STAT 216 Coursepack

Spring 2025
Montana State University

Melinda Yager
Jade Schmidt
Stacey Hancock

This resource was developed by Melinda Yager, Jade Schmidt, and Stacey Hancock in 2021 to accompany the online textbook: Hancock, S., Carnegie, N., Meyer, E., Schmidt, J., and Yager, M. (2021). *Montana State Introductory Statistics with R*. Montana State University. https://mtstateintrostats.github.io/IntroStatTextbook/.

This resource is released under a Creative Commons BY-NC-SA 4.0 license unless otherwise noted.

Contents

Preface

This coursepack accompanies the textbook for STAT 216: Montana State Introductory Statistics with R, which can be found at https://mtstateintrostats.github.io/IntroStatTextbook/. The syllabus for the course (including the course calendar), data sets, and links to D2L Brightspace, Gradescope, and the MSU RStudio server can be found on the course webpage: https://math.montana.edu/courses/s216/. Other notes and review materials are linked in D2L.

Each of the activities in this workbook is designed to target specific learning outcomes of the course, giving you practice with important statistical concepts in a group setting with instructor guidance. In addition to the in-class activities for the course, video notes are provided to aid in taking notes while you complete the required videos. Bring this workbook with you to class each class period, and take notes in the workbook as you would your own notes. A well-written completed workbook will provide an optimal study guide for exams!

All activities and labs in this coursepack will be completed during class time. Parts of each lab will be turned in on Gradescope. To aid in your understanding, read through the introduction for each activity before attending class each day.

STAT 216 is a 3-credit in-person course. In our experience, it takes six to nine hours per week outside of class to achieve a good grade in this class. By "good" we mean at least a C because a grade of D or below does not count toward fulfilling degree requirements. Many of you set your goals higher than just getting a C, and we fully support that. You need roughly nine hours per week to review past activities, read feedback on previous assignments, complete current assignments, and prepare for the next day's class. A typical week in the life of a STAT 216 student looks like:

- *Prior to class meeting*:
  - Read assigned sections of the textbook, using the provided reading guides to take notes on the material.
  - Watch the provided videos, taking notes in the coursepack.
  - Read through the introduction to the day's in-class activity.
  - Read through the week's homework assignment and note any questions you may have on the content.
- *During class meeting*:
  - Work through the guided activity, in-class activity or weekly lab with your classmates and instructor, taking detailed notes on your answers to each question in the activity.
- *After class meeting*:
  - Complete any parts of the activity you did not complete in class.
  - Review the activity solutions in the Math and Stat Center, and take notes on key points.
  - Complete any remaining assigned readings for the week.
  - Complete the week's homework assignment.

MODULE 1

Inference for a Quantitative Response with Independent Samples

1.1 Vocabulary Review and Key Topics

Review the Golden Ticket posted in the resources at the end of the coursepack for a summary of a categorical explanatory variable and a quantitative response variable for independent samples. Module 12 will cover both simulation and theory-based methods of inference.

Types of plot for independent variables

- **Side-by-side boxplots**: plots a boxplot of the five number summary for each categorical level
- R code to create side-by-side boxplots:

```
object %>%  # Data set piped into...
  ggplot(aes(y = response, x = explanatory))+  # Identify variables
  geom_boxplot()+  # Tell it to make a box plot
  labs(title = "Don't forget to include a title",  # Title: should include the type of plot,
       # observational units, variables
       x = "x-axis label",    # x-axis label
       y = "y-axis label")  # y-axis label
```

- **Stacked histogram**: plots one histogram for each level of the categorical variable
- **Stacked dotplots**: plots one dotplot for each level of the categorical variable
- Four characteristics to compare boxplots
  - Shape (symmetric or skewed)
  - Center
  - Spread
  - Outliers?

Summary measure

- **Difference in mean**: measures the difference in mean values between the two categorical groups
- Parameter notation for difference in means: $\mu_1 - \mu_2$, where 1 represents the 1st group of the explanatory variable and 2 represents the 2nd group
- Sample notation for difference in means: $\bar{x}_1 - \bar{x}_2$
- R code to find the summary statistics
  - Note: review the interpretations of the other summary measures from Module 6

```
object %>%
  reframe(favstats(response~explantory))
```

Hypothesis Testing

Hypotheses:

$$H_0 : \mu_1 - \mu_2 = 0 \text{ or } H_0 : \mu_1 = \mu_2$$

$$H_A : \mu_1 - \mu_2 \left\{ \begin{array}{c} < \\ \neq \\ < \end{array} \right\} 0 \text{ or } H_A : \mu_1 \left\{ \begin{array}{c} < \\ \neq \\ < \end{array} \right\} \mu_2$$

Simulation Hypothesis Testing

- R code for simulation methods to find the p-value using the `two_mean_test` function in the `catstats` package.

```
two_mean_test(response~explanatory, #Enter the names of the variables
               data = object,  # Enter the name of the dataset
              first_in_subtraction = "xx", # First outcome in order of subtraction
               number_repetitions = 10000,  # Number of simulations
               as_extreme_as = -xx,  # Observed statistic
               direction = "xx")  # Direction of alternative: "greater", "less", or "two-sided"
```

Simulation Confidence Interval

- R code to find the simulation confidence interval using the `twomean_bootstrap_CI` function from the `catstats` package.

```
two_mean_bootstrap_CI(response ~ explanatory, #Enter the name of the variables
                      data = object,  # Enter the name of the data set
                      first_in_subtraction = "xx", # First value in order of subtraction
                      number_repetitions = 10000,  # Number of simulations
                      confidence_level = xx)
```

- Review how to interpret the confidence interval for two groups from Module 8

Theory-based methods

- **Conditions for the sampling distribution of $\bar{x}_1 - \bar{x}_2$ to follow an approximate normal distribution**:
  - **Independence**: The sample's observations are independent, e.g., are from a simple random sample and there is independence between groups. (*Remember*: This also must be true to use simulation methods!)
  - **Normality Condition**: Either the sample observations come from a normally distributed population or we have a large enough sample size. When we have two samples, we need to check this condition for each group! To check this condition, use the following rules of thumb (for both n_1 and n_2):
    - $n < 30$: The distribution of the sample must be approximately normal with no outliers.
    - $30 \leq n < 100$: We can relax the condition a little; the distribution of the sample must have no extreme outliers or skewness.
    - $n \geq 100$: Can assume the sampling distribution of $\bar{x}$ is nearly normal, even if the underlying distribution of individual observations is not.
- **t-distribution**: a theoretical distribution that is symmetric with a given degrees of freedom smallest sample size minus 1 $(n - 1)$

$$t_{n-1}$$

- Calculation of standard error:

