

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	1
1 Exploring Quantitative Data: Exploratory Data Analysis and Hypothesis Testing for a Single Quantitative Variable	2
1.1 Vocabulary Review and Key Topics	2
1.2 Video Notes: Exploratory Data Analysis of Quantitative Variables	5
1.3 Activity 11: Summarizing Quantitative Variables	18
1.4 Activity 12: Hypothesis Testing of a Single Quantitative Variable	24
1.5 Activity 13: Body Temperature	28
2 Confidence Intervals for a Single Quantitative Variable	33
2.1 Vocabulary Review and Key Topics	33
2.2 Video Notes: Theory-based Inference for a single quantitative variable	35
2.3 Activity 14: Danceability of Songs	40
2.4 Activity 15: Errors and Power	44
2.5 Module 6 and 7 Lab: Arsenic	48
3 Exploratory Data Analysis and Simulation-based Inference for Two Categorical Variables	53
3.1 Vocabulary Review and Key Topics	53
3.2 Video Notes: Inference for Two Categorical Variables using Simulation-based Methods	55
3.3 Activity 16: Study Design	66
3.4 Activity 17: Summarizing Two Categorical Variables	72
3.5 Activity 18: The Good Samaritan	77
4 Inference for a Two Categorical Variable: Theory-based Methods	82
4.1 Vocabulary Review and Key Topics	82
4.2 Video Notes: Theoretical Inference for Two Categorical Variables	83
4.3 Activity 19: Winter Sports Helmet Use and Head Injuries — Theory-based Methods	89
4.4 Activity 20: Diabetes	94
4.5 Module 8 Lab: Poisonous Mushrooms	98
5 Unit 2 Review	102
5.1 Module 6 Review - One Mean Testing	103
5.2 Module 7 Review - One Mean Confidence Interval	107
5.3 Module 8 - 9 Review	111
5.4 Key Topics Exam 2	117
References	119

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.

Exploring Quantitative Data: Exploratory Data Analysis and Hypothesis Testing for a Single Quantitative Variable

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 single quantitative variable.

1.1.1 Key topics

Module 6 will introduce hypothesis testing using both simulation-based and theory-based methods for a single quantitative variable.

- The **summary measure** for one quantitative variable is the **mean**
- Additionally, we can find the five number summary (min, Q1, median, Q3, max) as well as the sample standard deviation

Exploratory data analysis

At the end of this module, you should understand how to calculate a summary statistic and plot a single quantitative variable.

- Notation for a sample mean: \bar{x}
- Notation for a population mean: μ
- Types of plots for a single categorical variable:
 - Histogram
 - Boxplot
 - Dotplot
- R code to find the summary statistics for a quantitative variable

```
object %>% # Data set piped into...
  summarise(favstats(variable))
```

Simulation-based Hypothesis Testing

Hypotheses for a single quantitative variable:

$$H_0 : \mu = \mu_0$$

$$H_A : \mu \left\{ \begin{array}{c} < \\ \neq \\ > \end{array} \right\} \mu_0$$

- R code to use for **simulation methods** for one quantitative variable to find the p-value, `one_mean_test`, is shown below. Review the comments (instructions after the `#`) to see what each should be entered for each line of code.

```
one_mean_test(object$variable, #Enter the object name and variable
  null_value = xx, #Enter the null value for the study
  summary_measure = "mean", #Can choose between mean or median
  shift = xx, #Difference between the null value and the sample mean
  as_extreme_as = xx, #Value of the summary statistic
  direction = "xx", #Specify direction of alternative hypothesis
  number_repetitions = 10000)
```

Theory-based Hypothesis Testing

- **Theory-based methods:** when specific conditions are met, the distribution of sample statistics if we were to repeatedly sample from the population can be fit with a theoretical distribution
- **Conditions for the sampling distribution of \bar{x} to follow an approximate normal distribution:**
 - **Independence:** The sample's observations are independent, e.g., are from a simple random sample. (*Remember:* This also must be true to use simulation methods!)
 - **Large enough sample size: Normality Condition:** The sample observations come from a normally distributed population. To check use the the following rules of thumb:
 - * $n < 30$: The distribution of the sample must be approximately normal with no outliers
 - * $30 \geq n < 100$: We can relax the condition a little; the distribution of the sample must have no extreme outliers or skewness
 - * $n > 100$: Can assume the sampling distribution of \bar{x} is nearly normal, even if the underlying distribuion of individual observations is not
- **t-distribution:** a theoretical distribution that is symmetric with a given degrees of freedom ($n - 1$)
 - t_{n-1}
- **Standardized sample mean:** standardized statistic for a single quantitative variable calculated using:

$$T = \frac{\bar{x} - \mu_0}{SE(\bar{x})},$$

- **Standard error of the sample mean assuming the null is true:** measures the how far each possible sample mean is from the true mean, on average and is calculated using the formula below:

$$SE(\bar{x}) = \frac{s}{\sqrt{n}}$$

- The following R code is used to find the p-value using theory based methods for a single quantitative variables.
 - `pt` will give you a p-value using the t-distribution with $n-1$ df (enter for `yy`)
 - Enter the value of the standardized statistic for `xx`
 - 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)
```

1.1.2 Vocabulary

- Quartile 1, Q_1 : value at the 25th percentile; approximately 25% of data values are at and below the value of Q_1
- Quartile 2, Q_3 : value at the 75th percentile; approximately 75% of data values are at and below the value of Q_3
- Sample standard deviation, s : on average, each value in the data set is s units from the mean of the data set
- **Interquartile range:** $IQR = Q_3 - Q_1$

Types of plot for one quantitative variables:

- **Histogram:** sorts a quantitative variable into bins of a certain width
- R code to create a histogram

```
object %>% # Data set piped into...
  ggplot(aes(x = variable)) + # Name variable to plot
  geom_histogram(binwidth = 10) + # Create histogram with specified binwidth
  labs(title = "Don't forget to title the plot!", # Title for plot
        x = "x-axis label", # Label for x axis
        y = "y-axis label") # Label for y axis
```

- **Boxplot:** plots the values of the five number summary and shows any outliers in the data set
- R code to create a boxplot

```
object %>% # Data set piped into...
  ggplot(aes(x = variable)) + # Name variable to plot
  geom_boxplot() + # Create boxplot
  labs(title = "Don't forget to title the plot!", # Title for plot
        x = "x-axis label", # Label for x axis
        y = "y-axis label") # Label for y axis
```

- **Dotplot:** plots each value as a dot along the x-axis
- Four characteristics of boxplots
 - Shape (symmetric or skewed)
 - Center
 - Spread
 - Outliers?

1.2 Video Notes: Exploratory Data Analysis of Quantitative Variables

Read Chapters 5 and 17 in the course textbook. Use the following videos to complete the video notes for Module 6.

1.2.1 Course Videos

- QuantitativeData
- 5.5to5.6
- 5.7
- 17.2
- 17.3TheoryTests

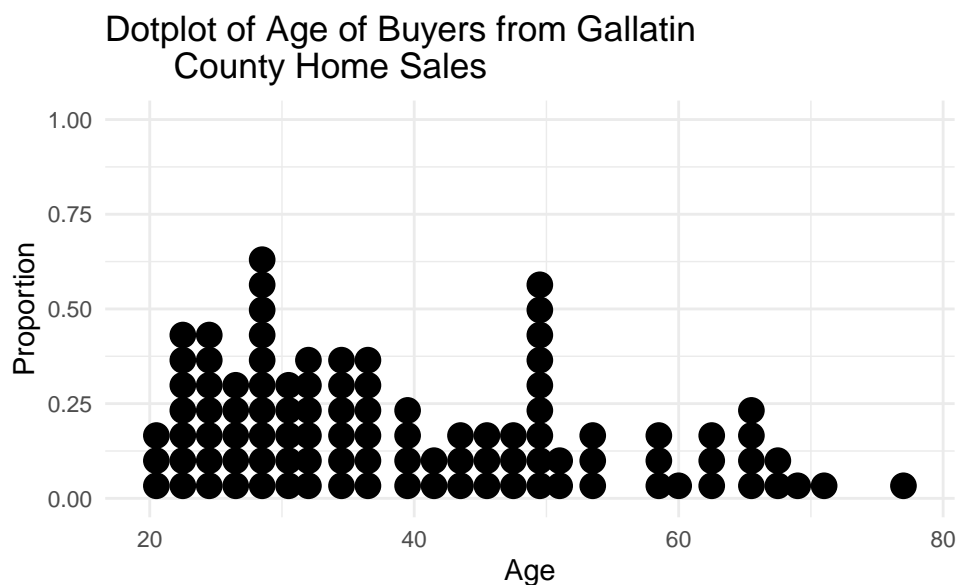
Summarizing quantitative data - Videos 5.2to5.4 and 5.5to5.6

Types of plots

We will revisit the moving to Montana data set and plot the age of the buyers.

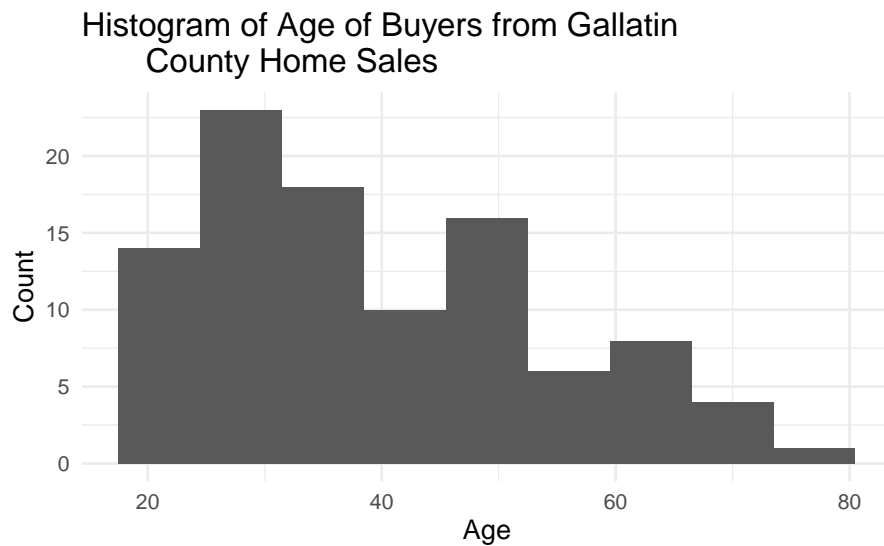
Dotplot:

```
moving %>%  
  ggplot(aes(x = Age)) + #Enter variable to plot  
  geom_dotplot() +  
  labs(title = "Dotplot of Age of Buyers from Gallatin  
    County Home Sales", #Title your plot  
    x = "Age", #x-axis label  
    y = "Proportion") #y-axis label
```



Histogram:

```
moving %>%  
  ggplot(aes(x = Age))+  
  geom_histogram(binwidth = 7) +  
  labs(title = "Histogram of Age of Buyers from Gallatin  
    County Home Sales",  
        #Title your plot  
        x = "Age",  
        y = "Count")
```



Quantitative data can be numerically summarized by finding:

Two measures of center:

- Mean: _____ of all the _____ in the data set.
 - Sum the values in the data set and divide the sum by the sample size
- Notation used for the population mean:
 - Single quantitative variable:
 - One categorical and one quantitative variable:
 - Subscripts represent the _____ variable groups
- Notation used for the sample mean:
 - Single quantitative variable:
 - One categorical and one quantitative variable:

- Median: Value at the _____ percentile
 - _____ % of values are at and _____ and at _____ the value of the _____.
 - Middle value in a list of ordered values

Two measures of spread:

- Standard deviation: Average _____ each data point is from the _____ of the data set.
 - Notation used for the population standard deviation
 - Notation used for the sample standard deviation

- Interquartile range: middle 50% of data values

Formula:

Quartile 3 (Q3) - value at the 75th percentile

- _____ % of values are at and _____ the value of Q3

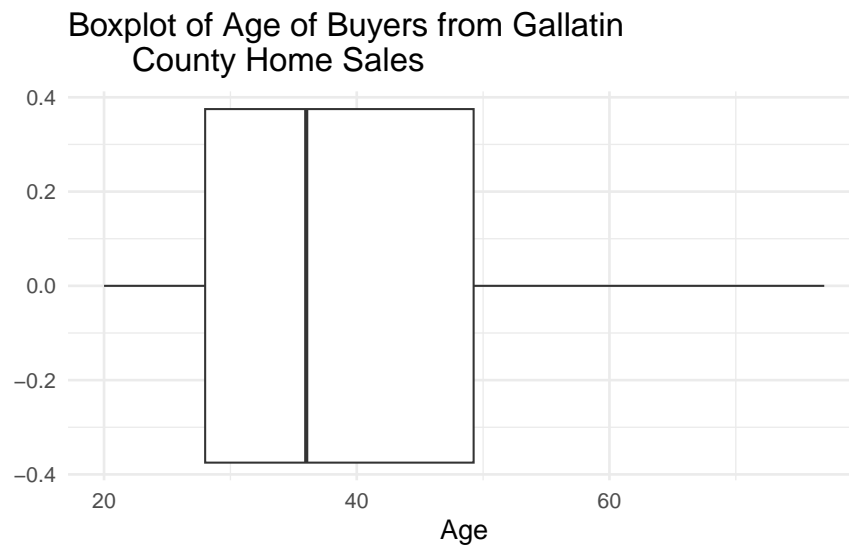
Quartile 1 (Q1) - value at the 25th percentile

- _____ % of values are at and _____ the value of Q1

Boxplot (3rd type of plot for quantitative variables)

- Five number summary: minimum, Q1, median, Q3, maximum

```
moving %>%  
  ggplot(aes(x = Age))+ #Enter variable to plot  
  geom_boxplot() +  
  labs(title = "Boxplot of Age of Buyers from Gallatin  
    County Home Sales", #Title your plot  
    x = "Age", #x-axis label  
    y = "") #y-axis label
```



```
favstats(moving$Age)
```

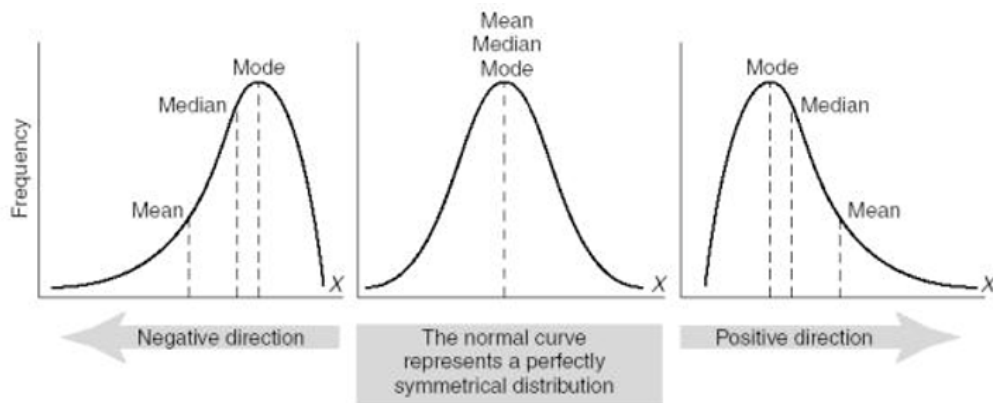
```
#>   min  Q1 median   Q3 max  mean    sd   n missing  
#>   20  28   36 49.25  77 39.77 14.35471 100      0
```

Interpret the value of Q_3 for the age of buyers.

Interpret the value of s for the age of buyers.

Four characteristics of plots for quantitative variables

- Shape: overall pattern of the data



- What is the shape of the distribution of age of buyers for Gallatin County home sales?

- Center:

Mean or Median

- Report the measure of center for the boxplot of age of buyers for Gallatin County home sales.

- Spread (or variability):

Standard deviation or IQR

- Report the IQR for the distribution of age of buyers from Gallatin County home sales.

- Outliers?

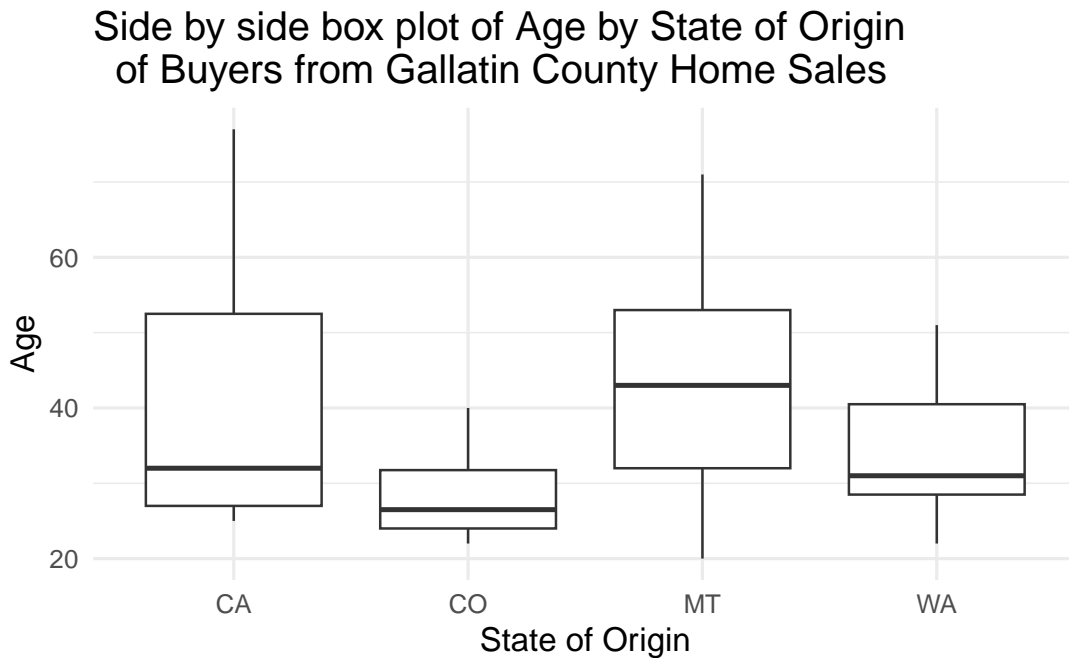
values $< Q_1 - 1.5 \times IQR$

values $> Q_3 + 1.5 \times IQR$

- Use these formulas to show that there are no outliers in the distribution of age of buyers from Gallatin County home sales.

Let's look at side-by-side boxplot of the variable age by state of origin moved from.

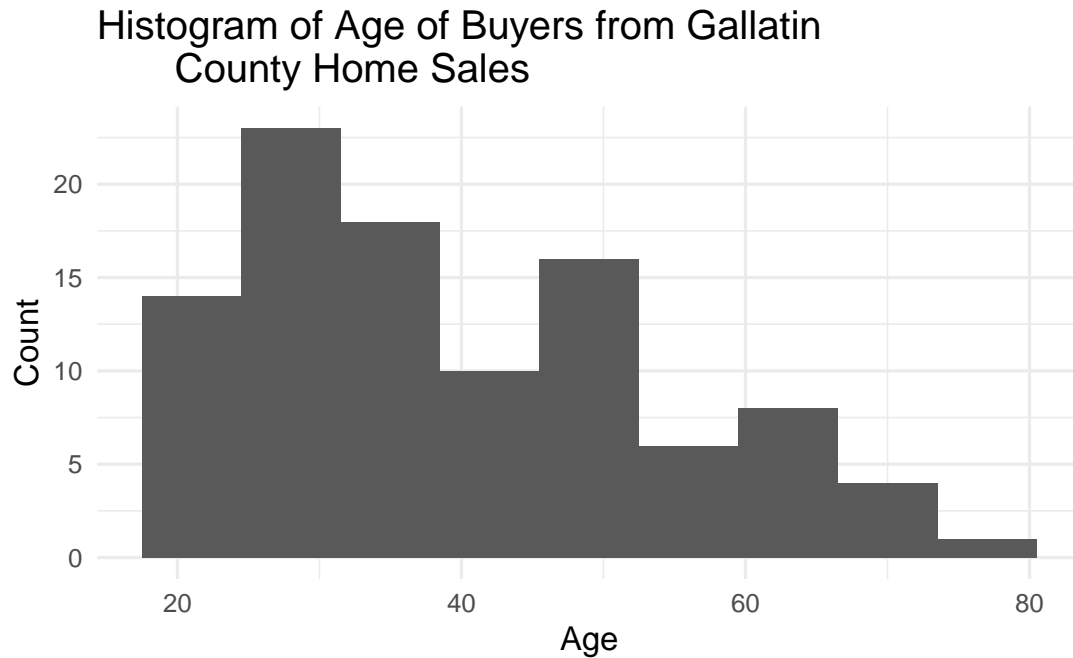
```
moving %>% # Data set piped into...
  ggplot(aes(y = Age, x = From)) + # Identify variables
  geom_boxplot() + # Tell it to make a box plot
  labs(title = "Side by side box plot of Age by State of Origin
of Buyers from Gallatin County Home Sales", # Title
       x = "State of Origin", # x-axis label
       y = "Age") # y-axis label
```



- Which state of origin had the oldest median age of buyers from sampled home sales?
- Which state of origin had the most variability in age of buyers from sampled home sales?
- Which state of origin had the most symmetric distribution of ages of buyers from sampled home sales?
- Which state of origin had outliers for the age of buyers from sampled home sales?

Robust statistics - Video 5.7

Let's review the summary statistics and histogram of age of buyers from sampled home sales.



