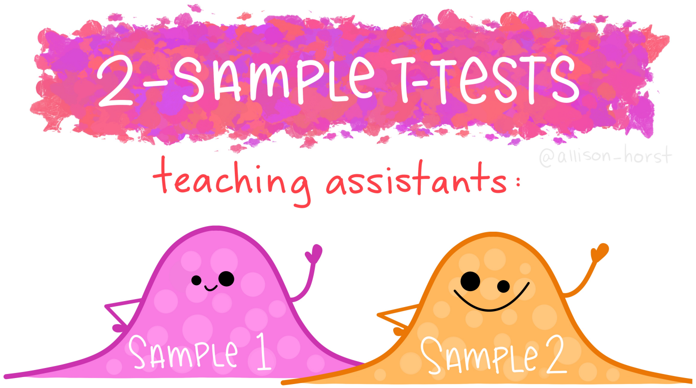
$$\frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- p-value: Calculated using the t distribution with degrees of freedom computed in R.

STA 210

Additional Resources

- Discussion in the scientific community about p-values: ["Scientists rise up against statistical significance" in *Nature*]
- Fun review of two-sample tests by @allison_horst: https://twitter.com/allison_horst/status/1216411185240690688



STA 210

By Thursday at noon

- Make sure you are a member of the course organization on GitHub
- Make sure you have access to RStudio
- If you are using RStudio on your local machine, make sure you have git configured and you can knit a PDF (need a Latex editor installed)