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

- Calculation of the standardized difference in sample mean:

$$t = \frac{\bar{x}_1 - \bar{x}_2 - 0}{SE(\bar{x}_1 - \bar{x}_2)}$$

- The p-value can be found by using the pt function.
  - Enter the value of the standardized statistic for xx
  - Enter the df smallest $(n-1)$ for yy
  - If a greater than alternative, change lower.tail = TRUE to FALSE.
  - If a two-sided test, multiply by 2.

```
pt(xx, df = yy, lower.tail=TRUE)
```

Theory-based methods to find the confidence interval

- Calculation of the confidence interval for a difference in sample means

$$\bar{x}_1 - \bar{x}_2 \pm t^* \times SE(\bar{x}_1 - \bar{x}_2)$$

- R code to find the multiplier for the confidence interval using theory-based methods.
  - qt will give you the multiplier using the t-distribution with smallest $n-1$ df (enter for yy)
  - Enter the percentile for the given confidence level

```
qt(percentile, df=yy, lower.tail=FALSE)
```

1.2 Video Notes: Inference for Independent Samples

Read Chapters 19 and 20 in the course textbook. Use the following videos to complete the video notes for Module 10.

1.2.1 Course Videos

- 19.1
- 19.2
- 19.3TheoryTests
- 19.4TheoryInterval

Single categorical, single quantitative variable with independent samples

- In this module, we will study inference for a ________________________________ explanatory variable and a ____________________________________ response variable where the two groups are ____.
- Independent groups: When the measurements in one sample are not related to the measurements in the other sample.
- Two random samples taken separately from two populations and the same response variable is recorded. Compare the average number of sick days off from work for people who had a flu shot and people who didn't.
- Participants are randomly assigned to one of two treatment conditions, and the same response variable is recorded.

Rather than analyzing the differences as a single mean we will calculate summary statistics on each sample.

Example: Fifty-one (51) college students volunteered to look at impacts on memorization, specifically if putting letters into recognizable patterns (like FBI, CIA, EDA, CDC, etc.) would increase the number letters memorized. (Miller 1956) The college students were randomly assigned to either a recognizable or non-recognizable letter group. After a period of study time, the number of letters memorized was collected on each study. Is there evidence that putting letters into recognizable letter groups improve memory?

- The summary measure for two independent groups is the ________________________________ in ____.

Notation for Independent Groups

- Population mean for group 1:
- Population mean for group 2:
- Sample mean for group 1:
- Sample mean for group 2:
- Sample difference in means:

- Population standard deviation for group 1:
- Population standard deviation for group 2:
- Sample standard deviation for group 1:
- Sample standard deviation for group 2:
- Sample size for group 1:
- Sample size for group 2:

Why should we treat this as two independent groups rather than paired data?

Hypothesis Testing

Conditions:

- Independence: the response for one observational unit will not influence the outcome for another observational unit

Null hypothesis assumes "no effect", "no difference", "nothing interesting happening", etc.

Always of form: "parameter" = null value

H_0 :

H_A :

- Research question determines the alternative hypothesis.

Write the null and alternative hypotheses for the letters study:

In notation:

H_0 :

H_A :

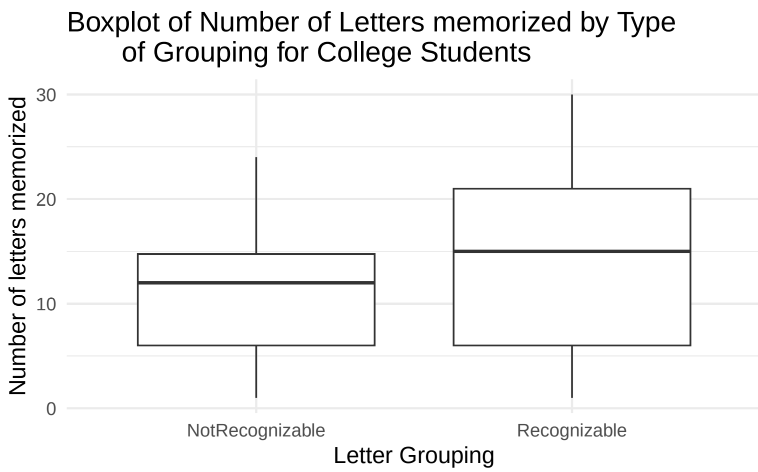
```
letters<-read.csv("data/letters.csv")
letters %>%
    reframe(favstats(Memorized~Grouped))
```

```
#>           Grouped min Q1 median    Q3 max     mean       sd  n missing
#> 1 NotRecognizable   1  6     12 14.75  24 11.15385 6.576883 26       0
#> 2    Recognizable   1  6     15 21.00  30 14.32000 8.518216 25       0
```

Summary statistic:

Interpret the summary statistic in context of the problem:

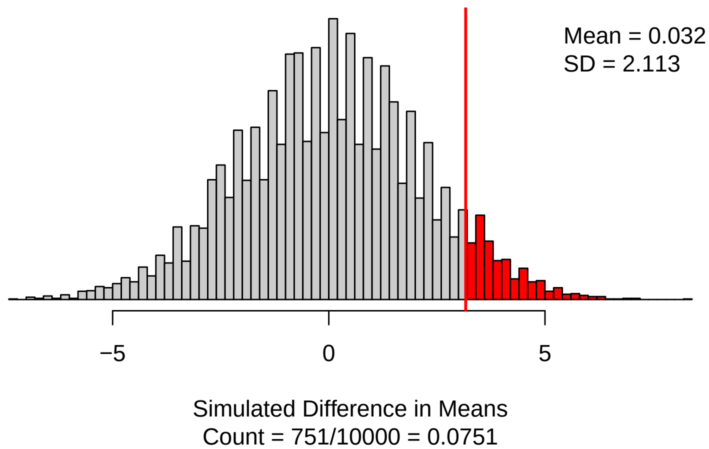
```
letters%>%
  ggplot(aes(y = Memorized, x = Grouped))  + #Enter the name of the explanatory and response variable
  geom_boxplot()+
  labs(title = "Boxplot of Number of Letters memorized by Type
       of Grouping for College Students", #Title your plot
       y = "Number of letters memorized", #y-axis label
       x = "Letter Grouping") #x-axis label
```