```
#>   min  Q1 median    Q3  max  mean    sd  n missing
#>   20  28    36 49.25  77 39.77 14.35471 100      0
```

Notice that the _____ has been pulled in the direction of the _____.

- The _____ is a robust measure of center.
- The _____ is a robust measure of spread.
- Robust means not _____ by outliers.

When the distribution is symmetric use the _____ as the measure of center and the _____ as the measure of spread.

When the distribution is skewed with outliers use the _____ as the measure of center and the _____ as the measure of spread.

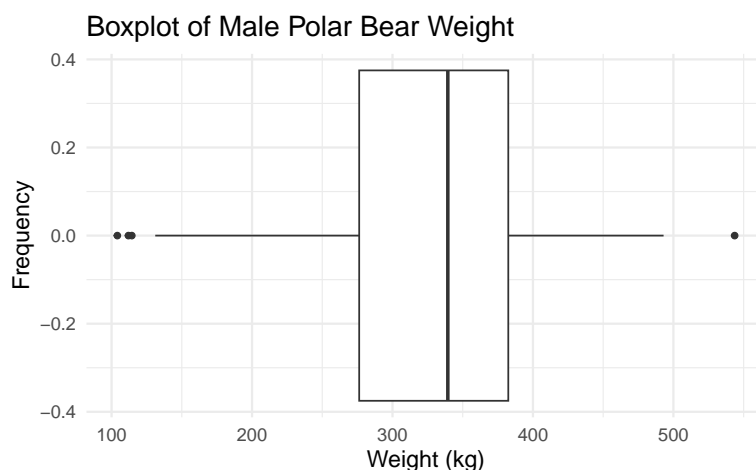
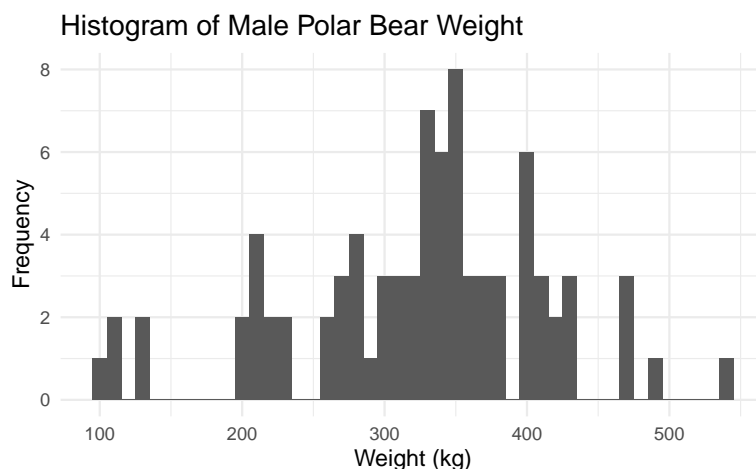
1.2.2 Video notes single quantitative variable inference

Example: What is the average weight of adult male polar bears? The weight was measured on a representative sample of 83 male polar bears from the Southern Beaufort Sea.

```
pb <- read.csv("https://math.montana.edu/courses/s216/data/polarbear.csv")
```

Plots of the data:

```
pb %>%  
  ggplot(aes(x = Weight)) + # Name variable to plot  
  geom_histogram(binwidth = 10) + # Create histogram with specified binwidth  
  labs(title = "Histogram of Male Polar Bear Weight", # Title for plot  
        x = "Weight (kg)", # Label for x axis  
        y = "Frequency") # Label for y axis  
  
pb %>% # Data set piped into...  
  ggplot(aes(x = Weight)) + # Name variable to plot  
  geom_boxplot() + # Create boxplot  
  labs(title = "Boxplot of Male Polar Bear Weight", # Title for plot  
        x = "Weight (kg)", # Label for x axis  
        y = "Frequency") # Label for y axis
```



Summary Statistics:

```
pb %>%  
  summarise(favstats(Weight)) #Gives the summary statistics  
#>      min      Q1 median      Q3      max      mean      sd  n missing  
#> 1 104.1 276.3 339.4 382.45 543.6 324.5988 88.32615 83      0
```

Hypothesis testing

- Hypotheses are always written about the _____. For a single mean we will use the notation _____.

Null Hypothesis:

H_0 :

Alternative Hypothesis:

H_A :

- Direction of the alternative depends on the _____.

Simulation-based method

- Simulate many samples assuming $H_0 : \mu = \mu_0$
 - Shift the data by the difference between μ_0 and \bar{x}
 - Sample with replacement n times from the shifted data
 - Plot the simulated shifted sample mean from each simulation
 - Repeat 1000 times (simulations) to create the null distribution
 - Find the proportion of simulations at least as extreme as \bar{x}

Example: Is there evidence that male polar bears weigh less than 370kg (previously recorded measure), on average? The weight was measured on a representative sample of 83 male polar bears from the Southern Beaufort Sea.

Hypotheses:

In notation:

H_0 :

H_A :

In words:

H_0 :

H_A :

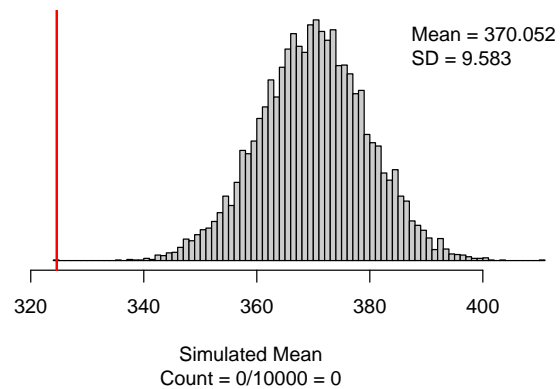
Reminder of summary statistics:

```
pb %>%  
  summarise(favstats(Weight)) #Gives the summary statistics  
#>   min    Q1 median    Q3   max    mean     sd  n missing  
#> 1 104.1 276.3 339.4 382.45 543.6 324.5988 88.32615 83      0
```

Find the difference:

$\mu_0 - \bar{x} =$

```
set.seed(216)  
one_mean_test(pb$Weight, #Enter the object name and variable  
  null_value = 370, #Enter null value for the study  
  summary_measure = "mean", #Can choose between mean or median  
  shift = 45.4, # Shift needed for bootstrap hypothesis test  
  as_extreme_as = 324.6, # Observed statistic  
  direction = "less", # Direction of alternative  
  number_repetitions = 10000) # Number of simulated samples for null distribution
```



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

Theory-based method

Conditions for inference using theory-based methods:

- Independence:
- Large enough sample size:

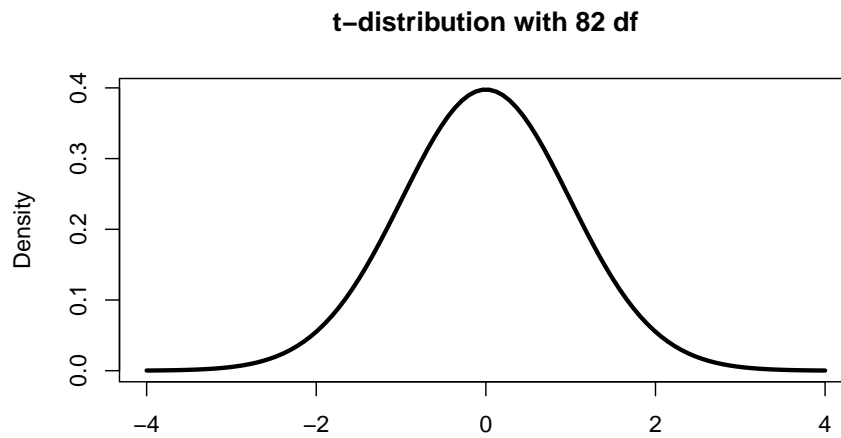
T - distribution

In the theoretical approach, we use the CLT to tell us that the distribution of sample means will be approximately normal, centered at the assumed true mean under H_0 and with standard deviation $\frac{\sigma}{\sqrt{n}}$.

$$\bar{x} \sim N(\mu_0, \frac{\sigma}{\sqrt{n}})$$

- Estimate the population standard deviation, σ , with the _____ standard deviation, _____.
- For a single quantitative variable we use the _____ - distribution with _____ degrees of freedom to approximate the sampling distribution.

The t^* multiplier is the value at the given percentile of the t-distribution with $n - 1$ degrees of freedom.

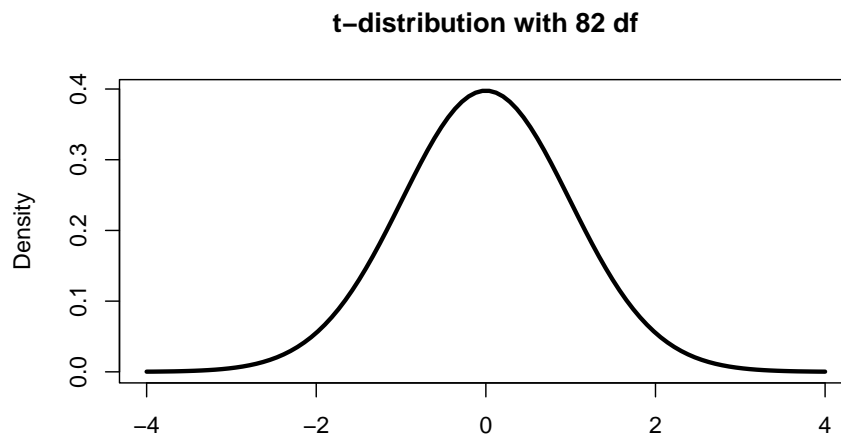


- Calculate the standardized statistic
- Find the area under the t-distribution with $n - 1$ df at least as extreme as the standardized statistic

Equation for the standard error of the sample mean:

Equation for the standardized sample mean:

Calculate the standardized sample mean weight of adult male polar bears:



Interpret the standardized sample mean weight:

To find the theory-based p-value:

```
pt(-4.683, df=82, lower.tail=TRUE)
#> [1] 5.531605e-06
```

1.2.3 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. What plots can be used to summarize quantitative data?
2. Which measure of center is robust to outliers?

1.3 Activity 11: Summarizing Quantitative Variables

1.3.1 Learning outcomes

- Identify and create appropriate summary statistics and plots given a data set or research question for quantitative data.
- Interpret the following summary statistics in context: median, lower quartile, upper quartile, standard deviation, interquartile range.

1.3.2 Terminology review

In today's activity, we will review summary measures and plots for quantitative variables. Some terms covered in this activity are:

- Two measures of center: mean, median
- Two measures of spread (variability): standard deviation, interquartile range (IQR)
- Plots of quantitative variables: dotplots, boxplots, histograms
- Given a plot or set of plots, describe and compare the distribution(s) of quantitative variables (center, variability, shape, outliers).

To review these concepts, see Chapter 5 in the textbook.

1.3.3 The Integrated Postsecondary Education Data System (IPEDS)

These data were collected on a subset of institutions that met the following selection criteria (Education Statistics 2018):

- Degree granting
- United States only
- Title IV participating
- Not for profit
- 2-year or 4-year or above
- Has full-time first-time undergraduates

Some of the variables collected and their descriptions are below. Note that several variables have missing values for some institutions (denoted by "NA").

Variable	Description
UnitID	Unique institution identifier
Name	Institution name
State	State abbreviation
Sector	whether public or private
LandGrant	Is this a land-grant institution (Yes/No)
Size	Institution size category based on total student enrolled for credit, Fall 2018: Under 1,000, 1,000\$-4,999, 5,000-9,999, 10,000-\$19,999, 20,000 and above
Cost_OutofState	Cost of attendance for full-time out-of-state undergraduate students
Cost_InState	Cost of attendance for full-time in-state undergraduate students
Retention	Retention rate is the percent of the undergraduate students that re-enroll in the next year
Graduation_Rate	6-year graduation rate for undergraduate students

Variable	Description
SATMath_75	75th percentile Math SAT score
ACT_75	75th percentile ACT score

Identifying Variables in a data set

Look through the provided chart showing the description of variables measured. The UnitID and Name are identifiers for each observational unit, *US degree granting institutions in 2018*.

1. Identify in the chart which variables collected on the US institutions are categorical (C) and which variables are quantitative (Q).

Summarizing quantitative variables

The `favstats()` function from the `mosaic` package gives the summary statistics for a quantitative variable. The R output below provides the summary statistics for the variable `Graduation_Rate`. The summary statistics provided are the two measures of center (mean and median) and two measures of spread (standard deviation and the quartile values to calculate the IQR) for IMDB score.

- Highlight and run lines 1 – 12 in the provided R script file to load the data set. Check that the summary statistics match the output given in the coursepack.
- Notice that the 2-year institutions were removed so the observational units for this study are **4-year higher education institutions**.

```
IPEDS <- read.csv("https://www.math.montana.edu/courses/s216/data/IPEDS_2018.csv")
IPEDS <- IPEDS %>%
  filter(Sector != "Public 2-year") # Filters the data set to remove Public 2-year
IPEDS <- IPEDS %>%
  filter(Sector != "Private 2-year") # Filters the data set to remove Private 2-year
IPEDS %>%
  summarize(favstats(Graduation_Rate))

#>   min Q1 median Q3 max      mean      sd    n missing
#> 1   0  38     53  67 100 52.48749 20.63192 1918      49
```

2. Report the values for the two measures of center (mean and median).
3. Calculate the interquartile range ($IQR = Q_3 - Q_1$) of Graduation Rates.
4. Report the value of the standard deviation and interpret this value in context of the problem.
5. Interpret the value of Q_3 in context of the study.

Displaying a single quantitative variable

There are three type of plots used to plot a single quantitative variable: a dotplot, a histogram or a boxplot. A dotplot of graduation rate would plot a dot for the graduation rate for each 4-year US higher education institution.

First, let's create a histogram of the variable `Graduation_Rate`.

- Enter the name of the variable in line 19 for `variable` in the R script file.
- Replace the word title for the plot in line 21 between the quotations with a descriptive title. **A title should include: type of plot, variable or variables plotted, and observational units.**
- Highlight and run lines 18 – ?? to create the histogram.

```
IPEDS %>% # Data set piped into...
ggplot(aes(x = xx)) + # Name variable to plot
  geom_histogram(binwidth = 10) + # Create histogram with specified binwidth
  labs(title = "Don't forget to title the plot!", # Title for plot
        x = "Graduation Rate", # Label for x axis
        y = "Frequency") # Label for y axis
```

Notice that the **bin width** for the histogram is 10. For example the first bin consists of the number of institutions in the data set with a graduation rate of 0 to 10%. It is important to note that a graduation rate on the boundary of a bin will fall into the bin above it; for example, 20 would be counted in the bin 20–30.

6. Which range of Graduation Rates have the highest frequency?

Next we will create a boxplot of the variable `Graduation_Rate`.

- Enter the name of the variable in line 19 for `variable` in the R script file.
- Highlight and run lines....

```
IPEDS %>% # Data set piped into...
ggplot(aes(x = variable)) + # Name variable to plot
  geom_boxplot() + # Create boxplot with specified binwidth
  labs(title = "Boxplot of Graduation Rates for 4-year Higher Education Institutions", # Title for plot
        x = "Graduation_Rate", # Label for x axis
        y = "") + # Remove y axis label
  theme(axis.text.y = element_blank(),
        axis.ticks.y = element_blank()) # Removes y-axis ticks
```

7. Sketch the boxplot created and identify the values of the 5-number summary (minimum value, Q1, median, Q3, maximum value) on the plot. Use the following formulas to find the invisible fence on both ends of the distribution. Draw a dotted line at the invisible fence to show how the outliers were found.

Lower Fence: values $\leq Q1 - 1.5 \times IQR$

Upper Fence: values $\geq Q3 + 1.5 \times IQR$

When describing plots of quantitative variables we discuss the shape (symmetric or skewed), the center (mean or median), spread (standard deviation or IQR), and if there are outliers present.

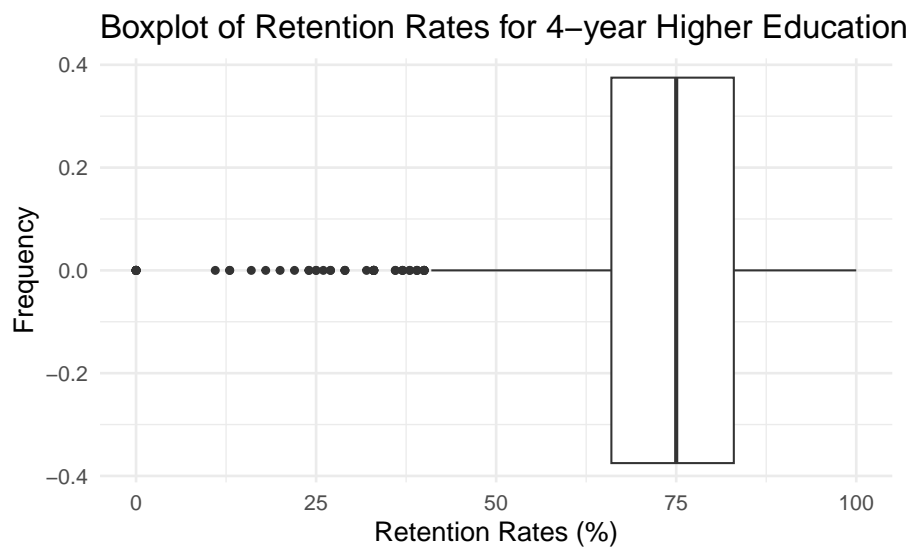
8. What is the shape of the distribution of graduation rates?
9. From which plot (histogram or boxplot) is it easier to determine the shape of the distribution?
10. From which plot is it easier to determine if there are outliers?

Robust Statistics

Let's examine how the presence of outliers affect the values of center and spread. For this part of the activity we will look at the variable retention rate in the IPEDS data set.

```
IPEDS %>% # Data set piped into...
  summarise(favstats(Retention))
#>   min Q1 median Q3 max   mean    sd   n missing
#> 1    0 66    75 83 100 73.8525 15.14323 1817    150

IPEDS %>% # Data set piped into...
  ggplot(aes(x = Retention)) + # Name variable to plot
  geom_boxplot() + # Create boxplot
  labs(title = "Boxplot of Retention Rates for 4-year Higher Education Institutions", # Title for plot
       x = "Retention Rates (%)", # Label for x axis
       y = "Frequency") # Label for y axis
#> Warning: Removed 150 rows containing non-finite outside the scale range
#> (`stat_boxplot()`).
```



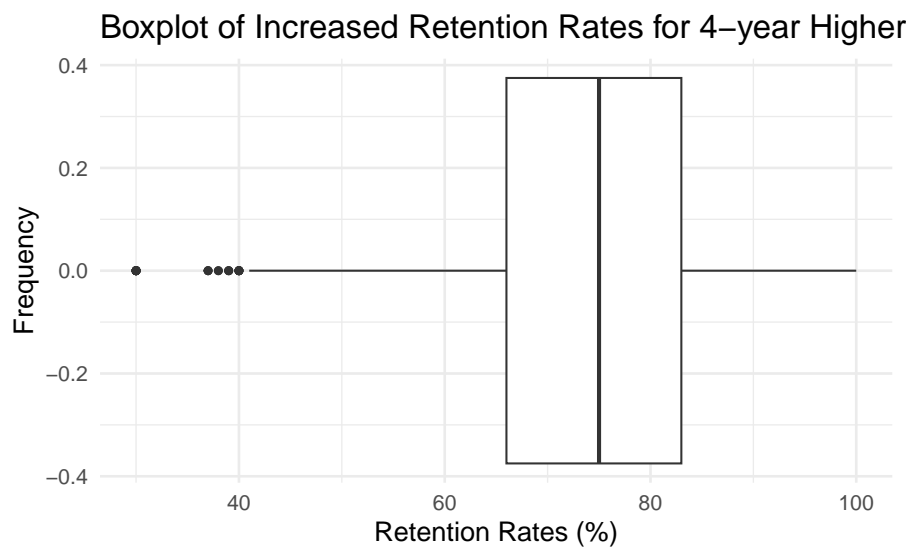
11. Report the two measures of center for these data.

12. Report the two measures of spread for these data.

To show the effect of outliers on the measures of center and spread, the smallest values of retention rate in the data set were increased by 30%. This variable is called `Retention_Inc`.