Simulation-based method

- Simulate many samples assuming $H_0 : \mu_1 = \mu_2$
  - Write the response variable values on cards
  - Mix the explanatory variable groups together
  - Shuffle cards into two explanatory variable groups to represent the sample size in each group (n_1 and n_2)
  - Calculate and plot the simulated difference in sample means from each simulation
  - Repeat 1000 times (simulations) to create the null distribution
  - Find the proportion of simulations at least as extreme as $\bar{x}_1 - \bar{x}_2$

```
set.seed(216)
two_mean_test(Memorized~Grouped, #Enter the names of the variables
              data = letters,  # Enter the name of the dataset
              first_in_subtraction = "Recognizable", # First outcome in order of subtraction
              number_repetitions = 10000,  # Number of simulations
              as_extreme_as = 3.166,  # Observed statistic
              direction = "greater")  # Direction of alternative: "greater", "less", or "two-sided"
```



Explain why the null distribution is centered at the value of zero:

Interpretation of the p-value:

- Statement about probability or proportion of samples
- Statistic (summary measure and value)
- Direction of the alternative
- Null hypothesis (in context)

Conclusion:

- Amount of evidence
- Parameter of interest
- Direction of the alternative hypothesis

Confidence interval

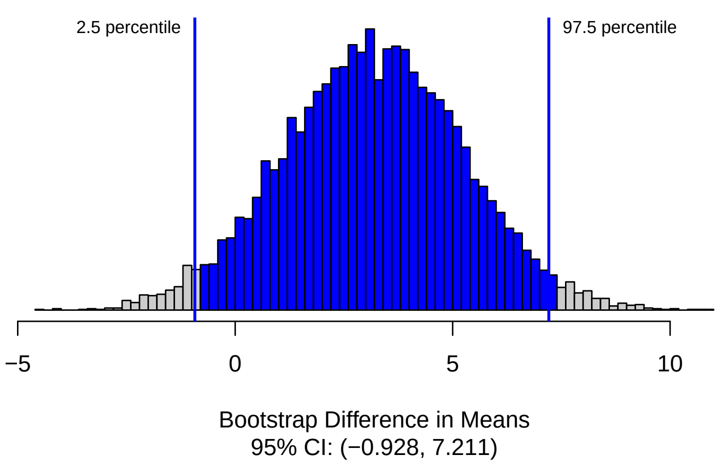
To estimate the difference in true mean we will create a confidence interval.

Simulation-based method - Video 19.2

- Write the response variable values on cards
- Keep explanatory variable groups separate
- Sample with replacement n_1 times in explanatory variable group 1 and n_2 times in explanatory variable group 2
- Calculate and plot the simulated difference in sample means from each simulation
- Repeat 1000 times (simulations) to create the bootstrap distribution
- Find the cut-offs for the middle X% (confidence level) in a bootstrap distribution.

For the letters example, we will estimate the difference in true mean number of letters recognized for students given recognizable letter groupings and students given non-recognizable letter groupings.

```
set.seed(216)
two_mean_bootstrap_CI(Memorized ~ Grouped, #Enter the name of the variables
                      data = letters,  # Enter the name of the data set
                      first_in_subtraction = "Recognizable", # First value in order of subtraction
                      number_repetitions = 10000,  # Number of simulations
                      confidence_level = 0.95)
```



Confidence interval interpretation:

- How confident you are (e.g., 90%, 95%, 98%, 99%)
- Parameter of interest
- Calculated interval
- Order of subtraction when comparing two groups

Theory-based method - Video 19.3TheoryTests

Example: Every year, orange and black monarch butterflies migrate from their summer breeding grounds in the US and Canada to mountain forests in central Mexico, where they hibernate for the winter. Due to abnormal weather patterns and drought affecting monarch habitats and feeding grounds, the population of monarch butterflies is estimated to have decreased by 53% from the 2018-2019 wintering season to the 2019-2020 wintering season (WWF, 2020). While conservationists often resort to captive-rearing with the goal of raising biologically indistinct individuals for release into the wild, tagging studies have shown that captive-reared monarchs have lower migratory success compared to wild monarchs. For this study, the researchers raised 67 monarchs (descended from wild monarchs) from eggs to maturity and then compared them to a group of 40 wild-caught monarchs. The researchers want to explore whether the maximum grip strength (how many Newtons a butterfly exerts at the moment of release when gently tugged from a mesh-covered perch) differs between captive-reared and wild-caught monarchs. Use Captive – Wild for order of subtraction.

Write the null and alternative hypotheses in notation.

H_0 :

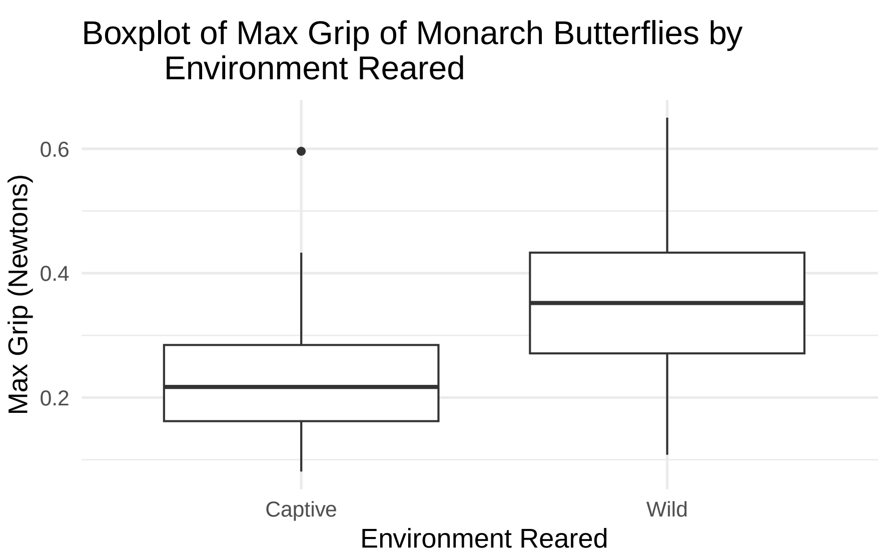
H_A :

```r
butterfly <-read.csv("data/butterfly1.csv")

butterflies <- butterfly %>% na.omit() %>%
    rename(Monarch_Group = "Monarch.Group",
           MaxGrip = "Max.Grip.Strength..N.") %>%
    mutate(Monarch_Group = factor(Monarch_Group),
           Sex = factor(Sex)) %>%
    mutate(Monarch_Group = fct_collapse(Monarch_Group,"Captive" = c("Incubator - Fall conditions", "Reari

butterflies %>%
    reframe(favstats(MaxGrip~Monarch_Group))
```

```
#>   Monarch_Group   min    Q1 median     Q3   max      mean         sd  n missing
#> 1       Captive 0.081 0.162  0.217 0.2845 0.596 0.2363731 0.09412948 67       0
#> 2          Wild 0.108 0.271  0.352 0.4330 0.650 0.3607500 0.14066796 40       0
```



Conditions:

- Independence: the response for one observational unit will not influence the outcome for another observational unit
- Large enough sample size

Like with paired data the t-distribution can be used to model the difference in means.