```
IPEDS %>% # Data set piped into...
  summarise(favstats(Retention_Inc))
#>   min Q1 median Q3 max    mean    sd   n missing
#> 1   30 66    75 83 100 74.49642 13.41255 1817    150

IPEDS %>% # Data set piped into...
  ggplot(aes(x = Retention_Inc)) + # Name variable to plot
  geom_boxplot() + # Create histogram
labs(title = "Boxplot of Increased Retention Rates for 4-year Higher Education Institutions", # Title for plot
     x = "Retention Rates (%)", # Label for x axis
     y = "Frequency") # Label for y axis
#> Warning: Removed 150 rows containing non-finite outside the scale range
#> (`stat_boxplot()`).
```



13. Report the two measures of center for this new data set.

14. Report the two measures of spread for this new data set.

15. Which measure of center is robust to outliers? Explain your answer.

16. Which measure of spread is robust to outliers? Explain your answer.

1.3.4 Take-home messages

1. Histograms, box plots, and dot plots can all be used to graphically display a single quantitative variable.
2. The box plot is created using the five number summary: minimum value, quartile 1, median, quartile 3, and maximum value. Values in the data set that are less than $Q_1 - 1.5 \times \text{IQR}$ and greater than $Q_3 + 1.5 \times \text{IQR}$ are considered outliers and are graphically represented by a dot outside of the whiskers on the box plot.
3. Data should be summarized numerically and displayed graphically to give us information about the study.
4. When comparing distributions of quantitative variables we look at the shape, center, spread, and for outliers. There are two measures of center: mean and the median and two measures of spread: standard deviation and the interquartile range, $\text{IQR} = Q_3 - Q_1$.

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 12: Hypothesis Testing of a Single Quantitative Variable

1.4.1 Learning outcomes

- Given a research question involving one quantitative variable, construct the null and alternative hypotheses in words and using appropriate statistical symbols.
- Investigate the process of creating a null distribution for one quantitative variable
- Find, evaluate, and interpret a p-value from the null distribution

1.4.2 Terminology review

In today's activity, we will simulation and theory-based methods to analyze a single quantitative variable. Some terms covered in this activity are:

- Null hypothesis
- Alternative hypothesis

To review these concepts, see Chapter 17 in the textbook.

1.4.3 College student sleep habits

According to the an article in *Sleep* (Watson 2015), experts recommend adults (>18) get at least 7 hours of sleep per night. A survey was sent to students in four sections of Stat 216 asking about their sleep habits. Is there evidence that sleep college students get less than the recommended 7 hours of sleep per night, on average?

Summarizing quantitative variables

- Download the R script file and data file for this activity
- Upload both files to the RStudio server and open the R script file
- Enter the name of the dataset for datasetname.csv
- Highlight and run lines 1–8 to load the data

```
sleep <- read.csv("datasetname.csv")
```

Ask a research question

1. Write the parameter of interest in context of the study.
2. Write the null hypothesis in words in context of the study.
3. Write the alternative hypothesis in notation.

Summarize and visualize the data

The `favstats()` function from the `mosaic` package gives the summary statistics for a quantitative variable.

- Enter the variable name, `SleepHours` for variable in line 13
- Highlight and run lines 12–13

```
sleep %>%  
  summarize(favstats(variable))
```

4. How far is each number of hours of sleep for a Stat 216 student from the mean number of hours of sleep, on average?

Create a boxplot of the variable `SleepHours`.

- Enter the name of the variable in line 19 for `variable` in the R script file.
- Enter a title in line 21 for the plot between the quotations
- Highlight and run lines 18 - 25

```
sleep %>% # Data set piped into...  
  ggplot(aes(x = variable)) + # Name variable to plot  
  geom_boxplot() + # Create boxplot with specified binwidth  
  labs(title = "Don't forget to title your plot!", # Title for plot  
        x = "Amount of sleep (hrs)", # Label for x axis  
        y = "") + # Remove y axis label  
  theme(axis.text.y = element_blank(),  
        axis.ticks.y = element_blank()) # Removes y-axis ticks
```

5. Describe the boxplot using the four characteristics of boxplots.

Simulation methods

To simulate the null distribution of sample means we will use a bootstrapping method. Recall that the null distribution must be created under the assumption that the null hypothesis is true. Therefore, before bootstrapping, we will need to *shift* each data point by the difference $\mu_0 - \bar{x}$. This will ensure that the mean of the shifted data is μ_0 (rather than the mean of the original data, \bar{x}), and that the simulated null distribution will be centered at the null value.

6. Calculate the difference $\mu_0 - \bar{x}$. Will we need to shift the data up or down?
 - Open the data set (`sleep_college`) in Excel
 - Create a new column labeled Shift
 - In the column, Shift, add the shifted value to each value in the column, `SleepHours`
 - Save the file and upload again to the RStudio server
 - Find the `favstats` of the variable, Shift

- Highlight and run lines 30–32

```
sleep <- read.csv("sleep_college.csv")
sleep %>%
  summarize(favstats(Shift))
```

7. Report the mean of the Shift variable. Why does it make sense that this value is the same as the null value?

8. Report the standard deviation of the Shift variable. How does this compare to the standard deviation for the variable SleepHours? Explain why these values are the same?

9. What inputs should be entered for each of the following to create the simulation?

- Null Value (What is the null value for the study?):
- Summary measure (“mean” or “median”):
- Shift (Difference between $\mu_0 - \bar{x}$):
- As extreme as (enter the value for the sample difference in proportions):
- Direction (“greater”, “less”, or “two-sided”):
- Number of repetitions:

Using the R script file for this activity...

- Enter your answers for question 9 in place of the xx’s to produce the null distribution with 10000 simulations
- Highlight and run lines 361–42.

```
one_mean_test(sleep$SleepHours, #Enter the object name and variable
  null_value = xx,
  summary_measure = "xx", #Can choose between mean or median
  shift = xx, #Difference between the null value and the sample mean
  as_extreme_as = xx, #Value of the summary statistic
  direction = "xx", #Specify direction of alternative hypothesis
  number_repetitions = 10000)
```

10. Interpret the p-value of the test in context of the problem.

11. Write a conclusion to the test in context of the problem.

1.4.4 Take-home messages

1. Histograms, box plots, and dot plots can all be used to graphically display a single quantitative variable.
2. The box plot is created using the five number summary: minimum value, quartile 1, median, quartile 3, and maximum value. Values in the data set that are less than $Q_1 - 1.5 \times \text{IQR}$ and greater than $Q_3 + 1.5 \times \text{IQR}$ are considered outliers and are graphically represented by a dot outside of the whiskers on the box plot.
3. Data should be summarized numerically and displayed graphically to give us information about the study.
4. When comparing distributions of quantitative variables we look at the shape, center, spread, and for outliers. There are two measures of center: mean and the median and two measures of spread: standard deviation and the interquartile range, $\text{IQR} = Q_3 - Q_1$.

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 Activity 13: Body Temperature

1.5.1 Learning outcomes

- Given a research question involving a quantitative variable, construct the null and alternative hypotheses in words and using appropriate statistical symbols.
- Describe and perform a theory-based hypothesis test for a single mean.
- Interpret and evaluate a p-value for a theory-based hypothesis test for a single mean.

1.5.2 Terminology review

In today's activity, we will analyze quantitative data using theory-based methods. Some terms covered in this activity are:

- Normality
- t -distribution
- Degrees of freedom
- T-score

To review these concepts, see Chapter 5 and ? in the textbook.

1.5.3 Body Temperature

It has long been reported that the mean body temperature of adults is 98.6°F. There have been a few articles that challenge this assertion. In 2018, a sample of 52 Stat 216 undergraduates, were asked to report their body temperature. Is there evidence that body temperatures of adults differ from the known temperature of 98.6°F?

Ask a research question

1. Write out the null hypothesis in proper notation for this study.
2. Write out the null hypothesis in words for this study.

In general, the sampling distribution for a sample mean, \bar{x} , based on a sample of size n from a population with a true mean μ and true standard deviation σ can be modeled using a Normal distribution when certain conditions are met.

Conditions for the sampling distribution of \bar{x} to follow an approximate Normal distribution:

- **Independence:** The sample's observations are independent. For paired data, that means each pairwise difference should be independent.
- **Normality:** The data should be approximately normal or the sample size should be large.
 - $n < 30$: If the sample size n is less than 30 and the distribution of the data is approximately normal with 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 in the data, 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.

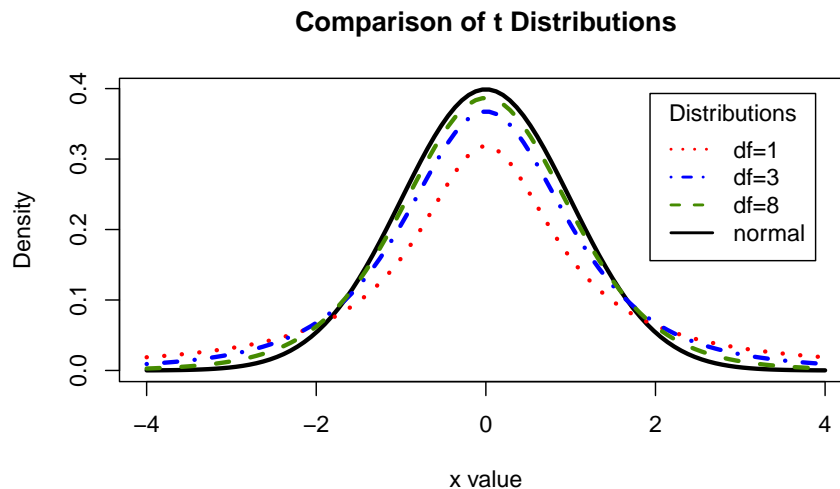


Figure 1.1: Comparison of the standard Normal vs t -distribution with various degrees of freedom

Like we saw in Chapter 5, we will not know the values of the parameters and must use the sample data to estimate them. Unlike with proportions, in which we only needed to estimate the population proportion, π , quantitative sample data must be used to estimate both a population mean μ and a population standard deviation σ . This additional uncertainty will require us to use a theoretical distribution that is just a bit wider than the Normal distribution. Enter the **t -distribution**!

As you can see from Figure 1.1, the t -distributions (dashed and dotted lines) are centered at 0 just like a standard Normal distribution (solid line), but are slightly wider. The variability of a t -distribution depends on its degrees of freedom, which is calculated from the sample size of a study. (For a single sample of n observations or paired differences, the degrees of freedom is equal to $n - 1$.) Recall from previous classes that larger sample sizes tend to result in narrower sampling distributions. We see that here as well. The larger the sample size, the larger the degrees of freedom, the narrower the t -distribution. (In fact, a t -distribution with infinite degrees of freedom actually IS the standard Normal distribution!)

Summarize and visualize the data

The following code is used to create a boxplot of the data.

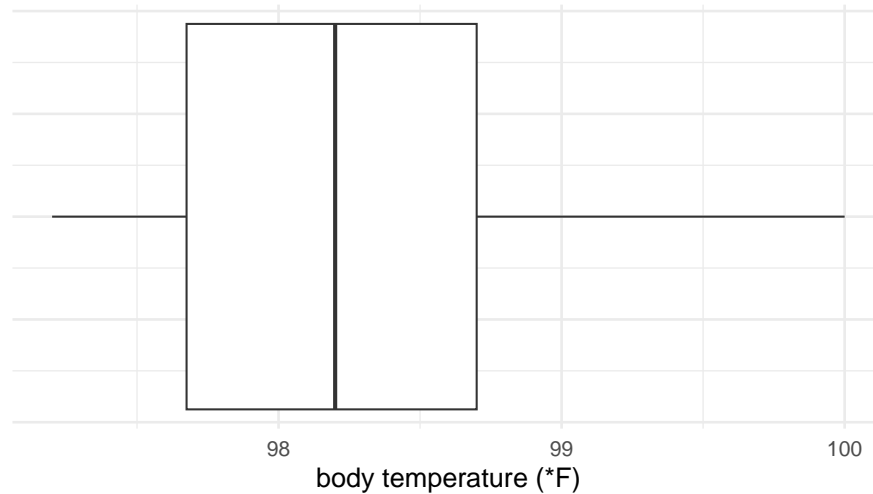
- Download the R script file upload to the R studio server.
- Open the R script file and highlight and run lines 1–14

```
bodytemp <- read.csv("https://math.montana.edu/courses/s216/data/normal_temperature.csv")
bodytemp %>%
  ggplot(aes(x = Temp))+
  geom_boxplot()+
  labs(title="Boxplot of Body Temperatures for Stat 216 Students",
       x = "body temperature (*F)") +
```



```
theme(axis.text.y = element_blank(),
      axis.ticks.y = element_blank()) # Removes y-axis ticks
```

Boxplot of Body Temperatures for Stat 216 Students



- Highlight and run lines 17 - 18 to get the summary statistics for the variable Temp.

```
bodytemp %>%
  summarise(favstats(Temp))
```

```
#>   min      Q1 median   Q3 max    mean      sd  n missing
#> 1  97.2  97.675   98.2  98.7 100 98.28462 0.6823789 52      0
```

Check theoretical conditions

3. Report the sample size of the study. Give appropriate notation.
4. Report the sample mean of the study. Give appropriate notation.
5. How do you know the independence condition is met for these data?
6. Is the normality condition met to use the theory-based methods for analysis? Explain your answer.

Use statistical inferential methods to draw inferences from the data

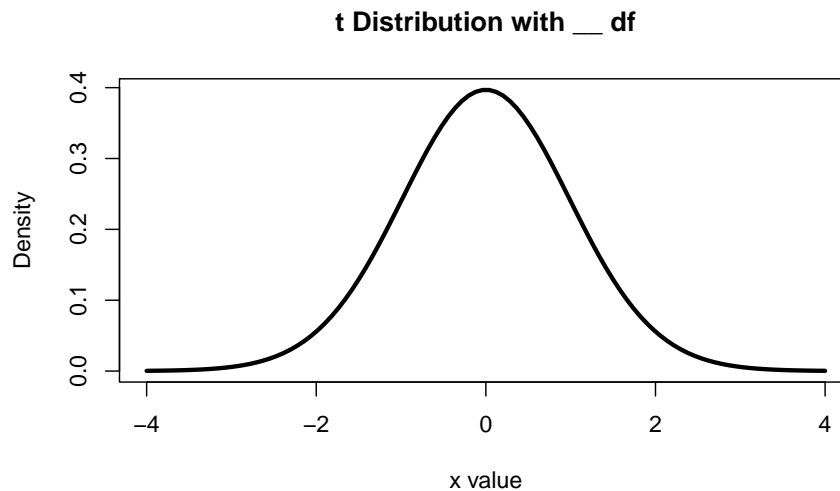
To find the standardized statistic for the mean we will use the following formula:

$$T = \frac{\bar{x} - \mu_0}{SE(\bar{x})},$$

where the standard error of the sample mean difference is:

$$SE(\bar{x}) = \frac{s}{\sqrt{n}}.$$

7. Calculate the standard error of the sample mean.
8. Interpret the standard error in context of the study.
9. Calculate the standardized mean.
10. We model a single mean with a t-distribution with $n - 1$ degrees of freedom. Calculate the degrees of freedom for this study.
11. Mark the value of the standardized statistic on the t-distribution and illustrate how the p-value is found.



To find the p-value for the theory-based test:

- Enter the value for the standardized statistic for `xx` in the `pt` function.
- Enter the `df` for `yy` in the `pt` function.
- Highlight and run line 24

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

12. What does this p-value mean, in the context of the study? Hint: it is the probability of what...assuming what?
13. Write a conclusion to the test in context of the study.
14. Can we generalize the results of the study be generalized to all adults? Explain your answer.

1.5.4 Take-home messages

1. In order to use theory-based methods for dependent groups (paired data), the independent observational units and normality conditions must be met.
2. A T-score is compared to a t -distribution with $n - 1$ 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 $n - 1$ df.

1.5.5 Additional notes

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

Confidence Intervals for a Single Quantitative Variable

2.1 Vocabulary Review and Key Topics

Review the Golden Ticket posted in the resources at the end of the coursepack for a summary of a single quantitative variable.

2.1.1 Key topics

Module 7 will cover creating confidence intervals using both simulation-based and theory-based methods. Additionally, we learn about types of errors and power in hypothesis testing.

Simulation Confidence Interval

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

```
one_mean_CI(object$variable, #Enter the name of the variable
             summary_measure = "mean", #choose the mean or median
             number_repetitions = 10000, # Number of simulations
             confidence_level = xx)
```

- Interpretation of the confidence interval is very similar as for a single proportion only the context and summary measure has changed
 - To write in context include:
 - * How confident you are (e.g., 90%, 95%, 98%, 99%)
 - * Parameter of interest
 - * Calculated interval

Theory-based Confidence Interval

- Calculation of the confidence interval for a sample mean

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

- 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 $n - 1$ df (enter for yy)
 - Enter the percentile for the given confidence level

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

Errors and Power

- **Significance level (α):** a given cut-off value that we compare the p-value to determine a decision of a test.
- **Decisions:**
 - If the p-value is less than the significance level, we make the decision to reject the null hypothesis
 - If the p-value is greater than the significance level, we make the decision to fail to reject the null hypothesis
- **Type I Error:** concluding there is evidence to reject the null hypothesis, when the null is actually true.
- **Type II Error:** concluding there is no evidence to reject the null hypothesis, when the null is actually false.
- **Power:** probability of concluding there is evidence to reject the null hypothesis, when the null is actually false

2.2 Video Notes: Theory-based Inference for a single quantitative variable

Read Chapters 5 and 17 in the course textbook. Use the following videos to complete the video notes for Module 7.

2.2.1 Course Videos

- 17.1
- 17.3 Theory Intervals

2.2.2 Single quantitative variable

- Reminder: review summary measures and plots discussed in the Module 6 material and Chapter 5 of the textbook.
- The summary measure for a single quantitative variable is the _____.

Notation:

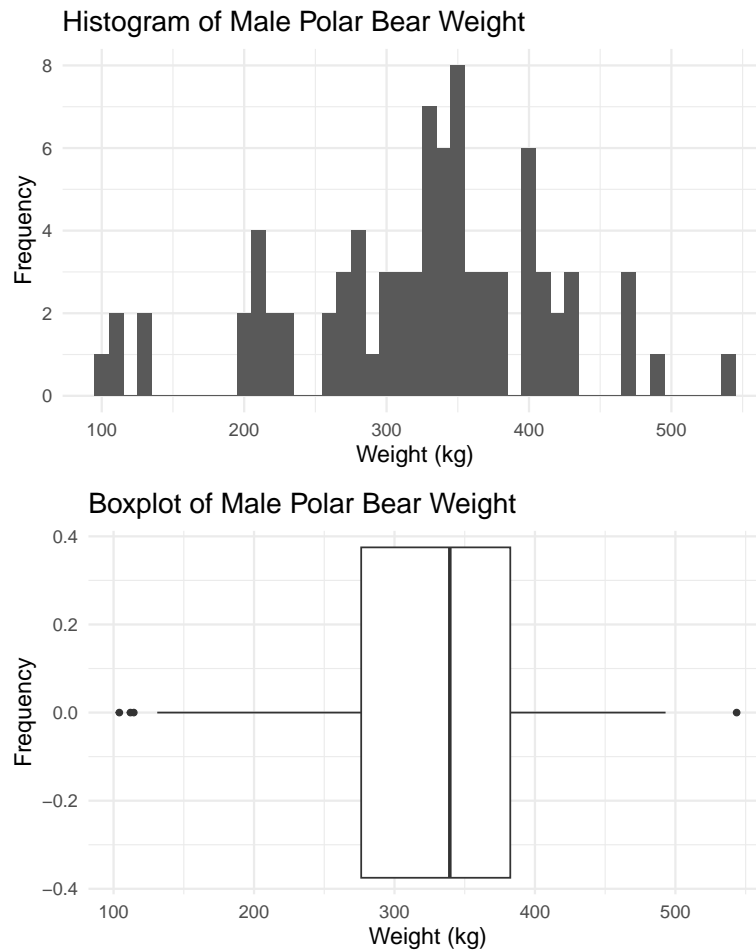
- Population mean:
- Population standard deviation:
- Sample mean:
- Sample standard deviation:
- Sample size:

Example: What is the average weight of adult male polar bears? The weight was measured on a representative sample of 83 male polar bears from the Southern Beaufort Sea.

```
pb <- read.csv("https://math.montana.edu/courses/s216/data/polarbear.csv")
```

Plots of the data:

```
pb %>%  
  ggplot(aes(x = Weight)) + # Name variable to plot  
  geom_histogram(binwidth = 10) + # Create histogram with specified binwidth  
  labs(title = "Histogram of Male Polar Bear Weight", # Title for plot  
        x = "Weight (kg)", # Label for x axis  
        y = "Frequency") # Label for y axis  
  
pb %>% # Data set piped into...  
  ggplot(aes(x = Weight)) + # Name variable to plot  
  geom_boxplot() + # Create boxplot  
  labs(title = "Boxplot of Male Polar Bear Weight", # Title for plot  
        x = "Weight (kg)", # Label for x axis  
        y = "Frequency") # Label for y axis
```



Summary Statistics:

```
pb %>%
  summarise(favstats(Weight)) #Gives the summary statistics
#>   min    Q1 median    Q3   max   mean    sd  n missing
#> 1 104.1 276.3 339.4 382.45 543.6 324.5988 88.32615 83      0
```

Confidence interval

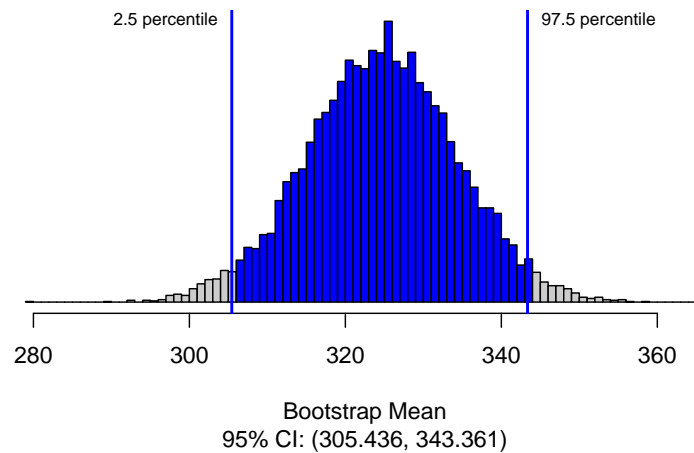
Simulation-based method

- Label cards with the values from the data set
- Sample with replacement (bootstrap) from the original sample n times
- Plot the simulated sample mean on the bootstrap distribution
- Repeat at least 1000 times (simulations)
- Find the cut-offs for the middle X% (confidence level) in a bootstrap distribution.
- ie. 95% CI = (2.5th percentile, 97.5th percentile)

Conditions for inference for a single mean:

- Independence:

```
set.seed(216)
one_mean_CI(pb$Weight,
  summary_measure = "mean",
  number_repetitions = 10000,
  confidence_level = 0.95)
```



The confidence interval estimates the _____ of _____.

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

- Calculate the interval centered at the sample statistic
statistic \pm margin of error

Conditions for inference using theory-based methods:

- Independence:
- Large enough sample size:

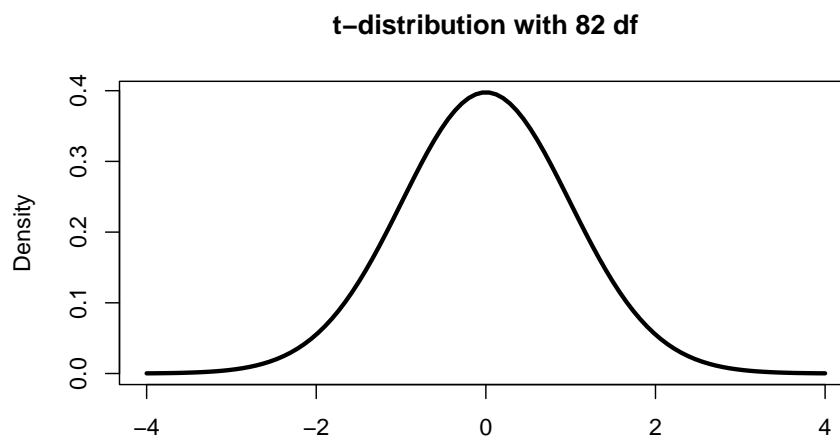
T - distribution

In the theoretical approach, we use the CLT to tell us that the distribution of sample means will be approximately normal, centered at the assumed true mean under H_0 and with standard deviation $\frac{\sigma}{\sqrt{n}}$.

$$\bar{x} \sim N(\mu_0, \frac{\sigma}{\sqrt{n}})$$

- Estimate the population standard deviation, σ , with the _____ standard deviation, _____.
- For a single quantitative variable we use the _____ - distribution with _____ degrees of freedom to approximate the sampling distribution.

The t^* multiplier is the value at the given percentile of the t-distribution with $n - 1$ degrees of freedom.



To find the t^* multiplier for a 95% confidence interval:

```
qt(0.975, df = 82)  
#> [1] 1.989319
```

Calculation of the confidence interval for the true mean weight of polar bears from the Southern Beaufort Sea:

2.3 Activity 14: Danceability of Songs

2.3.1 Learning outcomes

- Use simulation methods to find a confidence interval for a single mean
- Use theory-based methods to find a confidence interval for a single mean.
- Interpret a confidence interval for a single mean.
- Use a confidence interval to determine the conclusion of a hypothesis test.

2.3.2 Terminology review

In today's activity, we will estimate the parameter of interest using simulation and theory-based methods. Some terms covered in this activity are:

- Bootstrap distribution
- t -distribution
- Degrees of freedom
- T-score

To review these concepts, see Chapter 15 in the textbook.

2.3.3 Danceability

Spotify created a list of the top songs around the world for the past 10 years and several different audio features of those songs. One of the variables measured on these songs is Danceability. Danceability measures how easy it is to dance to a song; the higher the point value the easier it is to dance to the song. Estimate the average danceability of top songs from Spotify.

- Download the R script file for this activity from D2L and upload to the RStudio server
- Open the R script file, highlight and run

```
#>   min Q1 median Q3 max      mean      sd    n missing
#> 1    0  57     66  73  97 64.37977 13.37872 603         0

songs %>% # Data set piped into...
  ggplot(aes(x = Danceability)) + # Name variable to plot
  geom_boxplot() + # Create boxplot with specified binwidth
  labs(title = "Boxplot of Danceability Score for Top Spotify Songs", # Title for plot
       x = "danceability score (points)", # Label for x axis
       y = "") + # Remove y axis label
  theme(axis.text.y = element_blank(),
        axis.ticks.y = element_blank()) # Removes y-axis ticks
```

Summarizing quantitative variables

1. Describe the boxplot of danceability of top songs over the past 10 years on Spotify.

2. Write the parameter of interest in context of the study.

Simulation methods to create a confidence interval

Unlike creation of the null, the bootstrap distribution is found by sampling with replacement from the original sample.

- Write the original values for the variable on the cards
- Sample with replacement n times
- Plot the mean from each resampled sample on the distribution

Use the provided R script file to find a 95% confidence interval

- Enter the name of the variable for variable
- Enter the appropriate confidence interval
- Highlight and run lines 22–25

```
one_mean_CI(songs$variable, #Enter the name of the variable
             summary_measure = "mean", #choose the mean or median
             number_repetitions = 10000, # Number of simulations
             confidence_level = xx)
```

3. Report the 95% confidence interval for the parameter of interest.

Theory-based methods to create a confidence interval

- **Conditions for the sampling distribution of \bar{x} to follow an approximate normal distribution:**
 - **Independence:** The sample's observations are independent, e.g., are from a simple random sample. (*Remember:* This also must be true to use simulation methods!)
 - **Large enough sample size: Normality Condition:** The sample observations come from a normally distributed population. To check use the the following rules of thumb:
 - * $n < 30$: The distribution of the sample must be approximately normal with no outliers
 - * $30 \geq n < 100$: We can relax the condition a little; the distribution of the sample must have no extreme outliers or skewness
 - * $n > 100$: Can assume the sampling distribution of \bar{x} is nearly normal, even if the underlying distribution of individual observational is not

Next we will calculate a theory-based confidence interval. To calculate a theory-based confidence interval for the a single mean, use the following formula:

$$\bar{x} \pm t^* \times SE(\bar{x}).$$

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

- Enter the appropriate percentile in the R code to find the multiplier for a 95% confidence interval.
- Enter the df for yy. *The degrees of freedom for a single mean is $n - 1$*
- Highlight and run line 31

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

4. Mark on the t-distribution found below the values of $\pm t^*$. Draw a line at each multiplier and write the percentiles used to find each.

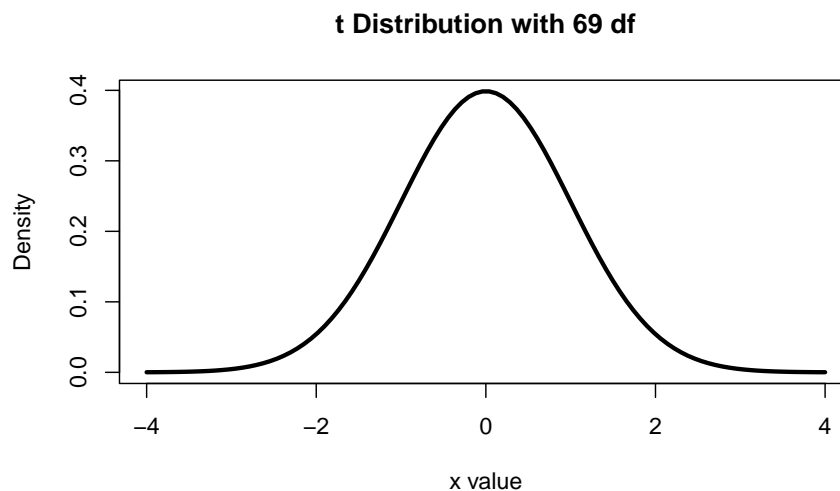


Figure 2.1: t-distribution with 602 degrees of freedom

5. Calculate the margin of error for the true mean using theory-based methods.
6. Calculate the confidence interval for the true mean using theory-based methods.
7. Interpret the confidence interval in context of the study.
8. Explain why the CI with theory-based methods is similar to the simulation CI.

2.3.4 Take-home messages

1. In order to use theory-based methods for a single mean, the independent observational units and normality conditions must be met.
2. The simulation based confidence interval and theory-based confidence interval should be similar if the normality condition is met.
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 $n - 1$ df.

2.3.5 Additional notes

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

2.4 Activity 15: Errors and Power

2.4.1 Learning outcomes

- Explain Type I and Type 2 Errors in the context of a study.
- Explain the power of a test in the context of a study.
- Understand how changes in sample size, significance level, and the difference between the null value and the parameter value impact the power of a test.
- Understand how significance level impacts the probability of a Type 1 Error.
- Understand the relationship between the probability of a Type 2 Error and power.
- Be able to distinguish between practical importance and statistical significance.

2.4.2 Terminology review

In this activity, we will examine the possible errors that can be made based on the decision in a hypothesis test as well as factors influencing the power of the test. Some terms covered in this activity are:

- Significance level
- Type 1 Error
- Type 2 Error
- Power

To review these concepts, see Chapter 12 in the textbook.

2.4.3 College Textbook Cost

A college student spends on average \$280 on textbooks per year. Many universities have started using open-source resources to help defray the cost of textbooks. One such university is hoping to show they have successfully reduced costs by \$100, on average.

1. Write the parameter of interest (μ) in words, in the context of this problem.
2. Use proper notation to write the null and alternative hypothesis the university would need to test in order to check their claim.

After determining hypotheses and prior to collecting data, researchers should set a **significance level** for a hypothesis test. The significance level, represented by α and most commonly 0.01, 0.05, or 0.10, is a cut-off for determining whether a p-value is small or not. The *smaller* the p-value, the *stronger* the evidence against the null hypothesis, so a p-value that is smaller than or equal to the significance level is strong enough evidence to *reject the null hypothesis*. Similarly, the *larger* the p-value, the *weaker* the evidence against the null hypothesis, so a p-value that is larger than the significance level does not provide enough evidence against the null hypothesis and the researcher would *fail to reject the null hypothesis*. Rejecting the null hypothesis or failing to reject the null hypothesis are the two **decisions** that can be made based on the data collected.

As you have already learned in this course, sample size of a study is extremely important. Often times, researchers will conduct what is called a power analysis to determine the appropriate sample size based on the goals of

their research, including a desired **power** of their test. Power is the probability of correctly rejecting the null hypothesis, or the probability of the data providing strong evidence against the null hypothesis *when the null hypothesis is false*.

The remainder of this activity will be spent investigating how different factors influence the power of a test, after which you will complete a power analysis for this physical therapy company.

- Navigate to <https://istats.shinyapps.io/power/>.
- Choose the tab **Population Mean**
- Use the scale under “Null Hypothesis value μ_0 ” to change the value to your null value from question 2. *Note we will convert this to a scale \$100 dollars. In other words, use the null value of 2.8.
- Change the “Alternative Hypothesis” to the direction you wrote in question 2.
- Leave all boxes un-checked.
- Set the “True value of μ ” to 2.8 as well
- Do not change the scales for “Sample size n” or “Type I Error α ”

The red distribution you see is the scaled-Normal distribution representing the null distribution for this hypothesis test, if the sample size was 30 and the significance level was 0.05. This means the red distribution is showing the probability of each possible sample mean of college students who spent \$280 on textbooks per year (\bar{x}) if we assume the null hypothesis is true.

3. Based off this distribution and your alternative hypothesis, give one possible sample mean which you think would lead to rejecting the null hypothesis. Explain how you decided on your value.
4. Check the box for “Show Critical Value(s) and Rejection Region(s)”. You will now see a vertical line on the plot indicating the *minimum* sample mean which would lead to reject the null hypothesis. What is this value?
5. Notice that there are some sample means under the red line (when the null hypothesis is true) which would lead us to reject the null hypothesis. Give the range of sample means which would lead to rejecting the null hypothesis when the null hypothesis is true? What is the statistical name for this mistake?

Check the “Type I Error” box under **Display**. This should verify (or correct) your answer to question 5! The area shaded in red represents the probability of making a **Type 1 Error** in our hypothesis test. Recall that a Type 1 Error is when we reject the null hypothesis even though the null hypothesis is true. To reject the null hypothesis, the p-value, which was found assuming the null hypothesis is true, must be less than or equal to the significance level. Therefore the significance level is the maximum probability of rejecting the null hypothesis when the null hypothesis is true, so the significance level IS the probability of making a Type 1 Error in a hypothesis test!

6. Based on the current applet settings, What percent of the null distribution is shaded red (what is the probability of making a Type 1 Error)?

Let’s say this university believes their program can reduce the cost of textbooks for college students by \$100 per year. In the applet, set the scale under “True value of μ ” to 1.8.

7. Where is the blue distribution centered?

The blue distribution that appears represents what the university believes, that \$180 (not \$280) is the true mean textbook cost for college students at this university. This blue distribution represents the idea that the **null hypothesis is false**.

8. Consider the definition of power provided earlier in this lab. Do you believe the power of the test will be an area within the blue distribution or red distribution? How do you know? What about the probability of making a Type 2 Error?

- Check the “Type II Error” and “Power” boxes under **Display**. This should verify (or correct) your answers to question 8! The area shaded in blue represents the probability of making a **Type 2 Error** in our hypothesis test (failing to reject the null hypothesis even though the null hypothesis is false). The area shaded in green represents the power of the test. Notice that the Type 1 and Type 2 Error rates and the power of the test are provided above the distribution.
9. Complete the following equation: Power + Type 2 Error Rate = . Explain why that equation makes sense. *Hint: Consider what power and Type 2 Error are conditional on.*

Now let’s investigate how changes in different factors influence the power of a test.

10. Using the same sample size and significance level, change the “True value of μ ” to see the effect on Power.

True value of p	2.0	1.5	1.0	0.05
Power				

11. What is changing about the simulated distributions pictured as you change the “True value of μ ”?

12. How does increasing the distance between the null and believed true mean affect the power of the test?

13. Using the same significance level, set the “True value of μ ” to 1.8 and change the sample size to see the effect on Power.

Sample Size	20	40	50	60	80
Power					

14. What is changing about the simulated distributions pictured as you change the sample size?

15. How does increasing the sample size affect the power of the test?

16. Using the same “True value of μ ”, set the sample size to 30 and change the “Type I Error α ” to see the effect on Power.

Type I Error α	0.01	0.03	0.05	0.10	0.15
Power					

17. What is changing about the simulated distributions pictured as you change the significance level?
18. How does increasing the significance level affect the power of the test?
19. Complete the power analysis for this university. The university believes they can reduce the cost of textbooks for their students by \$100. They want to limit the probability of a type 1 error to 10% and the probability of a type 2 error to 15%. What is the minimum number of students the university will need to collect data from in order to meet these goals? Use the applet to answer this question, then download your image created and upload the file to Gradescope.
20. Based on the goals outlined in question 19, which mistake below is the university more concerned about? In other words, which error were the researchers trying to minimize. Explain your answer.
- Not being able to show their textbook cost is lower, on average, when their textbook cost really is lower.
 - Advertising their textbook cost is lower, on average, even though it is not.

2.4.4 Take-home messages

1. There is a possibility of Type I Error when we make the decision to reject the null hypothesis. Type I Error - reject the null hypothesis when the null hypothesis is true.
2. There is a possibility of Type II Error when we make the decision to fail to reject the null hypothesis. Type II Error - fail to reject the null hypothesis when the null hypothesis is false.
3. Increasing the sample size will increase the power of the test.

2.4.5 Additional notes

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

2.5 Module 6 and 7 Lab: Arsenic

2.5.1 Learning outcomes

- Given a research question involving one quantitative variable, construct the null and alternative hypotheses in words and using appropriate statistical symbols.
- Investigate the process of creating a null distribution for one quantitative variable
- Find, evaluate, and interpret a p-value from the null distribution
- Use simulation methods to find a confidence interval for a single mean
- Interpret a confidence interval for a single mean.
- Use a confidence interval to determine the conclusion of a hypothesis test.

2.5.2 Arsenic

Scientists have devised a new way to measure a person's level of arsenic poisoning by examining toenail clippings. Scientists measured the arsenic levels (in parts per million or ppm) in toenail clippings from 19 randomly selected individuals with private wells in New Hampshire (data in the table below). An arsenic level greater than 0.150 ppm is considered hazardous. Is there evidence the ground water in New Hampshire has hazardous levels of arsenic concentration (as seen in the arsenic levels of New Hampshire residents)? How high is the arsenic concentration for New Hampshire residents with a private well?

1. What does μ represent in the context of this study?
 2. Notice that there are two research questions for this study. Identify which research question is best answered by finding a confidence interval and which is best answered by completing a hypothesis test?
 3. Write out the null hypothesis in proper notation for this study.
 4. What sign ($<$, $>$, or \neq) would you use in the alternative hypothesis for this study? Explain your choice.
- Upload and open the R script file for Week 12 lab.
 - Upload and import the csv file, **arsenic**.
 - Enter the name of the data set (see the environment tab) for `datasetname` in the R script file in line 8.
 - Highlight and run lines 1–9 to load the data and create a plot of the data.
 - **Upload a screenshot of your plot to Gradescope**

```
water <- read.csv("data/arsenic.csv")
water %>%
  summarise(favstats(level_arsenic))
```

```

water %>% # Data set piped into...
  ggplot(aes(x = variable)) + # Name variable to plot
  geom_boxplot() + # Create boxplot with specified binwidth
  labs(title = "Don't forget to title the plot!", # Title for plot
        x = "Enter an x-axis label! Don't forget the units!", # Label for x axis
        y = "") + # Remove y axis label
  theme(axis.text.y = element_blank(),
        axis.ticks.y = element_blank()) # Removes y-axis ticks

```

5. Based on the plot, does there appear to be some evidence in favor of the alternative hypothesis? How do you know?
6. Interpret the value of Q_3 in context of the study.
7. What is the value of \bar{x} ? What is the sample size?
8. How far, on average, is each arsenic level from the mean arsenic level? What is the appropriate notation for this value?

Use statistical inferential methods to draw inferences from the data

9. Using the provided graphs and summary statistics, determine if both theory-based methods and simulation methods could be used to analyze the data. Explain your reasoning.

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 the average arsenic levels are not hazardous.

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

10. Simulate a null distribution and compute the p-value, using the R script file for this lab.

```
one_mean_test(water$level_arsenic, #Enter the name of the variable
  null_value = 0.150, #Enter the name of the null value
  summary_measure = "mean", #Choose mean or median to test
  shift = -0.122, # Shift needed for bootstrap hypothesis test
  as_extreme_as = 0.272, # Observed statistic
  direction = "greater", # Direction of alternative
  number_repetitions = 10000) # Number of simulated samples for null distribution
```

Sketch the null distribution created using the `one_mean_test` code.

Communicate the results and answer the research question

11. Report the p-value. Based off of this p-value and a 1% significance level, what decision would you make about the null hypothesis? What potential error might you be making based on that decision?
12. Do you expect the 98% confidence interval to contain the null value of zero? Explain.

Confidence interval

We will use the `one_mean_CI()` function in R (in the `catstats` package) to simulate the bootstrap distribution of sample mean differences and calculate a confidence interval.

13. Using bootstrapping and the provided R script file, find a 98% confidence interval. Fill in the missing values/numbers in the `one_mean_CI()` function to create the 98% confidence interval. Highlight and run lines 37–40.