- For independent samples we use the ________- distribution with ______________________ degrees of freedom to approximate the sampling distribution.

Theory-based test:

- Calculate the standardized statistic
- Find the area under the t-distribution with the smallest $n-1$ df $[\min(n_1-1, n_2-1)]$ at least as extreme as the standardized statistic

Equation for the standard error of the difference in sample mean:

Equation for the standardized difference in sample mean:

Are the conditions met to analyze the butterfly data using theory based-methods?

Calculate the standardized difference in mean max grip strength.

- First calculate the $SE(\bar{x}_1 - \bar{x}_2)$

- Then calculate the T-score

What theoretical distribution should we use to find the p-value?

To find the theory-based p-value:

```
pt(-5, df=39, lower.tail=FALSE)*2
```

```
#> [1] 1.999987
```

Conclusion:

- Amount of evidence
- Parameter of interest
- Direction of the alternative hypothesis

Confidence Interval - Video 19.3TheoryIntervals

- Calculate the interval centered at the sample statistic

  statistic $\pm$ margin of error

Using the butterfly data, calculate the 99% confidence interval.

```
butterflies %>%
    reframe(favstats(MaxGrip~Monarch_Group))
```

```
#>   Monarch_Group   min    Q1 median     Q3   max      mean         sd  n missing
#> 1       Captive 0.081 0.162  0.217 0.2845 0.596 0.2363731 0.09412948 67       0
#> 2          Wild 0.108 0.271  0.352 0.4330 0.650 0.3607500 0.14066796 40       0
```

- Need the t^* multiplier for a 99% confidence interval from a t-distribution with _____________ df.

```
qt(0.995, df=39, lower.tail = TRUE)
```

```
#> [1] 2.707913
```

- We will use the same value for the $SE(\bar{x}_1 - \bar{x}_2)$ as calculated for the standardized statistic.

Calculate the margin of error for a 99% confidence interval for the parameter of interest.

Calculate a 99% confidence interval for the parameter of interest.

1.2.2 Concept Check

Be prepared for group discussion in the next class. One member from the table should write the answers to the following on the whiteboard.

1. Why is the recognizable letter study analyzed as two independent groups rather than paired data?

2. Write out the equation for the standard error for a difference in sample means.

1.3 Activity 24: Does behavior impact performance?

1.3.1 Learning outcomes

- Create a side-by-side boxplot of one categorical explanatory variable and one quantitative response variable
- Given a research question involving one categorical explanatory variable and one quantitative response variable, construct the null and alternative hypotheses in words and using appropriate statistical symbols.
- Describe and perform a simulation-based hypothesis test for a difference in means.
- Interpret and evaluate a p-value for a simulation-based hypothesis test for a difference in means.
- Use bootstrapping to find a confidence interval for a difference in means.
- Interpret a confidence interval for a difference in means.
- Use a confidence interval to determine the conclusion of a hypothesis test.

1.3.2 Terminology review

In today's activity, we will use simulation-based methods to analyze the association between one categorical explanatory variable and one quantitative response variable, where the groups formed by the categorical variable are independent. Some terms covered in this activity are:

- Independent groups
- Difference in means

To review these concepts, see Chapter 19 in the textbook.

1.3.3 Behavior and Performance

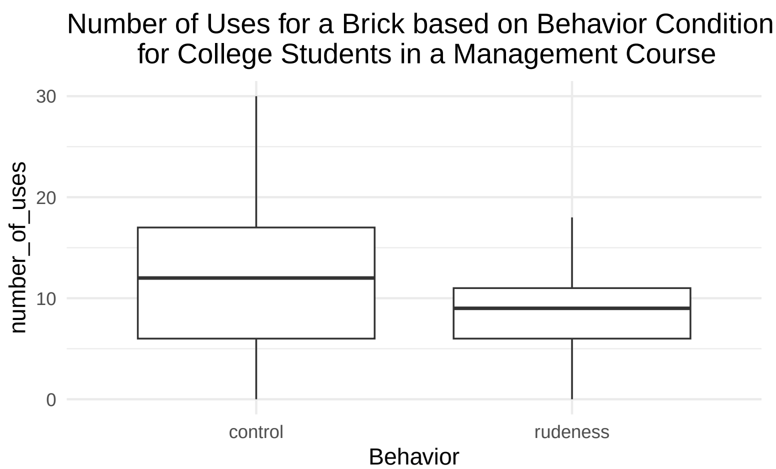
A study in the Academy of Management Journal (Porath 2017) investigated how rude behaviors influence a victim's task performance. Randomly selected college students enrolled in a management course were randomly assigned to one of two experimental conditions: rudeness condition (45 students) and control group (53 students). Each student was asked to write down as many uses for a brick as possible in five minutes; this value (total number of uses) was used as a performance measure for each student, where higher values indicate better performance. During this time another individual showed up late for class. For those students in the rudeness condition, the facilitator displayed rudeness by berating the students in general for being irresponsible and unprofessional (due to the late-arriving person). No comments were made about the late-arriving person for students in the control group. Is there evidence that the average performance score for students in the rudeness condition is lower than for students in the control group? Use the order of subtraction of rudeness – control.

- Download the R script file from D2L and upload to the RStudio server
- Highlight and run lines 1–7

```
# Read in data set
rude <- read.csv("https://math.montana.edu/courses/s216/data/rude.csv")
```

- Highlight and run lines 11–19

```
# Side-by-side box plots
rude %>%
ggplot(aes(x = condition, y = number_of_uses)) +
    geom_boxplot() +
    labs(title = "Number of Uses for a Brick based on Behavior Condition
         for College Students in a Management Course",
         x = "Behavior")
# Summary statistics
rude %>%
     reframe(favstats(number_of_uses ~ condition))
#>    condition min Q1 median Q3 max      mean       sd  n missing
#> 1    control   0  6     12 17  30 11.811321 7.382559 53       0
#> 2   rudeness   0  6      9 11  18  8.511111 3.992164 45       0
```



Quantitative variables review

1. Compare the distributions of the number of bricks between the two treatment conditions.
    - What is the shape of each group?
    - Which group has the higher center?
    - What group has the larger spread?
    - Does either distribution have outliers?
2. Is this an experiment or an observational study? Justify your answer.