```
one_mean_CI(data = water$level_variable, # Enter vector of differences
  summary_measure = "mean", # Not needed when entering vector of differences
  number_repetitions = 10000, # Number of bootstrap samples for CI
  confidence_level = xx) # Confidence level in decimal form
```

Report the 98% confidence interval in interval notation.

14. Write a paragraph summarizing the results of the study. **Upload a copy of your group's paragraph to Gradescope.** Be sure to describe:
- Summary statistic and interpretation
 - Summary measure (in context)
 - Value of the statistic
 - Order of subtraction when comparing two groups
 - 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
 - To what group of observational units do the results apply (target population or observational units similar to the sample)?
 - What type of inference is appropriate (causal or non-causal)?

Paragraph (continued):

Exploratory Data Analysis and Simulation-based Inference for Two Categorical Variables

3.1 Vocabulary Review and Key Topics

Review the Golden Ticket posted in the resources at the end of the coursepack for a summary of a two categorical variables.

3.1.1 Key topics

Module 8 will introduce exploratory data analysis and simulation-based inference for two categorical variables. We also explore study design and confounding variables.

Types of plots for two categorical variables

- **Segmented bar plot:** plots the conditional proportion of the response outcomes in each explanatory variable group
 - The plot shows no association between the variables, if the height of each segment is approximately the same in each group
- **Mosaic plot:** similar to the segmented bar plot but the sample size is reflected by the width of the bars

Summary measures

- **Difference in proportion:** calculation of the difference in two conditional proportions
 - Parameter notation: $\pi_1 - \pi_2$
 - Sample notation: $\hat{p}_1 - \hat{p}_2$
- **Relative risk:** the ratio of the conditional proportions
 - Relative Risk = $\frac{\hat{p}_1}{\hat{p}_2}$

Interpretation of relative risk:

- The risk of success in group 1 is relative risk times the risk of success in group 2
- Can also interpret as a percent increase or percent decrease in risk

–

$$(RR - 1) \times 100\%$$

- The risk of success in group 1 is xx% higher or lower than the risk of success in group 2
- Explanatory variable: the variable researchers think *may be* affecting the other variable.
- Response variable: the variable researchers think *may be* influenced by the other variable.
- Confounding variable:
 - associated with both the explanatory and the response variable
 - explains the association shown by the data


Study Design

- Observational study:
- Randomized experiment:


Scope of Inference Table:

Scope of Inference: If evidence of an association is found in our sample, what can be concluded?

Selection of cases	Study Type		
	Randomized experiment	Observational study	
Random sample (and no other sampling bias)	Causal relationship, and can generalize results to population.	Cannot conclude causal relationship, but can generalize results to population.	→ Inferences to population can be made
No random sample (or other sampling bias)	Causal relationship, but cannot generalize results to a population.	Cannot conclude causal relationship, and cannot generalize results to a population.	→ Can only generalize to those similar to the sample due to potential sampling bias



Can draw cause-and-effect conclusions



Can only discuss association due to potential confounding variables

- Conditions necessary to use simulation methods for inference for two categorical variables
 - There must be independence of observational units within groups and between groups

3.2 Video Notes: Inference for Two Categorical Variables using Simulation-based Methods

Read Sections 2.2 - 2.4, 15.1, 15.2 and Chapter 16 in the course textbook. Use the following videos to complete the video notes for Module 8.

3.2.1 Course Videos

- 2.2to2.4
- 15.1
- 15.2
- RelativeRisk

Observational studies, experiments, and scope of inference: Video 2.2to2.4

- Review
 - Explanatory variable: the variable researchers think *may be* affecting the other variable.
 - Response variable: the variable researchers think *may be* influenced by the other variable.
- Confounding variable:
 - associated with both the explanatory and the response variable
 - explains the association shown by the data

Example:

Study design

- Observational study:

- Experiment:

Principles of experimental design

- Control: hold other differences constant across groups
- Randomization: randomized experiment
- Replication: large sample size or repeat of study
- Blocking: group based on certain characteristics

Example: It is well known that humans have more difficulty differentiating between faces of people from different races than people within their own race. A 2018 study published in the *Journal of Experimental Psychology* (Levin 2000): *Human Perception and Performance* investigated a similar phenomenon with gender. In the study, volunteers were shown several pictures of strangers. Half the volunteers were randomly assigned to rate the attractiveness of the individuals pictured. The other half were told to rate the distinctiveness of the faces seen. Both groups were then shown a slideshow of faces (some that had been rated in the first part of the study, some that were new to the volunteer) and asked to determine if each face was old or new. Researchers found people were better able to recognize faces of their own gender when asked to rate the distinctiveness of the faces, compared to when asked to rate the attractiveness of the faces.

- What is the study design?

Example: In the Physician's Health Study ("Physician's Health Study," n.d.), male physicians participated in a study to determine whether taking a daily low-dose aspirin reduced the risk of heart attacks. The male physicians were randomly assigned to the treatment groups. After five years, 104 of the 11,037 male physicians taking a daily low-dose aspirin had experienced a heart attack while 189 of the 11,034 male physicians taking a placebo had experienced a heart attack.

- What is the study design?
- Assuming these data provide evidence that the low-dose aspirin group had a lower rate of heart attacks than the placebo group, is it valid for the researchers to conclude the lower rate of heart attacks was caused by the daily low-dose aspirin regimen?

Scope of Inference

1. How was the sample selected?
 - Random sample with no sampling bias:
 - Non-random sample with sampling bias:
2. What is the study design?
 - Randomized experiment:
 - Observational study:

Scope of Inference Table:

Scope of Inference: If evidence of an association is found in our sample, what can be concluded?

	Study Type	
Selection of cases	Randomized experiment	Observational study
Random sample (and no other sampling bias)	Causal relationship, and can generalize results to population.	Cannot conclude causal relationship, but can generalize results to population.
No random sample (or other sampling bias)	Causal relationship, but cannot generalize results to a population.	Cannot conclude causal relationship, and cannot generalize results to a population.

↓

Can draw cause-and-effect conclusions

↓

Can only discuss association due to potential confounding variables

→ Inferences to population can be made

→ Can only generalize to those similar to the sample due to potential sampling bias

Example: It is well known that humans have more difficulty differentiating between faces of people from different races than people within their own race. A 2018 study published in the Journal of Experimental Psychology (Levin 2000): Human Perception and Performance investigated a similar phenomenon with gender. In the study, volunteers were shown several pictures of strangers. Half the volunteers were randomly assigned to rate the attractiveness of the individuals pictured. The other half were told to rate the distinctiveness of the faces seen. Both groups were then shown a slideshow of faces (some that had been rated in the first part of the study, some that were new to the volunteer) and asked to determine if each face was old or new. Researchers found people were better able to recognize faces of their own gender when asked to rate the distinctiveness of the faces, compared to when asked to rate the attractiveness of the faces.

- What is the scope of inference for this study?

Two categorical variables - Video 15.1

- In this module, we will study inference for a _____ explanatory variable and a _____ response.
- The summary measure for two categorical variables is the _____ in _____.

Example: In a double-blind experiment (Weiss 1988) on 48 cocaine addicts hoping to overcome their addiction, half were randomly assigned to a drug called desipramine and the other half a placebo. The addicts were followed for 6 weeks to see whether they were still clean. Is desipramine more effective at helping cocaine addicts overcome their addiction than the placebo?

Observational units:

Explanatory variable:

Response variable:

Notation:

- Population proportion for group 1:
- Population proportion for group 2:
- Sample proportion for group 1:
- Sample proportion for group 2:
- Sample difference in proportions:
- Sample size for group 1:
- Sample size for group 2:

Hypothesis Testing

Conditions:

- Independence: the response for one observational unit will not influence 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 direction of the alternative hypothesis.

Write the null and alternative hypotheses for the cocaine study:

In notation:

H_0 :

H_A :

Summary statistics and plot

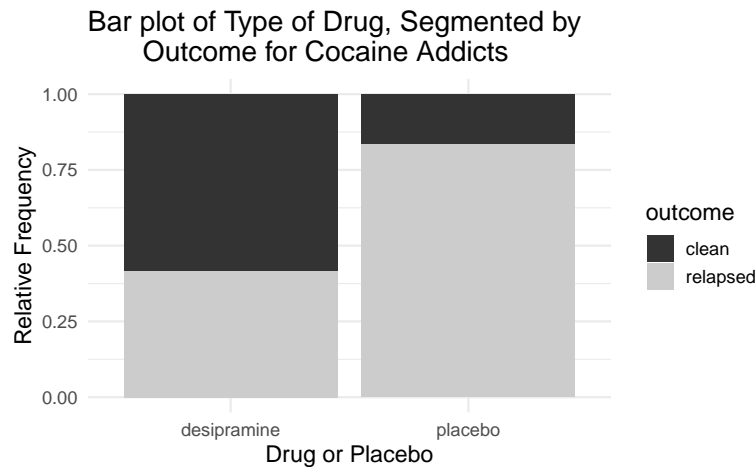
```
cocaine %>% group_by(drug) %>% count(outcome)
```

```
#> # A tibble: 4 x 3
#> # Groups:   drug [2]
#>   drug      outcome      n
#>   <chr>      <chr>   <int>
#> 1 desipramine clean     14
#> 2 desipramine relapsed  10
#> 3 placebo     clean      4
#> 4 placebo     relapsed  20
```

Summary statistic:

Interpretation:

```
cocaine%>%
  ggplot(aes(x = drug, fill = outcome))+
  geom_bar(stat = "count", position = "fill") +
  labs(title = "Bar plot of Type of Drug, Segmented by
    Outcome for Cocaine Addicts",
    y = "Relative Frequency",
    x = "Drug or Placebo") +
  scale_fill_grey()
```

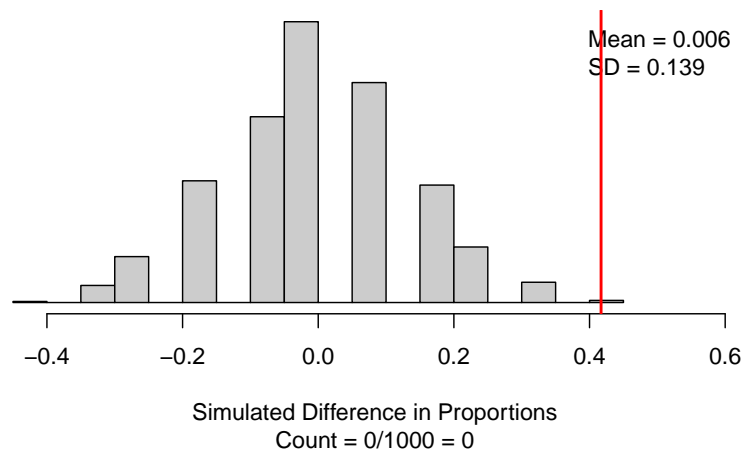


Is the independence condition met for simulation inference?

Simulation-based method

- Simulate many samples assuming $H_0 : \pi_1 = \pi_2$
 - 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 proportions from each simulation
 - Repeat 1000 times (simulations) to create the null distribution
 - Find the proportion of simulations at least as extreme as $\hat{p}_1 - \hat{p}_2$

```
set.seed(216)
two_proportion_test(formula = outcome~drug, # response ~ explanatory
  data = cocaine, # Name of data set
  first_in_subtraction = "desipramine", # Order of subtraction: enter the name of Group 1
  number_repetitions = 1000, # Always use a minimum of 1000 repetitions
  response_value_numerator = "clean", # Define which outcome is a success
  as_extreme_as = 0.417, # Calculated observed statistic (difference in sample proportions)
  direction="greater") # Alternative hypothesis direction ("greater", "less", "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 with scope of inference:

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

Confidence interval - Video 15.2

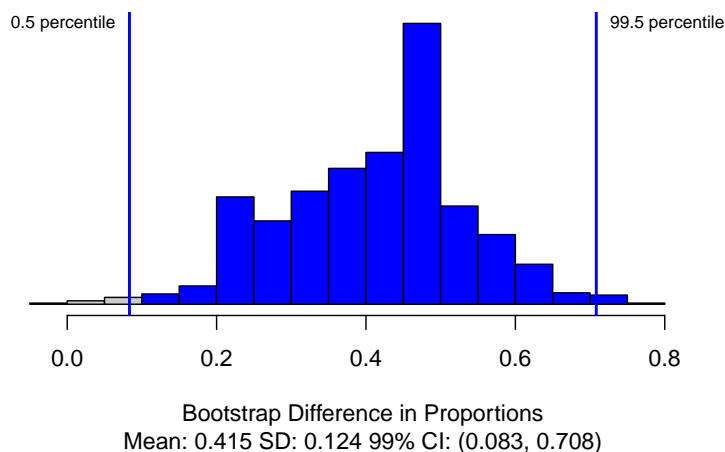
To estimate the difference in true proportion we will create a confidence interval.

Simulation-based method

- 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 proportions 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.

Returning to the cocaine example, we will estimate the difference in true proportion of cocaine addicts that stay clean for those on the desipramine and those on the placebo.

```
set.seed(216)
two_proportion_bootstrap_CI(formula = outcome ~ drug,
  data=cocaine, # Name of data set
  first_in_subtraction = "desipramine", # Order of subtraction: enter the name of Group 1
  response_value_numerator = "clean", # Define which outcome is a success
  number_repetitions = 1000, # Always use a minimum of 1000 repetitions
  confidence_level = 0.99) # Enter the level of confidence as a decimal
```



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

Relative Risk - Video Relative Risk

- Relative risk is the ratio of the risks in two different categories of an explanatory variable.

Relative Risk:

Example: In a study reported in the New England Journal of Medicine (Du Toit 2015), one-hundred fifty (150) children who had shown sensitivity to peanuts were randomized to receive a flour containing a peanut protein or a placebo flour for 2.5 years. At age 5 years, children were tested with a standard skin prick to see if they had an allergic reaction to peanut protein (yes or no). 71% of those in the peanut flour group no longer demonstrated a peanut allergy compared to 2% of those in the placebo group.

- Calculate the relative risk of desensitization comparing the peanut flour group to the placebo group.

- Interpretation:

- The proportion of successes in group 1 is the RR _____ the proportion of successes in group 2.

Increase in risk:

- Interpretation:

- The proportion of successes in group 1 is the $(RR - 1)$ _____ higher/lower than the proportion of successes in group 2.

Percent increase in risk:

- Interpretation:

- The proportion of successes in group 1 is the $(RR - 1) \times 100$ _____ higher/lower than the proportion of successes in group 2.

- Interpret the value of relative risk from the peanut study in context of the problem.

- Find the increase (or decrease) in risk of desensitization and interpret this value in context of the problem.

- Find the percent increase (or decrease) in risk of desensitization and interpret this value in context of the problem.

Within the peanut flour group, the percent desensitized within each age group (at start of study) is as follows:

1-year-olds: 71%; 2-year-olds: 35%; 3-year-olds: 19%

- Calculate the relative risk of desensitization comparing the 3 year olds to the 2 year olds within the peanut flour group.
- Interpret the percent increase (or decrease) in risk of desensitization comparing the 3 year olds to the 2 year olds within the peanut flour group.

Relative risk in the news

People 50 and older who have had a mild case of covid-19 are 15% more likely to develop shingles (herpes zoster) within six months than are those who have not been infected by the coronavirus, according to research published in the journal Open Forum Infectious Diseases (Bhavsar 2022).

- What was the calculated relative risk of developing shingles when comparing those who has mild COVID-19 to those who had not had COVID-19, among the 50 and older population?

Testing Relative Risk

In Unit 2, we tested for a difference in proportion. We could also test for relative risk.

Null Hypothesis:

$$H_0 :$$

Alternative Hypothesis:

$$H_A :$$

3.2.2 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. Explain why the null distribution is centered at the value of zero.
2. Does the confidence interval agree with the p-value?

3.3 Activity 16: Study Design

3.3.1 Learning outcomes

- Explain the purpose of random assignment and its effect on scope of inference.
- Identify whether a study design is observational or an experiment.
- Identify confounding variables in observational studies and explain why they are confounding.

3.3.2 Terminology review

In this activity, we will examine different study designs, confounding variables, and how to determine the scope of inference for a study. Some terms covered in this activity are:

- Scope of inference
- Explanatory variable
- Response variable
- Confounding variable
- Experiment
- Observational study

To review these concepts, see Sections 2.2 through 2.5 in the textbook.

3.3.3 Atrial fibrillation

Atrial fibrillation is an irregular and often elevated heart rate. In some people, atrial fibrillation will come and go on its own, but others will experience this condition on a permanent basis. When atrial fibrillation is constant, medications are required to stabilize the patient's heart rate and to help prevent blood clots from forming. Pharmaceutical scientists at a large pharmaceutical company believe they have developed a new medication that effectively stabilizes heart rates in people with permanent atrial fibrillation. They set out to conduct a trial study to investigate the new drug. The scientists will need to compare the proportion of patients whose heart rate is stabilized between two groups of subjects, one of whom is given a placebo and the other given the new medication.

1. Identify the explanatory and response variable in this trial study.

Explanatory variable:

Response variable:

Suppose 24 subjects with permanent atrial fibrillation have volunteered to participate in this study. There are 16 subjects that self-identified as male and 8 subjects that self-identified as female.

2. One way to separate into two groups would be to give all the males the placebo and all the females the new drug. Explain why this is not a reasonable strategy.

3. Could the scientists fix the problem with the strategy presented in question 2 by creating equal sized groups by putting 4 males and 8 females into the drug group and the remaining 12 males in the placebo group? Explain your answer.

4. A third strategy would be to **block** on sex. In this type of study, the scientists would assign 4 females and 8 males to each group. Using this strategy, out of the 12 individuals in each group what **proportion** are males?

5. Assume the scientists used the strategy in question 4, but they put the four tallest females and eight tallest males into the drug group and the remaining subjects into the placebo group. They found that the proportion of patients whose heart rate stabilized is higher in the drug group than the placebo group.

Could that difference be due to the sex of the subjects? Explain your answer.

Could it be due to other variables? Explain your answer.

While the strategy presented in question 5 controlled for the sex of the subject, there are more potential **confounding variables** in the study. A confounding variable is a variable that is *both*

1. associated with the explanatory variable, *and*
2. associated with the response variable.

When both these conditions are met, if we observe an association between the explanatory variable and the response variable in the data, we cannot be sure if this association is due to the explanatory variable or the confounding variable—the explanatory and confounding variables are “confounded.”

Random assignment means that subjects in a study have an equally likely chance of receiving any of the available treatments.

6. You will now investigate how randomly assigning subjects impacts a study's scope of inference.

- Navigate to the “Randomizing Subjects” applet under the “Other Applets” heading at: <http://www.rossmanchance.com/ISIapplets.html>. This applet lists the sex and height of each of the 24 subjects. Click “Show Graphs” to see a bar chart showing the sex of each subject. Currently, the applet is showing the strategy outlined in question 3.
- Click “Randomize”.

In this random assignment, what proportion of males are in group 1 (the placebo group)?

What proportion of males are in group 2 (the drug group)?

What is the difference in proportion of males between the two groups (placebo - drug)?

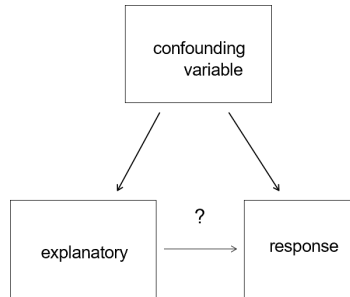
7. Notice the difference in the two proportions is shown as a dot in the plot at the bottom of the web page. Un-check the box for Animate above “Randomize” and click “Randomize” again. Did you get the same difference in proportion of males between the placebo and drug groups?

8. Change “Replications” to 998 (for 1000 total). Click “Randomize” again. Sketch the plot of the distribution of difference in proportions from each of the 1000 random assignments here. Be sure to include a descriptive x -axis label.

9. Does random assignment *always* balance the placebo and drug groups based on the sex of the participants? Does random assignment *tend* to make the placebo and drug groups *roughly* the same with respect to the distribution of sex? Use your plot from question 8 to justify your answers.

10. Change the drop-down menu below Group 2 from “sex” to “height”. The applet now calculates the average height in the placebo and drug groups for each of the 1000 random assignments. The dot plot displays the distribution of the difference in mean heights (placebo - drug) for each random assignment. Based on this dot plot, is height distributed equally, on average, between the two groups? Explain how you know.

The diagram below summarizes these ideas about confounding variables and random assignment. When a confounding variable is present (such as sex or height), and an association is found in a study, it is impossible to discern what caused the change in the response variable. Is the change the result of the explanatory variable or the confounding variable? However, if all confounding variables are *balanced* across the treatment groups, then only the explanatory variable differs between the groups and thus *must have caused* the change seen in the response variable.



11. What is the purpose of random assignment of the subjects in a study to the explanatory variable groups? Cross out the arrow in the figure above that is eliminated by random assignment.


12. Suppose in this study on atrial fibrillation, the scientists did randomly assign groups and found that the drug group has a higher proportion of subjects whose heart rates stabilized than the placebo group. Can the scientists conclude the new drug *caused* the increased chance of stabilization? Explain your answer.

13. Is the sample of subjects a simple random sample or a convenience sample?


14. Both the sampling method and the study design will help to determine the *scope of inference* for a study: To *whom* can we generalize, and can we conclude *causation or only association*? Use your answers to question 12 and 13 and the table on the next page to determine the scope of inference of this trial study described in question 12.

Scope of Inference: If evidence of an association is found in our sample, what can be concluded?

Selection of cases	Study Type		
	Randomized experiment	Observational study	
Random sample (and no other sampling bias)	Causal relationship, and can generalize results to population.	Cannot conclude causal relationship, but can generalize results to population.	→ Inferences to population can be made
No random sample (or other sampling bias)	Causal relationship, but cannot generalize results to a population.	Cannot conclude causal relationship, and cannot generalize results to a population.	→ Can only generalize to those similar to the sample due to potential sampling bias



Can draw cause-and-effect conclusions



Can only discuss association
due to potential confounding
variables

3.3.4 Scope of Inference

The two main study designs we will cover are **observational studies** and **experiments**. In observational studies, researchers have no influence over which subjects are in each group being compared (though they can control other variables in the study). An experiment is defined by assignment of the treatment groups of the *explanatory variable*, typically via random assignment.

For the next exercises identify the study design (observational study or experiment), the sampling method, and the scope of inference.

15. The pharmaceutical company Moderna Therapeutics, working in conjunction with the National Institutes of Health, conducted Phase 3 clinical trials of a vaccine for COVID-19 in the Fall of 2021. US clinical research sites enrolled 30,000 volunteers without COVID-19 to participate. Participants were randomly assigned to receive either the candidate vaccine or a saline placebo. They were then followed to assess whether or not they developed COVID-19. The trial was double-blind, so neither the investigators nor the participants knew who was assigned to which group.

Study design:

Sampling method:

Scope of inference:

16. In another study, a local health department randomly selected 1000 US adults without COVID-19 to participate in a health survey. Each participant was assessed at the beginning of the study and then followed for one year. They were interested to see which participants elected to receive a vaccination for COVID-19 and whether any participants developed COVID-19.

Study design:

Sampling method:

Scope of inference:

3.3.5 Take-home messages

1. The study design (observational study vs, experiment) determines if we can draw causal inferences or not. If an association is detected, a randomized experiment allows us to conclude that there is a causal (cause-and-effect) relationship between the explanatory and response variable. Observational studies have potential confounding variables within the study that prevent us from inferring a causal relationship between the variables studied.
2. Confounding variables are variables not included in the study that are related to both the explanatory and the response variables. When there are potential confounding variables in the study we cannot draw causal inferences.
3. Random assignment balances confounding variables across treatment groups. This eliminates any possible confounding variables by breaking the connections between the explanatory variable and the potential confounding variables.
4. Observational studies will always carry the possibility of confounding variables. Randomized experiments, which use random assignment, will have no confounding variables.

3.3.6 Additional notes

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

3.4 Activity 17: Summarizing Two Categorical Variables

3.4.1 Learning outcomes

- Identify and create appropriate summary statistics and plots given a data set or research question involving categorical variables.
- Plots for association between two categorical variables: segmented bar plot, mosaic plot.
- Calculate and interpret relative risk

3.4.2 Terminology review

In today's activity, we will review summary measures and plots for categorical variables. Some terms covered in this activity are:

- Conditional proportions
- Segmented bar plots
- Mosaic plots
- Relative risk

To review these concepts, see Chapter 4 in the textbook.

3.4.3 Graphing categorical variables

Follow these steps to upload the necessary R script file for today's activity:

- Download the RScript file for this Activity from D2L
- Upload and open the file on the server

the R script file from D2L

- Enter the name of the dataset ("myopia.csv") in line 6

Highlight and run lines 1–3 to load the packages needed for today's activity. Notice the use of the `#` symbol in the R script file. The `#` sign is not part of the R code. It is used by these authors to add comments to the R code and explain what each call is telling the program to do.

R will ignore everything after a `#` sign when executing the code. Refer to the instructions following the `#` sign to understand what you need to enter in the code.

Nightlight use and myopia

In a study reported in Nature (Quinn et al. 1999), a survey of 479 children found that those who had slept with a nightlight or in a fully lit room before the age of two had a higher incidence of nearsightedness (myopia) later in childhood.

In this study, there are two variables studied: **Light**: level of light in room at night (no light, nightlight, full light) and **Sight**: level of myopia developed later in childhood (high myopia, myopia, no myopia).

1. Which variable is the explanatory variable? Which is the response variable?

An important part of understanding data is to create visual pictures of what the data represent. In this activity,

we will create graphical representations of categorical data.

R code

The line of code shown below (line 6 in the R script file) reads in the data set and names the data set `myopia`. Highlight and run line 6 in the R script file to load the data from the Stat 216 webpage.

```
# This will read in the data set
myopia <- read.csv("https://math.montana.edu/courses/s216/data/ChildrenLightSight.csv")
```

2. Click on the data set name (`myopia`) in the Environment tab (upper right window). This will open the data set in a 2nd tab in the Editor window (upper left window). R is case sensitive, which means that you must always type the name of a variable EXACTLY as it is written in the data set including upper and lower case letters and without misspellings! Write down the name of each variable (column names) as it is written in the data set.

Summarizing two categorical variables

Is there an association between the level of light in a room and the development of myopia? Fill in the name of the explanatory variable, **Light** for explanatory and name of the response variable, **Sight** in line 29 in the R script file, highlight and run line 29 to get the counts for each combination of levels of variables.

```
myopia %>% group_by(explanatory) %>% count(response)
```

3. Fill in the following table with the values from the R output.

	Light Level			
Myopia Level	Full Light	Nightlight	No Light	Total
High Myopia				
Myopia				
No Myopia				
Total				

In the following questions, use the table to calculate the described proportions. Notation is important for each calculation. Since this is sample data, it is appropriate to use statistic notation for the proportion, \hat{p} . When calculating a proportion dependent on a single level of a variable, subscripts are needed when reporting the notation.

4. Calculate the proportion of children with no myopia. Use appropriate notation.
5. Calculate the proportion of children with no myopia among those that slept with full light. Use appropriate notation.
6. Calculate the proportion of children with no myopia among those that slept with no light. Use appropriate notation.

7. Calculate the difference in proportion of children with no myopia for those that slept with full light minus those who slept with no light. Give the appropriate notation. Use full light minus no light as the order of subtraction.
8. Interpret the calculated difference in proportion in context of the study.

Displaying two categorical variables

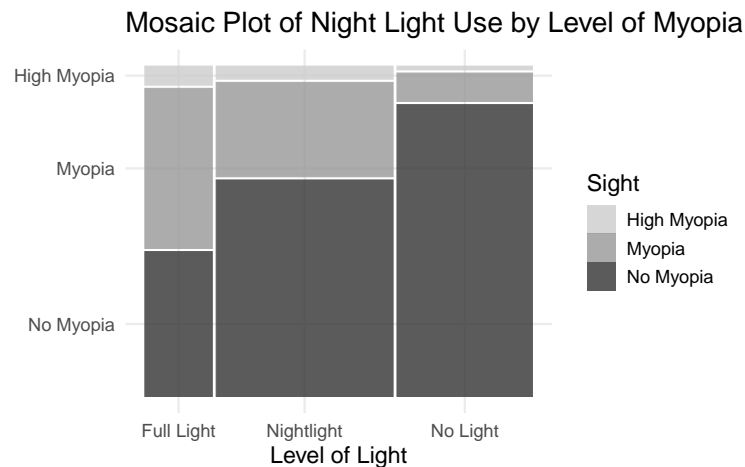
Two types of plots can be created to display two categorical variables. To examine the differences in level of myopia for the level of light, we will first create a segmented bar plot of **Light** segmented by **Sight**. To create the segmented bar plot enter the variable name, **Light** for **explanatory** and the variable name, **Sight** for **response** in the R script file in line 35. Highlight and run lines 34–40.

```
myopia %>% # Data set piped into...
ggplot(aes(x = explanatory, fill = response)) + # This specifies the variables
  geom_bar(stat = "count", position = "fill") + # Tell it to make a stacked bar plot
  labs(title = "Segmented Bar Plot of Night Light Use by Level of Myopia",
        # Make sure to title your plot
        x = "Level of Light", # Label the x axis
        y = "") + # Remove y axis label
  scale_fill_viridis_d() # Make figure color
```

9. Sketch the segmented bar plot created here. Be sure to label the axes.
10. From the segmented bar plot, which level of light has the highest proportion of No Myopia?
11. Based on the plot, is there an association between level of light and level of myopia?

We could also plot the data using a mosaic plot which is shown below.

```
myopia$Sight <- factor(myopia$Sight, levels = c("No Myopia", "Myopia", "High Myopia"))
myopia %>% # Data set piped into...
  ggplot() + # This specifies the variables
  geom_mosaic(aes(x=product(Light), fill = Sight)) + # Tell it to make a mosaic plot
  labs(title = "Mosaic Plot of Night Light Use by Level of Myopia", # Make sure to title your plot
       x = "Level of Light", # Label the x axis
       y = "") + # Remove y axis label
  scale_fill_grey(guide = guide_legend(reverse = TRUE)) # Make figure color
#> Warning: The `scale_name` argument of `continuous_scale()` is deprecated as of ggplot2
#> 3.5.0.
#> This warning is displayed once every 8 hours.
#> Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
#> generated.
#> Warning: The `trans` argument of `continuous_scale()` is deprecated as of ggplot2 3.5.0.
#> i Please use the `transform` argument instead.
#> This warning is displayed once every 8 hours.
#> Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
#> generated.
#> Warning: `unite_()` was deprecated in tidyr 1.2.0.
#> i Please use `unite()` instead.
#> i The deprecated feature was likely used in the ggmosaic package.
#> Please report the issue at <https://github.com/haleyjeppson/ggmosaic>.
#> This warning is displayed once every 8 hours.
#> Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
#> generated.
```



12. What is similar and what is different between the segmented bar chart and the mosaic bar chart?

13. Explain why the bar for **Nightlight** is the widest in the mosaic plot.

Relative Risk

14. Calculate the relative risk of myopia for children that slept with full light compared to those that slept with no light.
15. Interpret the value of relative risk in context of the problem.
16. Calculate the percent increase/decrease in risk of myopia for children that slept with full light compared to those that slept with no light.
17. Interpret as a percent increase/decrease in risk in context of the problem.

3.4.4 Take-home messages

1. Bar charts can be used to graphically display a single categorical variable either as counts or proportions. Segmented bar charts and mosaic plots are used to display two categorical variables.
2. Segmented bar charts always have a scale from 0 - 100%. The bars represent the outcomes of the explanatory variable. Each bar is segmented by the response variable. If the heights of each segment are the same for each bar there is no association between variables.
3. Mosaic plots are similar to segmented bar charts but the widths of the bars also show the number of observations within each outcome.

3.4.5 Additional notes

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

3.5 Activity 18: The Good Samaritan

3.5.1 Learning outcomes

- Given a research question involving two categorical variables, construct the null and alternative hypotheses in words and using appropriate statistical symbols.
- Investigate the process of creating a null distribution for two categorical variables
- Find and evaluate a p-value from the null distribution

3.5.2 Terminology review

In today's activity, we will use simulation-based methods to analyze two categorical variables. Some terms covered in this activity are:

- Conditional proportion
- Null hypothesis
- Alternative hypothesis

To review these concepts, see Chapter 15 in your textbook.

3.5.3 The Good Samaritan

Researchers at the Princeton University wanted to investigate influences on behavior (Darley and Batson 1973). The researchers randomly selected 67 students from the Princeton Theological Seminary to participate in a study. Only 47 students chose to participate in the study, and the data below includes 40 of those students (7 students were removed from the study for various reasons). As all participants were theology majors planning a career as a preacher, the expectation was that all would have a similar disposition when it comes to helping behavior. Each student was then shown a 5-minute presentation on the Good Samaritan, a parable in the Bible which emphasizes the importance of helping others. After the presentation, the students were told they needed to give a talk on the Good Samaritan parable at a building across campus. Half the students were told they were late for the presentation; the other half told they could take their time getting across campus (the condition was randomly assigned). On the way between buildings, an actor pretending to be a homeless person in distress asked the student for help. The researchers recorded whether the student helped the actor or not. The results of the study are shown in the table below. Do these data provide evidence that those in a hurry will be less likely to help people in need in this situation? Use the order of subtraction hurry – no hurry.

	Hurry Condition	No Hurry Condition	Total
Helped Actor	2	11	13
Did Not Help Actor	18	9	27
Total	20	20	40

These counts can be found in R by using the `count()` function:

- Download the R script file from D2L and upload to the RStudio server
- Highlight and run lines.....to get the counts for each group

```
# Read data set in
good <- read.csv("https://math.montana.edu/courses/s216/data/goodsam.csv")
good %>% group_by(Condition) %>% count(Behavior)
```

```
#> # A tibble: 4 x 3
#> # Groups:   Condition [2]
#>   Condition Behavior      n
#>   <chr>      <chr>    <int>
```



```
#> 1 Hurry      Help      2
#> 2 Hurry      No help   18
#> 3 No hurry   Help      11
#> 4 No hurry   No help    9
```

Ask a research question

The research question as stated above is: Do these data provide evidence that those in a hurry will be less likely to help people in need in this situation? In order to set up our hypotheses, we need to express this research question in terms of parameters.

Remember, we define the parameter for a single categorical variable as the true proportion of observational units that are labeled as a “success” in the response variable.

For this study we are identifying two parameters and looking at the difference between these two parameters.

- π_{hurry} = long-run proportion of Princeton Theological Seminary students assigned to hurry that helped the actor
- $\pi_{\text{no hurry}}$ = long-run proportion of Princeton Theological Seminary students assigned not to hurry that helped the actor
- $\pi_{\text{hurry}} - \pi_{\text{no hurry}}$ = the difference in long-run proportion of Princeton Theological Seminary Students that helped the actor between those who were assigned to hurry and those who were not assigned to hurry

When comparing two groups, we assume the two parameters are equal in the null hypothesis—there is no association between the variables.

1. Write the null hypothesis out in words.
2. Based on the research question, fill in the appropriate sign for the alternative hypothesis ($<$, $>$, or \neq):

$$H_A : \pi_{\text{hurry}} - \pi_{\text{no hurry}} \quad \underline{\hspace{1cm}} \quad 0$$

Summarize and visualize the data

To create the segmented bar plot:

- Enter the name of the explanatory variable for explanatory
- Enter the name of the response variable for response
- Highlight and run lines.....

```
good %>%
  ggplot(aes(x = explanatory, fill = response)) + #Enter the variables to plot
  geom_bar(stat = "count", position = "fill") +
  labs(title = "Segmented bar plot of Condition of Seminary \n Students by Behavior", #Title your plot
        y = "Relative Frequency", #y-axis label
        x = "Condition") + #x-axis label
  scale_fill_grey()
```

3. Based on the segmented bar plot, is there an association between whether a Seminary student helps the actor and condition assigned?

4. Using the two-way table given in the introduction, calculate the conditional proportion of students in the hurry condition who helped the actor. Use appropriate notation.
5. Using the two-way table given in the introduction, calculate the conditional proportion of students in the no hurry condition who helped the actor. Use appropriate notation.
6. Calculate the summary statistic (difference in sample proportion) for this study. Use Hurry - No hurry as the order of subtraction. Use appropriate notation.

Interpretation of the summary statistic:

The proportion of Princeton Theological Seminary students that helped the actor is 0.45 less for those assigned to hurry compared to those assigned not to hurry.

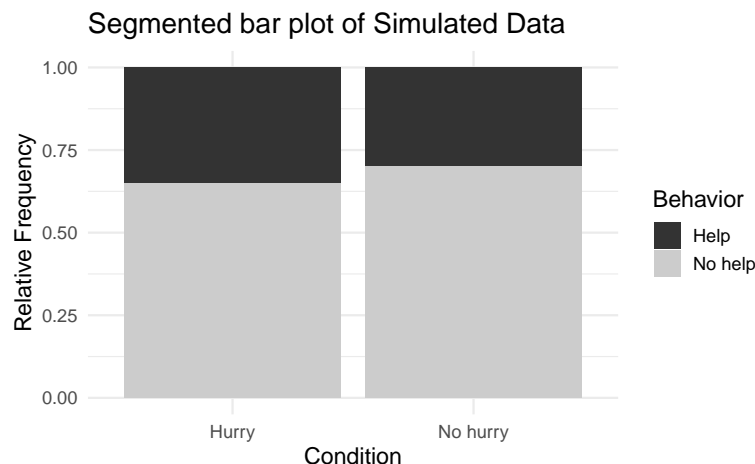
Hypothesis Test

We will now simulate a **null distribution** of sample differences in proportions. The null distribution is created under the assumption the null hypothesis is true.

Using the cards provided by your instructor, simulate one sample under the assumption the null hypothesis is true.

- Start with 40 cards (13 labeled helped, 27 labeled did not help)
- Mix the cards together
- Shuffle the cards into two piles (20 in hurry, 20 in no hurry)
- Calculate the proportion of simulated students that helped in each group.
- Report the difference in proportion of simulated students that helped (hurry - no hurry)

The segmented bar plot below shows the relationship between the variables for **one simulation assuming the null hypothesis is true**.



To create the null distribution of differences in sample proportions, we will use the `two_proportion_test()` function in R (in the `catstats` package). We will need to enter the response variable name and the explanatory variable name for the formula, the data set name (identified above as `good`), the outcome for the explanatory variable that is first in subtraction, number of repetitions, the outcome for the response variable that is a success (what the numerator counts when calculating a sample proportion), and the direction of the alternative hypothesis.

The response variable name is `Behavior` and the explanatory variable name is `Condition`.

8. What inputs should be entered for each of the following to create the simulation?

- First in subtraction (What is the outcome for the explanatory variable that is used as first in the order of subtraction? "Hurry" or "No hurry"):
- Number of repetitions:
- Response value numerator (What is the outcome for the response variable that is considered a success? "Help" or "No help"):
- As extreme as (enter the value for the sample difference in proportions):
- Direction ("greater", "less", or "two-sided"):

Using the R script file for this activity, enter your answers for question 16 in place of the `xx`'s to produce the null distribution with 1000 simulations; highlight and run lines 1–16.

```
two_proportion_test(formula = Behavior~Condition, # response ~ explanatory
  data = good, # Name of data set
  first_in_subtraction = "xx", # Order of subtraction: enter the name of Group 1
  number_repetitions = 1000, # Always use a minimum of 1000 repetitions
  response_value_numerator = "xx", # Define which outcome is a success
  as_extreme_as = xx, # Calculated observed statistic (difference in sample proportions)
  direction="xx") # Alternative hypothesis direction ("greater","less","two-sided")
```

9. Sketch the null distribution created here.

10. Explain why the null distribution is centered around the value of zero?

11. Interpret the p-value in context of the study.

12. Write a conclusion in context of the study.

3.5.4 Take-home messages

1. When comparing two groups, we are looking at the difference between two parameters. In the null hypothesis, we assume the two parameters are equal, or that there is no difference between the two proportions.
2. To create one simulated sample on the null distribution for a difference in sample proportions, label $n_1 + n_2$ cards with the response variable outcomes from the original data. Mix cards together and shuffle into two new groups of sizes n_1 and n_2 , representing the explanatory variable groups. Calculate and plot the difference in proportion of successes.

3.5.5 Additional notes

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

Inference for a Two Categorical Variable: Theory-based Methods

4.1 Vocabulary Review and Key Topics

Review the Golden Ticket posted in the resources at the end of the coursepack for a summary of a two categorical variables.

4.1.1 Key topics

Module 9 introduces theory-based hypothesis testing methods and both simulation-based and theory-based confidence intervals for two categorical variables.

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

- **Independence:** The data are independent within and between the two groups. (*Remember:* This also must be true to use simulation methods!)
- **Success-failure condition:** This condition is met if we have at least 10 successes and 10 failures in each sample. Equivalently, we check that all cells in the table have at least 10 observations.
- Calculation of standard error assuming the null is true:

$$SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\hat{p}_{pooled} \times (1 - \hat{p}_{pooled}) \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

- Calculation of the standardized difference in sample proportion:

$$t = \frac{\hat{p}_1 - \hat{p}_2 - 0}{SE(\hat{p}_1 - \hat{p}_2)}$$

- Measures the number of standard errors the sample difference in proportions is above or below the null value of zero
- Calculation of the difference in sample proportion not assuming the null is true

$$SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1 \times (1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2 \times (1 - \hat{p}_2)}{n_2}}$$

* Calculation of the confidence interval for a difference in sample proportions

$$\hat{p}_1 - \hat{p}_2 \pm z^* \times SE(\hat{p}_1 - \hat{p}_2)$$

4.2 Video Notes: Theoretical Inference for Two Categorical Variables

Read Sections 15.3 and 15.4 in the course textbook. Use the following videos to complete the video notes for Module 9.

4.2.1 Course Videos

- 15.3TheoryTests
- 15.3TheoryIntervals

Hypothesis testing using theory-based methods - Video 15.4TheoryTests

Example: In Modules 3 and 4, we investigated data on higher education institutions in the United States, collected by the Integrated Postsecondary Education Data System (IPEDS) for the National Center for Education Statistics (NCES) (Education Statistics 2018). A random sample of 2900+ higher education institutions in the United States was collected in 2018. Two variables measured on this data set is whether the institution is a land grant university and whether the institution offers tenure. Does the proportion of universities that offer tenure differ between land grant and non-land-grant institutions?

What is the explanatory variable?

What is the response variable?

Write the parameter of interest:

Hypotheses:

In notation:

H_0 :

H_A :

```
IPED <- read.csv("https://math.montana.edu/courses/s216/data/IPEDS_2018.csv")

IPEDS <- IPED %>%
  drop_na(Tenure)

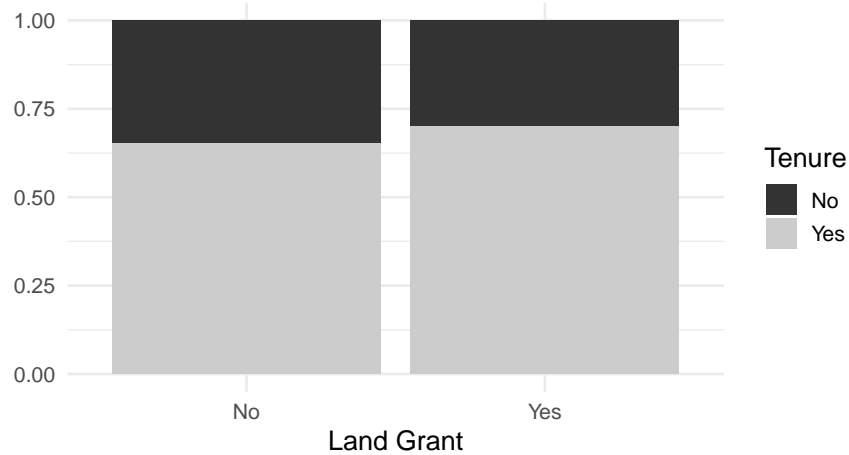
IPEDS %>% # Data set piped into...
  ggplot(aes(x = LandGrant, fill = Tenure)) + # This specifies the variables
  geom_bar(stat = "count", position = "fill") + # Tell it to make a stacked bar plot
  labs(title = "Segmented Bar Plot of Tenure Availability
    by Type of Institution for Higher Ed Institutions",
    # Make sure to title your plot
    x = "Land Grant", # Label the x axis
    y = "") + # Remove y axis label
```

```
scale_fill_grey()
```

```
IPEDS %>% group_by(LandGrant) %>% count(Tenure)
```

```
#> # A tibble: 4 x 3
#> # Groups:   LandGrant [2]
#>   LandGrant Tenure     n
#>   <chr>      <chr> <int>
#> 1 No        No      976
#> 2 No        Yes     1829
#> 3 Yes       No       31
#> 4 Yes       Yes       72
```

Segmented Bar Plot of Tenure Availability
by Type of Institution for Higher Ed Institutions



Report the summary statistic:

Conditions for inference using theory-based methods for two categorical variables:

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

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

Steps to use theory-based methods:

- Calculate the standardized statistic
- Find the area under the standard normal distribution at least as extreme as the standardized statistic

Equation for the standard error of the difference in sample proportions assuming the null hypothesis is true:

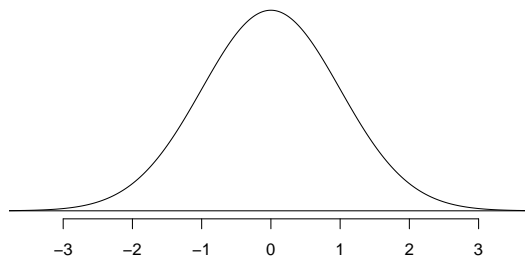
- This value measures how far each possible sample difference in proportions is from the null value, on average.