3. Explain why this is two independent samples and not paired data.

Ask a research question

In this study we are assessing the difference in true mean number of uses for a brick given by college students enrolled in a management course assigned to a rudeness condition and for those assigned to a control group.

4. What assumption are we making about the difference in true mean?

5. Write the alternative hypothesis in notation.

Numerically Summarize the data

6. Calculate the summary statistic of interest (difference in means). What is the appropriate notation for this statistic?

Interpret this calculated value.

7. Write out the parameter of interest for this study in context of the study.
   - To write in context:
     - Population word (true, long-run, population)
     - Summary measure (depends on the type of data)
     - Context
       - Observational units
       - Variable(s)

In this study we are assessing the difference in true mean number of uses for a brick given by college students enrolled in a management course assigned to a rudeness condition and for those assigned to a control group.

Use statistical inferential methods to draw inferences from the data

Hypothesis test Remember that the null distribution is created based on the assumption the null hypothesis is true. In this study, the null hypothesis states that there is no association between the two variables. This means that the values observed in the data set would have been the same regardless of the behavior condition.

To demonstrate this simulation, we could create cards to simulate a sample.

- Write the number of uses for a brick given by each student on one card.
- Mix together and shuffle into two piles, one with 45 cards to represent the rudeness condition and one with 53 cards to represent the control group.
- Calculate the difference in mean number of uses for a brick (rudeness - control)

We will use the `two_mean_test()` function in R (in the `catstats` package) to simulate the null distribution of differences in sample means and compute a p-value.

- Fill in the response and explanatory variable names
- Fill in the missing values/names for the xx's in the R script file
- Highlight and run lines 25–30

```
set.seed(216)
two_mean_test(response~explanatory, #Enter the names of the variables
              data = rude,  # Enter the name of the dataset
              first_in_subtraction = "xx", # First outcome in order of subtraction
              number_repetitions = 1000,  # Number of simulations
              as_extreme_as = xx,  # Observed statistic
              direction = "xx")  # Direction of alternative: "greater", "less", or "two-sided"
```

8. Report the p-value. Based off of this p-value, write a conclusion to the hypothesis test.

Confidence interval We will use the `two_proportion_bootstrap_CI()` function in R (in the `catstats` package) to simulate the bootstrap distribution of differences in sample proportions and calculate a confidence interval. We will need to enter the response variable name and the explanatory variable name for the formula, the data set name (identified above as `rude`), the outcome for the explanatory variable that is first in subtraction, number of repetitions, the outcome for the response variable that is a success (the count for the numerator when calculating a sample proportion), and the confidence level as a decimal.

The response variable name is `number_of_uses` and the explanatory variable name is `condition`.

9. What values should be entered for each of the following into the simulation to create a 99% confidence interval?

- First in subtraction (What is the outcome for the explanatory variable that is used as first in the order of subtraction? `"rudeness"` or `"control"`):
- Number of repetitions:

- Confidence level (entered as a decimal):

Using the R script file for this activity, enter your answers for question 9 in place of the `xx`'s to produce the bootstrap distribution with 1000 simulations; highlight and run lines 35–39.

```
two_mean_bootstrap_CI(response ~ explanatory, #Enter the name of the variables
                      data = rude,  # Enter the name of the data set
                      first_in_subtraction = "xx", # First value in order of subtraction
                      number_repetitions = 1000,  # Number of simulations
                      confidence_level = xx)
```

10. Where is the bootstrap distribution centered? Explain why.

11. Report the bootstrap 99% confidence interval.

12. What percentile of the bootstrap distribution does the upper value of the confidence interval represent?

13. Interpret the 99% confidence interval.

1.3.4 Take-home messages

1. This activity differs from the activities in Module 11 because the responses are independent, not paired. These data are analyzed as a difference in means, not a mean difference.
2. To create one simulated sample on the null distribution for a difference in sample means, label cards with the response variable values from the original data. Mix cards together and shuffle into two new groups of sizes n_1 and n_2. Calculate and plot the difference in means.
3. To create one simulated sample on the bootstrap distribution for a difference in sample means, label $n_1 + n_2$ cards with the original response values. Keep groups separate and randomly draw with replacement n_1 times from group 1 and n_2 times from group 2. Calculate and plot the resampled difference in means.

1.3.5 Additional notes

Use this space to summarize your thoughts and take additional notes on today's activity and material covered

1.4 Activity 25: Moon Phases and Virtual Reality

1.4.1 Learning outcomes

- Given a research question involving one categorical explanatory variable and one quantitative response variable, construct the null and alternative hypotheses in words and using appropriate statistical symbols.
- Describe and perform a theory-based hypothesis test for a difference in means.
- Interpret and evaluate a p-value for a theory-based hypothesis test for a difference in means.
- Use theory-based methods to find a confidence interval for a difference in means.
- Interpret a confidence interval for a difference in means.
- Use a confidence interval to determine the conclusion of a hypothesis test.

1.4.2 Terminology review

In today's activity, we will use theory-based methods to analyze the association between one categorical explanatory variable and one quantitative response variable, where the groups formed by the categorical variable are independent. Some terms covered in this activity are:

- Difference in means
- Independence within and between groups
- Normality

To review these concepts, see Chapter 19 in the textbook.

1.4.3 Moon Phases and Virtual Reality

In a study comparing immersive virtual reality (VR) to traditional hands-on methods, researchers recruited 115 undergraduate students to assess the effectiveness of these approaches in teaching complex scientific concepts like Moon phases (Madden 2020). Participants were randomly assigned to experience either a VR simulation replicating the Sun-Earth-Moon system or a hands-on activity where they physically manipulated models to observe Moon phases. The students were given a 14 multiple choice question quiz about Moon phases and the Moon's motion relative to the Earth to evaluate their understanding of Moon phases and the Moon's motion. Each question had only one correct answer, and the participant's score was the sum of the number of correct answers, with all questions weighted equally (with a maximum score of 14). Is there evidence of a difference, on average, in student learning comparing those using VR methods to those using the traditional method? Use order of subtraction VR – Hands-on.