Equation for the standardized difference in sample proportions:

- This value measures how many standard errors the sample difference in proportions is above/below the null value.

Calculate the standardized difference in sample proportion of higher education institutions that offer tenure between land grant universities and non-land grant universities.

- First calculate the standard error of the difference in proportion assuming the null hypothesis is true
- Then calculate the Z score



Interpret the standardized statistic

To find the p-value, find the area under the standard normal distribution at the standardized statistic and more extreme.

```
pnorm(0.985, lower.tail = FALSE)*2
```

```
#> [1] 0.3246241
```

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 with scope of inference:

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

Confidence interval - Video 15.3 Theory Intervals

- Estimate the _____ in true _____
- $CI = \text{statistic} \pm \text{margin of error}$

Theory-based method for a two categorical variables

- $CI = \hat{p}_1 - \hat{p}_2 \pm (z^* \times SE(\hat{p}_1 - \hat{p}_2))$
- When creating a confidence interval, we no longer assume the _____ hypothesis is true.
Use the sample _____ to calculate the sample to sample variability, rather than \hat{p}_{pooled} .

Equation for the standard error of the difference in sample proportions *NOT* assuming the null is true:

Example: Estimate the difference in true proportions of higher education institutions that offer tenure between land grant universities and non-land grant universities.

Find a 90% confidence interval:

- 1st find the z^* multiplier

```
qnorm(0.95, lower.tail=TRUE)
```

```
#> [1] 1.644854
```

- Next, calculate the standard error for the difference in proportions **NOT** assuming the null hypothesis is true

- Calculate the margin of error

- Calculate the endpoints of the 90% confidence interval

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

4.2.2 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. What conditions must be met to use the Normal Distribution to approximate the sampling distribution for the difference in sample proportions?
2. How is the value of relative risk calculated?

3. Explain why a theory-based confidence interval for the Good Samaritan study from last module would NOT be similar to the bootstrap interval created.

4.3 Activity 19: Winter Sports Helmet Use and Head Injuries — Theory-based Methods

4.3.1 Learning outcomes

- Assess the conditions to use the normal distribution model for a difference in proportions.
- Create and interpret a theory-based confidence interval for a difference in proportions.
- Calculate and interpret the standardized difference in sample proportion
- Use the standard normal distribution to find the p-value for the test

4.3.2 Terminology review

In today's activity, we will use theory-based methods to estimate the difference in two proportions. Some terms covered in this activity are:

- Standard normal distribution
- Independence and success-failure conditions

To review these concepts, see Chapter 15 in your textbook.

4.3.3 Winter sports helmet use and head injury

In this activity we will focus on theory-based methods to calculate a confidence interval. The sampling distribution of a difference in proportions can be mathematically modeled using the normal distribution if certain conditions are met.

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

- **Independence:** The data are independent within and between the two groups. (*Remember:* This also must be true to use simulation methods!)
- **Success-failure condition:** This condition is met if we have at least 10 successes and 10 failures in each sample. Equivalently, we check that all cells in the table have at least 10 observations.

A study was reported in “Helmet Use and Risk of Head Injuries in Alpine Skiers and Snowboarders” by Sullheim et. al., (Sulheim et al. 2017), on the use of helmets and head injuries for skiers and snowboarders involved in accidents. The summary results from a random sample of 3562 skiers and snowboarders involved in accidents is shown in the two-way table below.

	Helmet Use	No Helmet Use	Total
Head Injury	96	480	576
No Head Injury	656	2330	2986
Total	752	2810	3562

- Download the R script file from D2L and upload to the RStudio server
- Enter the name of the dataset
- Highlight and run...to import the data set and create the segmented bar plot

```
skiers <- read.csv("https://www.math.montana.edu/courses/s216/data/HeadInjuries.csv") # Read data set in
skiers %>% # Data set piped into...
  ggplot(aes(x = Helmet, fill = Outcome)) + # This specifies the variables
  geom_bar(stat = "count", position = "fill") + # Tell it to make a stacked bar plot
```

```
labs(title = "Segmented Bar Plot of Head Injuries for Skiers/Snowboarders
  Involved in Injuries between Helmet Use", # Make sure to title your plot
  x = "Helmet Use", # Label the x axis
  y = "") + # Remove y axis label
scale_fill_grey() # Make figure color
```

1. Verify the independence condition is met.
2. Verify the success failure condition is met to use theory-based methods.
3. Calculate the difference in sample proportion of skiers and snowboarders involved in accidents with a head injury for those who wear helmets and those who do not. Use appropriate notation with informative subscripts.

Hypothesis test

4. Write the null and alternative hypotheses in notation.

H_0 :

H_A :

Use statistical analysis methods to draw inferences from the data

To test the null hypothesis, we could use simulation-based methods as we did in the activities in Module 7. In this activity, we will focus on theory-based methods. Like with a single proportion, the sampling distribution of a difference in sample proportions can be mathematically modeled using the normal distribution if certain conditions are met.

To calculate the standardized statistic we use:

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - \text{null value}}{SE_0(\hat{p}_1 - \hat{p}_2)},$$

where the null standard error is calculated using the pooled proportion of successes:

$$SE_0(\hat{p}_1 - \hat{p}_2) = \sqrt{\hat{p}_{pool} \times (1 - \hat{p}_{pool}) \times \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}.$$

For this study we would first calculate the pooled proportion of successes.

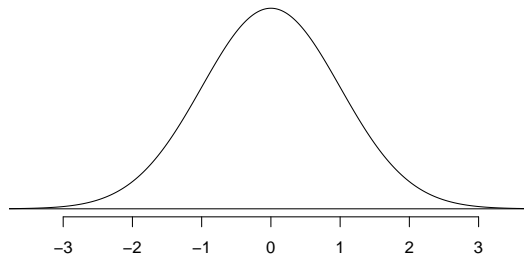
$$\hat{p}_{pool} = \frac{\text{number of "successes"}}{\text{number of cases}}$$

5. Calculate the pooled proportion of head injuries.

6. Use the value for the pooled proportion of successes to calculate the $SE_0(\hat{p}_1 - \hat{p}_2)$ assuming the null hypothesis is true.

7. Use the value of the null standard error to calculate the standardized statistic (standardized difference in proportion).

8. Mark the value of the standardized difference in proportion on the standard normal distribution shown below. Interpret this value in context of the problem.



We will use the `pnorm()` function in R to find the p-value.

- Enter the value of z for xx
- Highlight and run lines...

```
pnorm(xx, # Enter value of standardized statistic  
      m=0, s=1, # Using the standard normal mean = 0, sd = 1
```

```
lower.tail=TRUE) # Gives a p-value less than the standardized statistic
```

9. Write a conclusion to the test.

How would an increase in sample size impact the p-value of the test?

	Helmet Use	No Helmet Use	Total
Head Injury	135	674	809
No Head Injury	921	3270	4191
Total	1056	3944	5000

Note that the sample proportions for each group are the same as the smaller sample size.

$$\hat{p}_h = \frac{135}{1056} = 0.128, \quad \hat{p}_n = \frac{674}{3944} = 0.171$$

First calculate the pooled proportion of successes.

$$\hat{p}_{pool} = \frac{\text{number of "successes"}}{\text{number of cases}} = \frac{809}{5000} = 0.162$$

We use the value for the pooled proportion of successes to calculate the $SE_0(\hat{p}_1 - \hat{p}_2)$.

$$SE_0(\hat{p}_1 - \hat{p}_2) = \sqrt{0.162 \times (1 - 0.162) \times \left(\frac{1}{1056} + \frac{1}{3944} \right)} = 0.013$$

Standardized Statistic Calculation:

$$Z = \frac{0.128 - 0.171 - 0}{0.013} = -3.308$$

Use Rstudio to find the p-value for this new sample.

```
pnorm(-3.308, # Enter value of standardized statistic
      m=0, s=1, # Using the standard normal mean = 0, sd = 1
      lower.tail=TRUE) # Gives a p-value greater than the standardized statistic
```

```
#> [1] 0.000469824
```

10. How does the increase in sample size affect the p-value?

4.3.4 Take-home messages

1. Simulation-based methods and theory-based methods should give similar results for a study *if the validity conditions are met*. For both methods, observational units need to be independent. To use theory-based methods, additionally, the success-failure condition must be met. Check the validity conditions for each type of test to determine if theory-based methods can be used.
2. When calculating the standard error for the difference in sample proportions when doing a hypothesis test, we use the pooled proportion of successes, the best estimate for calculating the variability *under the assumption the null hypothesis is true*. For a confidence interval, we are not assuming a null hypothesis, so we use the values of the two conditional proportions to calculate the standard error. Make note of the difference in these two formulas.
3. Increasing sample size will result in less sample-to-sample variability in statistics, which will result in a smaller standard error, and a larger standardized statistic.

4.3.5 Additional notes

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

4.4 Activity 20: Diabetes

4.4.1 Learning outcomes

- Assess the conditions to use the normal distribution model for a difference in proportions.
- Describe and perform a simulation-based confidence interval for a difference in proportions.
- Create and interpret a theory-based confidence interval for a difference in proportions.

4.4.2 Glycemic control in diabetic adolescents

Researchers compared the efficacy of two treatment regimens to achieve durable glycemic control in children and adolescents with recent-onset type 2 diabetes (Group 2012). A convenience sample of patients 10 to 17 years of age with recent-onset type 2 diabetes were randomly assigned to either a medication (rosiglitazone) or a lifestyle-intervention program focusing on weight loss through eating and activity. Researchers measured whether the patient still needs insulin (failure) or had glycemic control (success). Of the 233 children who received the Rosiglitazone treatment, 143 had glycemic control, while of the 234 who went through the lifestyle-intervention program, 125 had glycemic control. Is there evidence that there is difference in proportion of patients that achieve durable glycemic control between the two treatments? Use Rosiglitazone – Lifestyle as the order of subtraction.

- Upload and open the R script file. Upload the csv file, **diabetes**.
- Enter the name of the data set for **datasetname.csv** in the R script file in line 7.
- Highlight and run lines 1–8 to get the counts for each combination of categories.

```
glycemic <- read.csv("datasetname.csv")
glycemic %>% group_by(treatment) %>% count(outcome)
```

1. Is this an experiment or an observational study?
2. Complete the following two-way table using the R output.

Outcome	Treatment		Total
	rosiglitazone	lifestyle	
glycemic control (success)			
insulin required (failure)			
Total			

3. Is the independence condition met for this study? Explain your answer.
4. Is the success failure condition met for this study? Explain your answer.
5. Write the parameter of interest for the research question.

6. Calculate the summary statistic (difference in proportions). Use appropriate notation.

Simulation methods

First we will use simulation methods to find the confidence interval. This will give an interval estimate for the parameter of inference.

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 90% 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 `glycemic`), the outcome for the explanatory variable that is first in subtraction, number of repetitions, the outcome for the response variable that is a success (what the numerator counts when calculating a sample proportion), and the confidence level as a decimal.

7. What inputs should be entered for each of the following to create the bootstrap simulation?
 - First in subtraction (What is the outcome for the explanatory variable that is used as first in the order of subtraction? "rosi" or "lifestyle"):
 - Number of repetitions:
 - Response value numerator (What is the outcome for the response variable that is considered a success? "success" or "failure"):
 - confidence_level:
 - Fill in the missing values/names in the R script file in the `two_proportion_bootstrap_CI` function to create a simulation 90% confidence interval.

```
two_proportion_bootstrap_CI(formula = response~explanatory,  
  data=mushrooms, # Name of data set  
  first_in_subtraction = "xx", # Order of subtraction: enter the name of Group 1  
  response_value_numerator = "xx", # Define which outcome is a success  
  number_repetitions = 1000, # Always use a minimum of 1000 repetitions  
  confidence_level = xx) # Enter the level of confidence as a decimal
```

8. Report the 90% confidence interval.
9. Interpret the confidence interval in context of the problem.

Theory-based Methods

Next we will use theory-based methods to find the 90% confidence interval.

10. Is the sample size large enough to use theory-based methods to find the confidence interval? Explain in context of the study,

To find a confidence interval for the difference in proportions we will add and subtract the margin of error from the point estimate to find the two endpoints.

$$\hat{p}_1 - \hat{p}_2 \pm z^* \times SE(\hat{p}_1 - \hat{p}_2), \text{ where}$$

$$SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1 \times (1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2 \times (1 - \hat{p}_2)}{n_2}}$$

In this formula, we use the sample proportions for each group to calculate the standard error for the difference in proportions since we are not assuming that the true difference is zero.

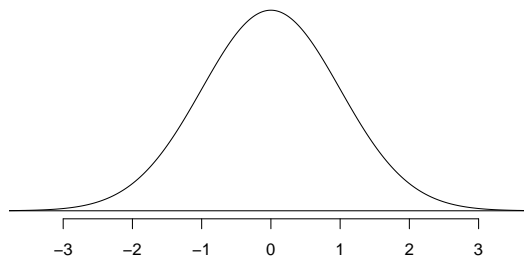
11. Calculate the standard error of the sample proportion not assuming the null hypothesis is true.

Recall that the z^* multiplier is the percentile of a standard normal distribution that corresponds to our confidence level. If our confidence level is 90%, we find the Z values that encompass the middle 90% of the standard normal distribution. If 90% of the standard normal distribution should be in the middle, that leaves 10% in the tails, or 5% in each tail. The `qnorm()` function in R will tell us the z^* value for the desired percentile (in this case, 90% + 5% = 95% percentile).

```
qnorm(0.95, lower.tail = TRUE) # Multiplier for 90% confidence interval
```

```
#> [1] 1.644854
```

12. Mark the value of the z^* multiplier and the percentages used to find this multiplier on the standard normal distribution shown below.



Remember that the margin of error is the value added and subtracted to the sample difference in proportions to find the endpoints for the confidence interval.

$$ME = z^* \times SE(\hat{p}_1 - \hat{p}_2)$$

13. Using the multiplier of $z^* = 1.645$ and the calculated standard error, calculate the margin of error for a 90% confidence interval.

14. Calculate the 90% confidence interval for the parameter of interest.

4.5 Module 8 Lab: Poisonous Mushrooms

4.5.1 Learning outcomes

- Given a research question involving two categorical variables, 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 proportions.
- Interpret and evaluate a p-value for a simulation-based hypothesis test for a difference in proportions.
- Interpret and evaluate a confidence interval for a simulation-based confidence interval for a difference in proportions.

4.5.2 Poisonous Mushrooms

Wild mushrooms, such as chanterelles or morels, are delicious, but eating wild mushrooms carries the risk of accidental poisoning. Even a single bite of the wrong mushroom can be enough to cause fatal poisoning. An amateur mushroom hunter is interested in finding an easy rule to differentiate poisonous and edible mushrooms. They think that the mushroom's gills (the part which holds and releases spores) might be related to a mushroom's edibility. They used a data set of 8124 mushrooms and their descriptions. For each mushroom, the data set includes whether it is edible (e) or poisonous (p) and the size of the gills (broad (b) or narrow (n)). Is there evidence gill size is associated with whether a mushroom is poisonous? PLEASE NOTE: According to The Audubon Society Field Guide to North American Mushrooms, there is no simple rule for determining the edibility of a mushroom; no rule like "leaflets three, let it be" for Poisonous Oak and Ivy.

- Upload and open the R script file for Week 8 lab. Upload and import the csv file, `mushrooms_edibility`.
- Enter the name of the data set (see the environment tab) for `datasetname` in the R script file in line 8.
- Highlight and run lines 1–9 to get the counts for each combination of categories.

```
mushrooms <- datasetname # Read data set in
mushrooms %>% group_by(gill_size) %>% count(edibility) #finds the counts in each group
```

1. What is the explanatory variable? How are the two levels of the explanatory variable written in the data set?
2. What is the response variable? How are the two levels of the response variable written in the data set?
3. Write the parameter of interest in words, in context of the study.
4. Write the null hypothesis for this study in notation.

5. Using the research question, write the alternative hypothesis in words.

6. Fill in the following two-way table using the R output.

	Gill Size		
Edibility	Broad (b)	Narrow (n)	Total
Poisonous (p)			
Edible (e)			
Total			

7. Calculate the difference in proportion of mushrooms that are poisonous for broad gill mushrooms and narrow gill mushrooms. Use broad - narrow for the order of subtraction. Use appropriate notation.

- Fill in the missing values/names in the R script file for the `two-proportion_test` function to create the null distribution and find the p-value for the test.

```
two_proportion_test(formula = response~explanatory, # response ~ explanatory
  data= mushrooms, # Name of data set
  first_in_subtraction = "xx", # Order of subtraction: enter the name of Group 1
  number_repetitions = 1000, # Always use a minimum of 1000 repetitions
  response_value_numerator = "xx", # Define which outcome is a success
  as_extreme_as = xx, # Calculated observed statistic (difference in sample proportions)
  direction="xx") # Alternative hypothesis direction ("greater","less","two-sided")
```

8. Report the p-value for the study.

9. Do you expect that a 90% confidence interval would contain the null value of zero? Explain your answer.

- Fill in the missing values/names in the R script file in the `two_proportion_bootstrap_CI` function to create a simulation 90% confidence interval.
- **Upload a copy of the bootstrap distribution to Gradescope.**

```
two_proportion_bootstrap_CI(formula = response~explanatory,
  data=mushrooms, # Name of data set
  first_in_subtraction = "xx", # Order of subtraction: enter the name of Group 1
  response_value_numerator = "xx", # Define which outcome is a success
  number_repetitions = 1000, # Always use a minimum of 1000 repetitions
  confidence_level = xx) # Enter the level of confidence as a decimal
```

10. Report the 90% confidence interval.
11. Write a paragraph summarizing the results of the study as if writing a press release. Be sure to describe:
 - Summary statistic and interpretation
 - Summary measure (in context)
 - Value of the statistic
 - Order of subtraction when comparing two groups
 - 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
 - To what group of observational units do the results apply (target population or observational units similar to the sample)?
 - What type of inference is appropriate (causal or non-causal)?

Upload your group's confidence interval interpretation and conclusion to Gradescope.

Paragraph:

Unit 2 Review

The following section contains both a list of key topics covered in Unit 2 as well as Module Review Worksheets.

5.0.1 Key Topics

Review the key topics for Unit 2 to review prior to the first exams. All of these topics will be covered in Modules 6–9.

5.0.2 Module Review

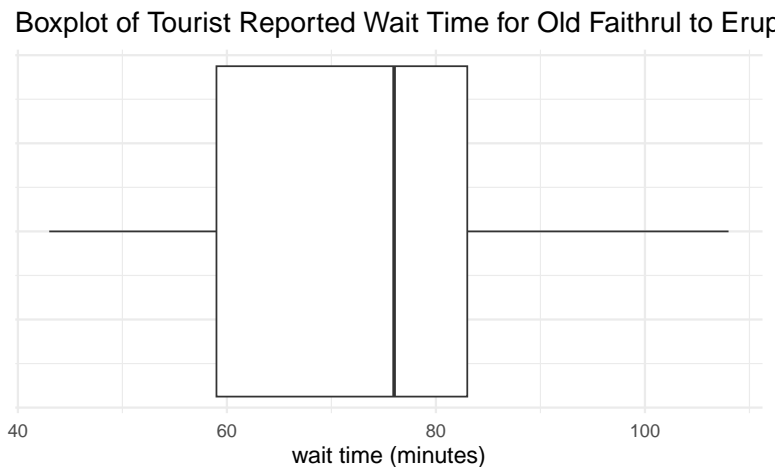
The following worksheets review each of the modules. These worksheets will be completed during Melinda's Study Sessions each week. Solutions will be posted on D2L in the Unit 2 Review folder after the study sessions.

5.1 Module 6 Review - One Mean Testing

There are about 4 million tourists to Yellowstone National Park per year. One of the most visited sites within the park is the Old Faithful Geyser. The reason this geyser is called old faithful is because of the regularity of eruptions. Tourists report a typical wait time of 30 minutes, on average. A sample of 299 tourists reported their wait time to see Old Faithful erupt. Is there evidence that the average wait time differs from 30 minutes?

```
#>   min  Q1 median  Q3 max    mean    sd  n missing
#> 1  43  59     76  83 108 72.31438 13.89032 299      0
```

The following code created the boxplot of waiting time.