- Download the RScript file and dataset from D2L and upload to the RStudio server
- Open the RScript file
- Enter the name of the dataset for datasetname in line

1. Write out the parameter of interest in words in context of the study.
    - To write in context:
        - Population word (true, long-run, population)
        - Summary measure (depends on the type of data)
        - Context
            - Observational units

* Variable(s)

2. Write out the null hypothesis in notation for this study. Be sure to clearly identify the subscripts.

3. Write out the alternative hypothesis in words for this study.

The sampling distribution for $\bar{x}_1 - \bar{x}_2$ can be modeled using a normal distribution when certain conditions are met.

Conditions for the sampling distribution of $\bar{x}_1 - \bar{x}_2$ to follow an approximate normal distribution:

- **Independence**: The sample's observations are independent
- **Normality**: Each sample should be approximately normal or have a large sample size. For *each* sample:
  - $n < 30$: If the sample size n is less than 30 and there are no clear outliers in the data, then we typically assume the data come from a nearly normal distribution to satisfy the condition.
  - $30 \leq n < 100$: If the sample size n is between 30 and 100 and there are no particularly extreme outliers, then we typically assume the sampling distribution of $\bar{x}$ is nearly normal, even if the underlying distribution of individual observations is not.
  - $n \geq 100$: If the sample size n is at least 100 (regardless of the presence of skew or outliers), we typically assume the sampling distribution of $\bar{x}$ is nearly normal, even if the underlying distribution of individual observations is not.

To create the plots of the data:

- Enter a title in line 15 for the plot between the quotations
- Highlight and run lines 12 - 17

```
moon <- read.csv("data/Moon_VR.csv")
moon %>%  # Data set piped into...
  ggplot(aes(y = TestScore, x = Method))+  # Identify variables
  geom_boxplot()+  # Tell it to make a box plot
  labs(title = "Boxplots of Test Scores for Undergraduate Students Comparing VR
       Teaching Methods and Traditional Teaching Methods",  # Title
       x = "Methods",    # x-axis label
       y = "Test Score (points)")  # y-axis label
```

To find the summary statistic:

- Enter the response and explanatory variable names in line 22
- Highlight and run lines 21–22

```r
moon %>%
  reframe(favstats(response~explanatory))
```

4. Can theory-based methods be used to analyze these data?

5. Calculate the summary statistic (difference in means) for this study. Use appropriate notation with clearly defined subscripts.

Use statistical inferential methods to draw inferences from the data

To find the standardized statistic for the difference in means we will calculate:

$$T = \frac{\bar{x}_1 - \bar{x}_2 - 0}{SE(\bar{x}_1 - \bar{x}_2)},$$

where the standard error of the difference in means is calculated using:

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

6. Calculate the standard error for the difference in sample means.

7. Calculate the standardized statistic for the difference in sample means.

To find the degrees of freedom to use for the t-distribution, we need to use the group with the smallest sample size and subtract 1. (`df` = minimum of $n_1 - 1$ or $n_2 - 1$).

- Enter the value of the standardized statistic for xx
- Enter the df for yy
- Highlight and run line 28

```r
2*pt(xx, df=yy, lower.tail=FALSE)
```

8. What is the p-value for the study?

To calculate a theory-based 95% confidence interval for a difference in means, use the formula:

$$(\bar{x}_1 - \bar{x}_2) \pm (t^* \times SE(\bar{x}_1 - \bar{x}_2))$$

We will need to find the t^* multiplier using the function `qt()`.

- Enter the percentile to find the multiplier for a 95% confidence level
- Enter the degrees of freedom for yy
- Highlight and run line 34

```
qt(percentile, df = yy, lower.tail=TRUE)
```

9. Calculate the margin of error for a 95% confidence interval.

10. Calculate the 95% confidence interval.

11. Write a conclusion to the test.

1.4.4 Take-home messages

1. In order to use theory-based methods for independent groups, the normality condition must be met for each sample.
2. A T-score is compared to a t-distribution with the minimum $n-1$ df in order to calculate a one-sided p-value. To find a two-sided p-value using theory-based methods we need to multiply the one-sided p-value by 2.
3. A t^* multiplier is found by obtaining the bounds of the middle X% (X being the desired confidence level) of a t-distribution with the minimum $n-1$ df.

1.4.5 Additional notes

Use this space to summarize your thoughts and take additional notes on today's activity and material covered

1.5 Module 12 Lab: Trustworthiness

1.5.1 Learning outcomes

- Given a research question involving one categorical explanatory variable and one quantitative response variable, construct the null and alternative hypotheses in words and using appropriate statistical symbols.
- Describe and perform a theory-based hypothesis test for a difference in means.
- Interpret and evaluate a p-value for a theory-based hypothesis test for a difference in means.
- Use theory-based methods to find a confidence interval for a difference in means.
- Interpret a confidence interval for a difference in means.
- Use a confidence interval to determine the conclusion of a hypothesis test.

1.5.2 Trustworthiness

Researchers in India wanted to find out how trustworthy famous YouTubers are (Kalra 2022). They went through a process in which they collected data on many videos from famous YouTubers to determine a trustworthiness score. Scientists randomly selected videos from famous YouTubers (>1000 subscribers) to include in the study. There were many different factors that went into calculating the trustworthiness score. Researchers also recorded if YouTubers were a subject matter expert (SME) or not a subject matter expert (non-SME). An example of an SME would be if one of your statistics professors made a YouTube video of how to do hypothesis testing. An example of someone who isn't an SME would be if one of your friends who has never taken a civil engineering class in their life decided to make a YouTube video about how to build a bridge. There were 621 Youtubers who are SMEs in the sample and 1026 who aren't SMEs. Is there evidence of a difference in mean trustworthiness score between subject matter experts (SME) YouTubers and non-SME YouTubers? Use SME − Non -SME as the order of subtraction

1. **Write out the parameter of interest in words in context of the study.**

2. Write out the null hypothesis in notation for this study. Be sure to clearly identify the subscripts.

3. Write out the alternative hypothesis in words for this study.

The sampling distribution for $\bar{x}_1 - \bar{x}_2$ can be modeled using a normal distribution when certain conditions are met.

Conditions for the sampling distribution of $\bar{x}_1 - \bar{x}_2$ to follow an approximate normal distribution:

- **Independence**: The sample's observations are independent
- **Normality**: Each sample should be approximately normal or have a large sample size. For *each* sample:

- $n < 30$: If the sample size n is less than 30 and there are no clear outliers in the data, then we typically assume the data come from a nearly normal distribution to satisfy the condition.
- $30 \leq n < 100$: If the sample size n is between 30 and 100 and there are no particularly extreme outliers, then we typically assume the sampling distribution of $\bar{x}$ is nearly normal, even if the underlying distribution of individual observations is not.
- $n \geq 100$: If the sample size n is at least 100 (regardless of the presence of skew or outliers), we typically assume the sampling distribution of $\bar{x}$ is nearly normal, even if the underlying distribution of individual observations is not.

- Upload and open the R script file for Module 10 lab. Upload the csv file, `Trustworthiness.csv`.
- Enter the name of the data set for datasetname in the R script file in line 10.
- Write a title for the boxplots in line 14.
- Highlight and run lines 1–16 to load the data and create plots of the data.

```
trust <- read.csv("datasetname")
trust %>%  # Data set piped into...
  ggplot(aes(y = Trustworthiness_Video, x = Creator_SME))+  # Identify variables
  geom_boxplot()+  # Tell it to make a box plot
  labs(title = "Don't forget to include a title",  # Title: should include the type of plot,
       # observational units, variables
       x = "Whether the Creator is SME",     # x-axis label
       y = "Trustworthiness Score")  # y-axis label
```

4. Is the independence condition met? Explain your answer.

5. Check that the normality condition is met to use theory-based methods to analyze these data.

- Enter the name of the explanatory variable for `explanatory` and the name of the response variable for `response` in line 22.
- Highlight and run lines 21–22 to get the summary statistics for the data.

```
trust %>%
  reframe(favstats(response~explantory))
```

6. **Calculate the summary measure (difference in means) for this study. Use appropriate notation with clearly defined subscripts.**

Use statistical inferential methods to draw inferences from the data

To find the standardized statistic for the difference in means we will calculate:

$$T = \frac{\bar{x}_1 - \bar{x}_2 - 0}{SE(\bar{x}_1 - \bar{x}_2)},$$

where the standard error of the difference in means is calculated using:

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

7. Calculate the standard error for the difference in sample means.

8. **Calculate the standardized statistic for the difference in sample means.**

9. When we are comparing two quantitative variables to find the degrees of freedom to use for the t-distribution, we need to use the group with the smallest sample size and subtract 1. (`df` = minimum of $n_1 - 1$ or $n_2 - 1$). Calculate the `df` for this study.

10. Using the provided R script file, enter the T-score (for `xx`) and the `df` calculated in question 9 for `yy` into the `pt()` function to find the p-value. Highlight and run line 27. Report the p-value calculated.

```
2*pt(xx, df=yy, lower.tail=FALSE)
```

11. **Explain why we multiplied by 2 in the code above.**

12. Do you expect the 95% confidence interval to contain the null value of zero? Explain your answer.

To calculate a theory-based 95% confidence interval for a difference in means, use the formula:

$$(\bar{x}_1 - \bar{x}_2) \pm (t^* \times SE(\bar{x}_1 - \bar{x}_2))$$

We will need to find the t^* multiplier using the function `qt()`. For a 95% confidence level, we are finding the t^* value at the 97.5th percentile with (`df` = minimum of $n_1 - 1$ or $n_2 - 1$).

- Enter the appropriate percentile value (as a decimal) for `xx` and degrees of freedom for `yy` into the `qt()` function at line 32 to find the appropriate t^* multiplier

```
qt(xx, df = yy, lower.tail=FALSE)
```

13. Report the t^* multiplier for the 95% confidence interval.

14. Calculate the 95% confidence interval using theory-based methods.

15. Do the results of the CI agree with the p-value? Explain your answer.

16. What type of error may be possible?

17. Write a paragraph summarizing the results of the study as if you are reporting the results to your supervisor. **Upload a copy of your paragraph to Gradescope for your group.** Be sure to describe:

- Summary statistic and interpretation
- P-value and interpretation
  - Statement about probability or proportion of samples
  - Statistic (summary measure and value)
  - Direction of the alternative
  - Null hypothesis (in context)
- Confidence interval and interpretation
  - How confident you are (e.g., 90%, 95%, 98%, 99%)
  - Parameter of interest
  - Calculated interval
  - Order of subtraction when comparing two groups
- Conclusion (written to answer the research question)
  - Amount of evidence
  - Parameter of interest
  - Direction of the alternative hypothesis
- Scope of inference

Paragraph continued:

"Average Driving Distance and Fairway Accuracy." 2008. https://www.pga.com/ and https://www.lpga.com/.

Banton, et al, S. 2022. "Jog with Your Dog: Dog Owner Exercise Routines Predict Dog Exercise Routines and Perception of Ideal Body Weight." *PLoS ONE* 17(8).

Bhavsar, et al, A. 2022. "Increased Risk of Herpes Zoster in Adults ≥50 Years Old Diagnosed with COVID-19 in the United States." *Open Forum Infectious Diseases* 9(5).

Bulmer, M. n.d. "Islands in Schools Project." https://sites.google.com/site/islandsinschoolsprojectwebsite/home.

"Bureau of Transportation Statistics." 2019. https://www.bts.gov/.

"Child Health and Development Studies." n.d. https://www.chdstudies.org/.

Darley, J. M., and C. D. Batson. 1973. ""From Jerusalem to Jericho": A Study of Situational and Dispositional Variables in Helping Behavior." *Journal of Personality and Social Psychology* 27: 100–108.

Davis, Smith, A. K. 2020. "A Poor Substitute for the Real Thing: Captive-Reared Monarch Butterflies Are Weaker, Paler and Have Less Elongated Wings Than Wild Migrants." *Biology Letters* 16.

Du Toit, et al, G. 2015. "Randomized Trial of Peanut Consumption in Infants at Risk for Peanut Allergy." *New England Journal of Medicine* 372.

Edmunds, et al, D. 2016. "Chronic Wasting Disease Drives Population Decline of White-Tailed Deer." *PLoS ONE* 11(8).

Education Statistics, National Center for. 2018. "IPEDS." https://nces.ed.gov/ipeds/.

"Great Britain Married Couples: Great Britain Office of Population Census and Surveys." n.d. https://discovery.nationalarchives.gov.uk/details/r/C13351.

Group, TODAY Study. 2012. "A Clinical Trial to Maintain Glycemic Control in Youth with Type 2 Diabetes." *New England Journal of Medicine* 366: 2247–56.