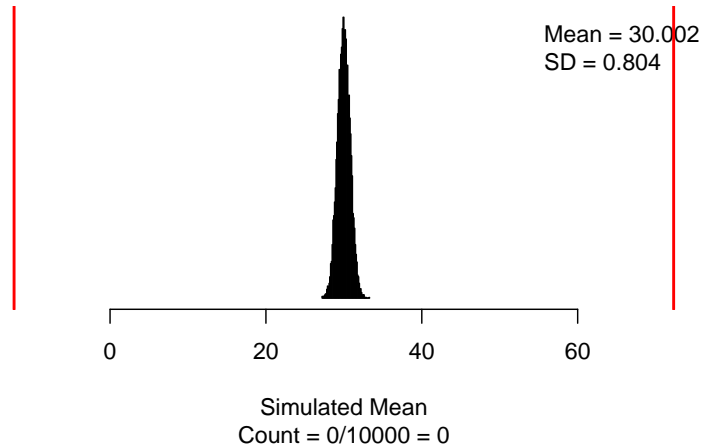


1. Report and interpret the value of Q_1 in context of the study.
2. Report and interpret the standard deviation of wait time in context of the study.
3. Describe the plot using the four characteristics for boxplots.
4. Write the parameter of interest for this study in context of the study.
5. Write the null hypothesis in notation.
6. Write the alternative hypothesis in words.

We will start with simulation methods.

7. Calculate the difference $\mu_0 - \bar{x}$. Will we need to shift the data up or down?

```
set.seed(216)
one_mean_test(data = geysers$waiting, #Object and variable
  null_value = 30, #null value for the study
  shift = -42.31438, #Shift needed for bootstrap hypothesis test
  summary_measure = "mean",
  as_extreme_as = 72.314, #Observed statistic
  direction = "two-sided", #Direction of alternative
  number_repetitions = 10000) #Number of simulated samples for null distribution
```



8. Interpret the p-value of the test.

Now let's focus on theory-based methods.

Conditions for the sampling distribution of \bar{x} to follow an approximate Normal distribution:

- **Independence:** The sample's observations are independent. For paired data, that means each pairwise difference should be independent.
- **Normality:** The data should be approximately normal or the sample size should be large.
 - $n < 30$: If the sample size n is less than 30 and the distribution of the data is approximately normal with 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 in the data, 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.

9. Is the independence condition met?

10. Is the normality condition met to use theory-based methods?

To find the standardized statistic for the mean we will use the following formula:

$$T = \frac{\bar{x} - \mu_0}{SE(\bar{x})},$$

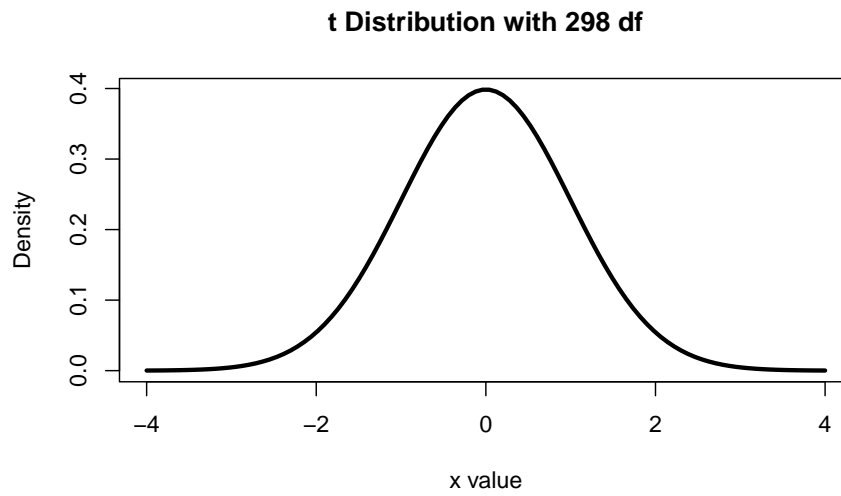
where the standard error of the sample mean difference is:

$$SE(\bar{x}) = \frac{s}{\sqrt{n}}.$$

11. Calculate the standard error of the sample mean.

12. Calculate the standardized mean for the study.

13. Mark on the t-distribution shown below on how to find the p-value of the test.



14. Interpret the standardized mean in context of the study.

The following code calculates the p-value for the study.

```
2*pt(-52.676, df=298, lower.tail=TRUE)  
#> [1] 5.045442e-153
```

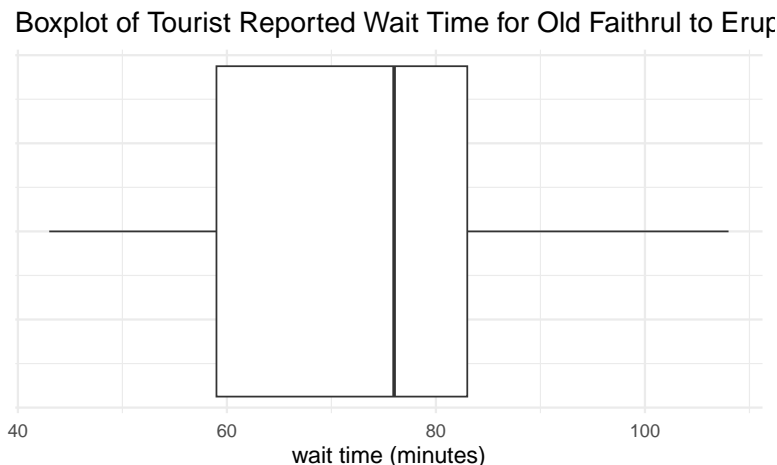
15. Write a conclusion to the test.

5.2 Module 7 Review - One Mean Confidence Interval

There are about 4 million tourists to Yellowstone National Park per year. One of the most visited sites within the park is the Old Faithful Geyser. The reason this geyser is called old faithful is because of the regularity of eruptions. Tourists report a typical wait time of 30 minutes, on average. A sample of 299 tourists reported their wait time to see Old Faithful erupt. How long, on average, do tourists wait for Old Faithful to erupt?

```
#>   min Q1 median Q3 max    mean    sd  n missing
#> 1  43  59    76  83 108 72.31438 13.89032 299      0
```

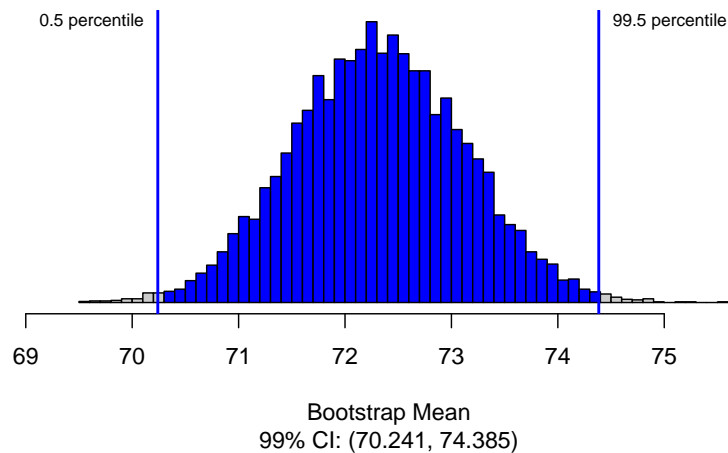
The following code created the boxplot of waiting time.



1. Write the parameter of interest in context of the study.
2. In the last module review, we saw very strong evidence that the true mean wait time reported by tourists for Old Faithful to erupt differs from 30 minutes. Do you expect the 99% confidence interval to contain the null value of zero? Explain your answer.

We will start with simulation methods to create the 99% confidence interval.

```
set.seed(216)
one_mean_CI(data = geyser$waiting,    #Object and variable
             summary_measure = "mean",
             confidence_level = 0.99, #Level of context as a decimal
             number_repetitions = 10000) #Number of simulated samples for null distribution
```



3. How many simulations are at and below the value of 70.241?

4. Report the 99% confidence interval.

Now let's focus on theory-based methods. **In the last module review, we verified the normality conditions were met.**

Conditions for the sampling distribution of \bar{x} to follow an approximate Normal distribution:

- **Independence:** The sample's observations are independent. For paired data, that means each pairwise difference should be independent.
- **Normality:** The data should be approximately normal or the sample size should be large.
 - $n < 30$: If the sample size n is less than 30 and the distribution of the data is approximately normal with 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 in the data, 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 calculate a theory-based confidence interval for the a single mean, use the following formula:

$$\bar{x} \pm t^* \times SE(\bar{x}).$$

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

- Enter the appropriate percentile (0.995) in the R code to find the multiplier for a 99% confidence interval.
- Enter the df $n - 1 = 299 - 1 = 298$

```
qt(0.995, df = 298, lower.tail=TRUE)
```

```
#> [1] 2.592428
```

5. Mark on the t-distribution found below the values of $\pm t^*$. Draw a line at each multiplier and write the percentiles used to find each.

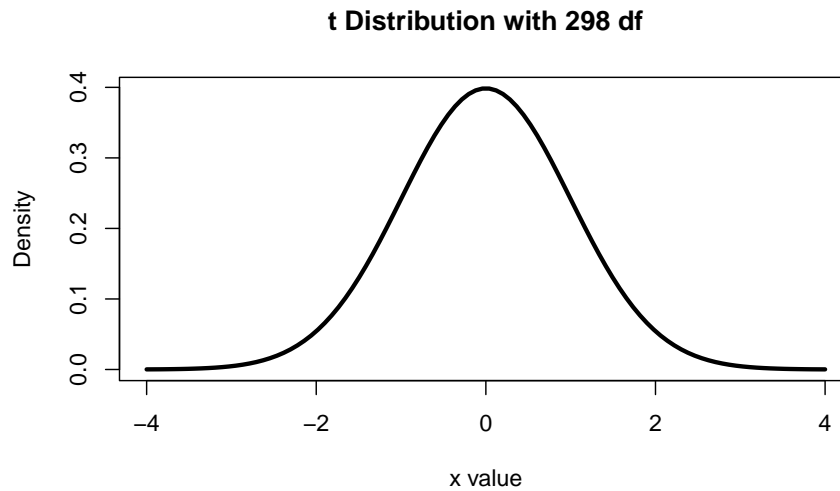


Figure 5.1: t-distribution with 602 degrees of freedom

6. Calculate the 99% confidence interval using theory-based methods.

7. Interpret the confidence interval in context of the study.

Types of Errors:

8. What type of error may have occurred for this study?
9. Interpret this error in context of the study.

5.3 Module 8 - 9 Review

```
allergy <- read.csv("https://math.montana.edu/courses/s216/data/PeanutAllergy.csv")
allergy %>% group_by(Treatment) %>% count(Allergy)
```

```
#> # A tibble: 4 x 3
#> # Groups:   Treatment [2]
#>   Treatment Allergy     n
#>   <chr>      <chr>   <int>
#> 1 Avoiders  No       220
#> 2 Avoiders  Yes       35
#> 3 Peanuts   No       240
#> 4 Peanuts   Yes        5
```

In the last 10 years, the proportion of children who are allergic to peanuts has doubled in Western countries. However, the allergy is not very common in some other countries where peanut protein is an important part of peoples' diets. The LEAP randomized trial, reported by Du Toit, et.al in the New England Journal of Medicine in February 2015 identified over 500 children ages 4 to 10 months who showed some sensitivity to peanut protein. They randomly assigned them to two groups:

- Peanut avoiders: parents were told to not give their kids any food which contained peanuts
- Peanut eaters: parents were given a snack containing peanut protein and told to feed it to their child several times per week (target dose was at least 6g of peanut protein per week).

At age 5 years, children were tested with a standard skin prick to see if they had an allergic reaction to peanut protein (yes or no). Is there evidence that exposure to peanuts reduces the likelihood of developing peanut allergies?

	Peanut Avoiders	Peanut Eaters	Total
Allergy	35	5	40
No Allergy	220	240	460
Total	255	245	500

For this study we will use the order of subtraction avoiders – eaters.

1. Fill in the blanks with one answer from each set of parentheses:

The variable whether or not a child is given peanut protein is the _____ (explanatory/response) variable and it is _____ (categorical/quantitative).

The variable whether or not a child developed a peanut allergy is the _____ (explanatory/response) variable and it is _____ (categorical/quantitative).

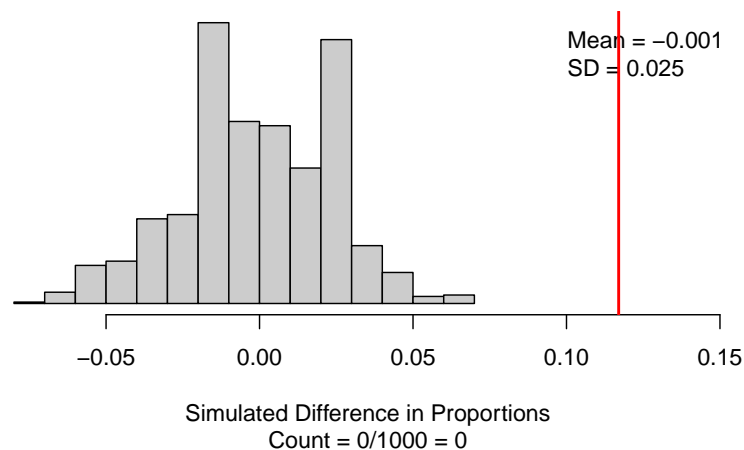
2. Write the parameter of interest for this study.

3. Write the null hypothesis in notation.

4. Write the alternative hypothesis in words.
5. Calculate the conditional proportion of children that developed a peanut allergy among those that avoided peanuts. Use proper notation.
6. Calculate the conditional proportion of children that developed a peanut allergy among those that ate peanuts. Use proper notation.
7. Calculate the difference in proportion of children that developed a peanut allergy for those that avoided peanuts and those who ate peanuts. Use proper notation.
8. First, let's think about how one simulation would be created on the null distribution using cards.
How many cards would you need?

What would be written on each card?
9. Next, we would mix the cards together and shuffle into two piles. How many cards would be in each pile?
What would each pile represent?
10. Once we have one simulated sample, what would we calculate and plot on the null distribution? *Hint:*
What statistic are we calculating from the data?

```
two_proportion_test(formula = Allergy ~ Treatment, #response~explanatory
  data=allergy, #name of dataset
  first_in_subtraction = "Avoiders", #order of subtraction: avoiders - peanuts
  number_repetitions = 1000, #always use a minimum of 1000 repetitions
  response_value_numerator = "Yes", #define a success as having an allergy
  as_extreme_as = 0.117, #type your calculated observed statistic (difference in sample
  direction="greater") #type your selected direction to match the alternative hypothesis
```

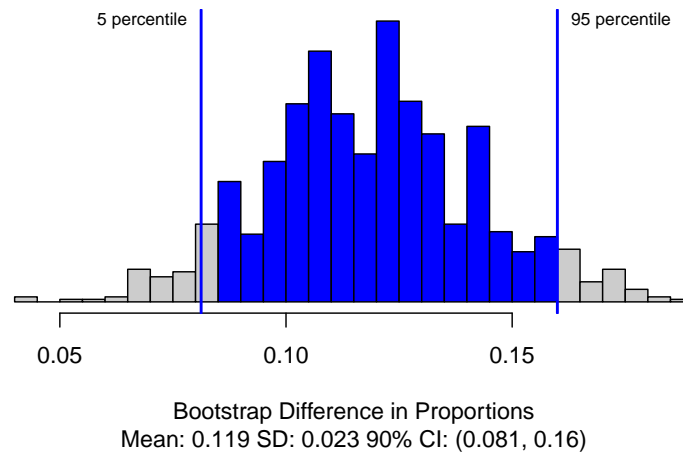


11. Interpret the p-value in context of the problem:

12. Write a conclusion to the test in context of the study.

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 `allergy`), the outcome for the explanatory variable that is first in subtraction, number of repetitions, the outcome for the response variable that is a success (what the numerator counts when calculating a sample proportion), and the confidence level as a decimal.

```
two_proportion_bootstrap_CI(formula = Allergy~Treatment,
  data=allergy, # Name of data set
  first_in_subtraction = "Avoiders", # Order of subtraction: enter the name of Group 1
  response_value_numerator = "Yes", # Define which outcome is a success
  number_repetitions = 1000, # Always use a minimum of 1000 repetitions
  confidence_level = 0.90) # Enter the level of confidence as a decimal
```



13. Interpret the 90% confidence interval in context of the problem.

Theory-based Methods

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

- **Independence:** The data are independent within and between the two groups. (*Remember:* This also must be true to use simulation methods!)
- **Success-failure condition:** This condition is met if we have at least 10 successes and 10 failures in each sample. Equivalently, we check that all cells in the table have at least 10 observations.

14. Are the conditions met to use theory-based methods?

To calculate the standardized statistic we use:

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - \text{null value}}{SE_0(\hat{p}_1 - \hat{p}_2)},$$

where the null standard error is calculated using the pooled proportion of successes:

$$SE_0(\hat{p}_1 - \hat{p}_2) = \sqrt{\hat{p}_{pool} \times (1 - \hat{p}_{pool}) \times \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}.$$

15. Calculate the null standard error of the difference in proportion.

16. Calculate the standardized statistic.

17. Interpret the standardized statistic in context of the problem.

```
pnorm(4.814, lower.tail = FALSE)
#> [1] 7.39694e-07
```

$$\hat{p}_1 - \hat{p}_2 \pm z^* \times SE(\hat{p}_1 - \hat{p}_2), \text{ where}$$

$$SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1 \times (1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2 \times (1 - \hat{p}_2)}{n_2}}$$

18. Calculate the standard error of the difference in proportions to calculate the confidence interval.

```
qnorm(0.90, lower.tail = TRUE)
#> [1] 1.281552
```

19. Calculate the 90% confidence interval.

20. What is the scope of inference for this study?

5.4 Key Topics Exam 2

Descriptive statistics and study design

1. Identify the observational units.
2. Identify the types of variables (categorical or quantitative).
3. Identify the explanatory variable (if present) and the response variable (roles of variables).
4. Identify the appropriate type of graph and summary measure.
5. Identify the study design (observational study or randomized experiment).
6. Identify the sampling method and potential types of sampling bias (non-response, response, selection).
7. Calculate and interpret the difference in proportions, relative risk, and percent increase/decrease in risk for a study involving two categorical variables.

Hypothesis testing

8. Identify which of the two scenarios applies to the study: one quantitative variable or two categorical variables.
9. Write the parameter of interest in words and correct notation.
10. Find the value of the observed statistic (point estimate, summary statistic). Use correct notation.
11. State the null and alternative hypotheses in words and in correct notation.
12. Verify the validity condition is met to use simulation-based methods to find a p-value.
13. Verify the validity conditions are met to use theory-based methods to find a p-value from the theoretical distribution.
14. In a simulation-based hypothesis test, describe how to create one dot on a dotplot of the null distribution using coins, cards, or spinners.
15. Explain where the null distribution is centered and why.
16. Describe and illustrate how R calculates the p-value for a simulation-based test.
17. Describe and illustrate how R calculates the p-value for a theory-based test.
18. Type of theoretical distribution (standard normal distribution or t-distribution with appropriate degrees of freedom) used to model the standardized statistic in a theory-based hypothesis test.
19. Calculate and interpret the standard error of the statistic under the null using the correct formula on the Golden ticket.
20. Calculate and interpret the appropriate standardized statistic using the correct formula on the Golden ticket.
21. Interpret the p-value in context of the study: it is the probability of _____, assuming _____.
22. Evaluate the p-value for strength of evidence against the null: how much evidence does the p-value provide against the null?
23. Write a conclusion about the research question based on the p-value.
24. Given a significance level, what decision can be made about the research question based on the p-value.
25. Describe which features of the study could be changed to increase power and how.
26. Describe which features of the study impact the p-value and how.

27. Write a Type I error in context of the problem.
28. Write a Type II error in context of the problem.
29. Interpret power in context of the problem.
30. Based on your p-value, identify what type of error could have occurred.

Confidence interval

31. Describe how to simulate one bootstrapped sample using cards.
32. Explain where the bootstrap distribution is centered and why.
33. Find an appropriate percentile confidence interval using a bootstrap distribution from R output.
34. Verify the validity condition is met to use simulation-based methods to find the confidence interval.
35. Verify the validity conditions are met to use theory-based methods to calculate a confidence interval.
36. Describe and illustrate how the bootstrap distribution is used to find the confidence interval for a given confidence level.
37. Describe and illustrate how the standard normal distribution or t-distribution is used to find the multiplier for a given confidence level.
38. Calculate and interpret the standard error of the statistic (not assuming the null hypothesis) using the correct formula on the Golden ticket
39. Calculate the appropriate margin of error and confidence interval using theory-based methods.
40. Interpret the confidence interval in context of the study.
41. Based on the interval, what decision can you make about the null hypothesis? Does the confidence interval agree with the results of the hypothesis test? Justify your answer.
42. Interpret the confidence level in context of the study. What does “confidence” mean?
43. Describe which features of the study have an effect on the width of the confidence interval and how.

References

- “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 Injuries 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., D. I. 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).
- 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-Confirmed 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 Healthy 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).