Hamblin, J. K., K. Wynn, and P. Bloom. 2007. "Social Evaluation by Preverbal Infants." *Nature* 450 (6288): 557–59.

Hirschfelder, A., and P. F. Molin. 2018. "I Is for Ignoble: Stereotyping Native Americans." Retrieved from https://www.ferris.edu/HTMLS/news/jimcrow/native/homepage.htm.

Hutchison, R. L., and M. A. Hirthler. 2013. "Upper Extremity Injuies in Homer's Iliad." *Journal of Hand Surgery (American Volume)* 38: 1790–93.

"IMDb Movies Extensive Dataset." 2016. https://kaggle.com/stefanoleone992/imdb-extensive-dataset.

Kalra, et al., Dl. 2022. "Trustworthiness of Indian Youtubers." Kaggle. https://doi.org/10.34740/KAGGLE/DSV/4426566.

Keating, D., N. Ahmed, F. Nirappil, Stanley-Becker I., and L. Bernstein. 2021. "Coronavirus Infections Dropping Where People Are Vaccinated, Rising Where They Are Not, Post Analysis Finds." *Washington Post*. https://www.washingtonpost.com/health/2021/06/14/covid-cases-vaccination-rates/.

Laeng, Mathisen, B. 2007. "Why Do Blue-Eyed Men Prefer Women with the Same Eye Color?" *Behavioral Ecology and Sociobiology* 61(3).

Levin, D. T. 2000. "Race as a Visual Feature: Using Visual Search and Perceptual Discrimination Tasks to Understand Face Categories and the Cross-Race Recognition Deficit." *Journal of Experimental Psychology* 129(4).

LUETKEMEIER, et al., M. 2017. "Skin Tattoos Alter Sweat Rate and Na+ Concentration." *Medicine and Science in Sports and Exercise* 49(7).

Madden, et al, J. 2020. "Ready Student One: Exploring the Predictors of Student Learning in Virtual Reality." *PLoS ONE* 15(3).

Miller, G. A. 1956. "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information." *Psychological Review* 63(2).

Moquin, W., and C. Van Doren. 1973. "Great Documents in American Indian History." Praeger.

"More Americans Are Joining the 'Cashless' Economy." 2022. https://www.pewresearch.org/short-reads/2022/10/05/more-americans-are-joining-the-cashless-economy/.

National Weather Service Corporate Image Web Team. n.d. "National Weather Service – NWS Billings." https://w2.weather.gov/climate/xmacis.php?wfo=byz.

O'Brien, Lynch, H. D. 2019. "Crocodylian Head Width Allometry and Phylogenetic Prediction of Body Size in Extinct Crocodyliforms." *Integrative Organismal Biology* 1.

"Ocean Temperature and Salinity Study." n.d. https://calcofi.org/.

"Older People Who Get Covid Are at Increased Risk of Getting Shingles." 2022. https://www.washingtonpost.com/health/2022/04/19/shingles-and-covid-over-50/.

"Physician's Health Study." n.d. https://phs.bwh.harvard.edu/.

Porath, Erez, C. 2017. "Does Rudeness Really Matter? The Effects of Rudeness on Task Performance and Helpfulness." *Academy of Management Journal* 50.

Quinn, G. E., C. H. Shin, M. G. Maguire, and R. A. Stone. 1999. "Myopia and Ambient Lighting at Night." *Nature* 399 (6732): 113–14. https://doi.org/10.1038/20094.

Ramachandran, V. 2007. "3 Clues to Understanding Your Brain." https://www.ted.com/talks/vs_ramachandran_3_clues_to_understanding_your_brain.

"Rates of Laboratory-Confimed COVID-19 Hospitalizations by Vaccination Status." 2021. CDC. https://covid.cdc.gov/covid-data-tracker/#covidnet-hospitalizations-vaccination.

Richardson, T., and R. T. Gilman. 2019. "Left-Handedness Is Associated with Greater Fighting Success in Humans." *Scientific Reports* 9 (1): 15402. https://doi.org/10.1038/s41598-019-51975-3.

Stephens, R., and O. Robertson. 2020. "Swearing as a Response to Pain: Assessing Hypoalgesic Effects of Novel "Swear" Words." *Frontiers in Psychology* 11: 643–62.

Stewart, E. H., B. Davis, B. L. Clemans-Taylor, B. Littenberg, C. A. Estrada, and R. M. Centor. 2014. "Rapid Antigen Group a Streptococcus Test to Diagnose Pharyngitis: A Systematic Review and Meta-Analysis" 9 (11). https://doi.org/10.1371/journal.pone.0111727.

Stroop, J. R. 1935. "Studies of Interference in Serial Verbal Reactions." *Journal of Experimental Psychology* 18: 643–62.

Subach, et al, A. 2022. "Foraging Behaviour, Habitat Use and Population Size of the Desert Horned Viper in the Negev Desert." *Soc.Open Sci* 9.

Sulheim, S., A. Ekeland, I. Holme, and R. Bahr. 2017. "Helmet Use and Risk of Head Injuries in Alpine Skiers and Snowboarders: Changes After an Interval of One Decade" 51 (1): 44–50. https://doi.org/10.1136/bjsports-2015-095798.

"Titanic." n.d. http://www.encyclopedia-titanica.org.

"US COVID-19 Vaccine Tracker: See Your State's Progress." 2021. Mayo Clinic. https://www.mayoclinic.org/coronavirus-covid-19/vaccine-tracker.

US Environmental Protection Agency. n.d. "Air Data – Daily Air Quality Tracker." https://www.epa.gov/outdoor-air-quality-data/air-data-daily-air-quality-tracker.

Wahlstrom, et al, K. 2014. "Examining the Impact of Later School Start Times on the Health and Academic Performance of High School Students: A Multi-Site Study." *Center for Applied Research and Educational Improvement*.

Watson, et al., N. 2015. "Recommended Amount of Sleep for a Heathy Adult: A Joint Consensus Statement of the American Academy of Sleep Medicine and Sleep Research Society." *Sleep* 38(6).

Weiss, R. D. 1988. "Relapse to Cocaine Abuse After Initiating Desipramine Treatment." *JAMA* 260(17).

"Welcome to the Navajo Nation Government: Official Site of the Navajo Nation." 2011.Retrieved from https://www.navajo-nsn.gov/.

Wilson, Woodruff, J. P. 2016. "Vertebral Adaptations to Large Body Size in Theropod Dinosaurs." *PLoS ONE* 11(7).