STAT 216 Coursepack

Spring 2025
Montana State University

Melinda Yager
Jade Schmidt
Stacey Hancock

Contents

Preface

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

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

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

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

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

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

1.1 Vocabulary Review and Key Topics

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

1.1.1 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 that 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.
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.

Inferences to population can be made

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

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

Read Sections 2.2 - 2.4, 15.1, 15.2, Chapter 4 and Chapter 16 in the course textbook. Use the following videos to complete the video notes for Module 8.

1.2.1 Course Videos

- 2.2to2.4
- 4.1_TwoProp
- 4.2_TwoProp
- 4.4
- 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?

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.
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.

Inferences to population can be made

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

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?

Summarizing two categorical variables - Video 4.1_TwoProp

- The summary measure for two categorical variables is the _____ in _____.

Notation used for the population difference in proportion:

- Two categorical variables:

    - Subscripts represent the _____ variable groups

Notation used for the sample difference in proportion:

- Two categorical variables

When we have two categorical variables we report the data in a _____ or two-way table with the _____ variable on the columns and the _____ variable on the rows.

For today's videos we will again use the `moving_to_mt` data set.

Example from the Video: Gallatin Valley is the fastest growing county in Montana. You'll often hear Bozeman residents complaining about the 'out-of-staters' moving in. A local real estate agent recorded data on a random sample of 100 home sales over the last year at her company and noted where the buyers were moving from as well as the age of the person or average age of a couple buying a home. The variable age was binned into two categories, "Under30" and "Over30." Additionally, the variable, state the buyers were moving from, was created as a binary variable, "Out" for a location out of state and "In" for a location in state.

The following code reads in the data set, `moving_to_mt` and names the object moving.

```
moving <- read.csv("data/moving_to_mt.csv")
```

To look at the relationship between the variable, `Age_Group` and the variable, `From` create the following two-way table using the `R` output below. Note, we are using `From` as the explanatory variable to predict whether a home sale has a buyer that is over or under the age of 30.

```
moving %>%
    group_by(Age_Group) %>% count(From) %>% print(n=8)
```

```
#> # A tibble: 8 x 3
#> # Groups:   Age_Group [2]
#>   Age_Group From      n
#>   <chr>     <chr> <int>
#> 1 Over30    CA        6
#> 2 Over30    CO        2
#> 3 Over30    MT       47
#> 4 Over30    WA       10
#> 5 Under30   CA        6
#> 6 Under30   CO        6
#> 7 Under30   MT       14
#> 8 Under30   WA        9
```

	State				
Age Group	CA	CO	MT	WA	Total
Over30	6	2	47	10	65
Under30	6	6	14	9	35
Total	12	8	61	19	100

- Using the table above, how many of the sampled home sales have buyers who were under 30 years old and from Montana?

If we want to know what proportion of each age group is from each state, we would calculate the proportion of home sales with buyers from each state within each age group. In other words, divide the number of home sales from each state with buyers that are over 30 by the total for row 1, the total number of home sales with buyers over 30.

- What proportion of sampled home sales with buyers under 30-years-old were from California?

- What notation should be used for this value?

Additionally, we could find the proportion of home sales with buyers in each state for each age group. Here we would calculate the proportion of home sales with buyers in each age group within each state. Divide the number of home sales with buyers in each age group from CA by the total for column 1, the total number of home sales with buyers from CA.

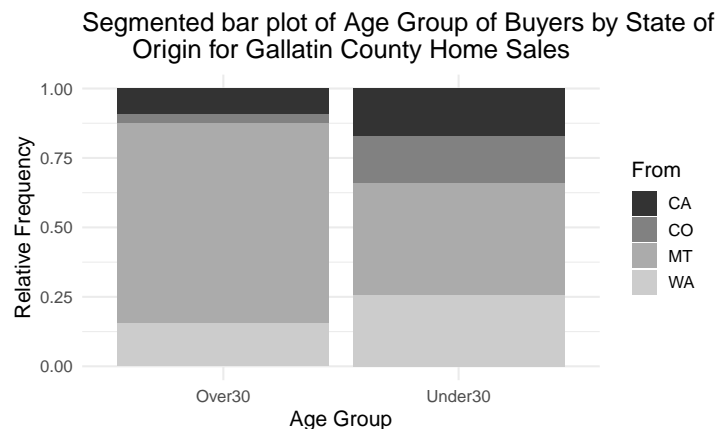|	State				
Age Group	CA	CO	MT	WA	Total
Over30	6	2	47	10	65
Under30	6	6	14	9	35
Total	12	8	61	19	100

- Using the table, calculate the proportion of home sales in Gallatin County with in-state buyers who are over 30 years old? Use appropriate notation with informative subscripts.

- Using the table, calculate the proportion of home sales in Gallatin County with California buyers who are over 30 years old? Use appropriate notation with informative subscripts.

- Calculate the difference in proportion of home sales in Gallatin County over 30 years old from other parts of Montana and from California. Use MT - CA as the order of subtraction. Give appropriate notation.

- Interpret the difference in proportion in context of the study.

## Plots for two categorical variables - Video 4.2_TwoProp

In a segmented bar plot, the bar for each category will sum to 1. In this first plot, we are plotting the row proportions calculated conditional on the age group.

```
moving %>%
  ggplot(aes(x = Age_Group, fill = From))+ #Enter the variables to plot
  geom_bar(stat = "count", position = "fill") +
  labs(title = "Segmented bar plot of Age Group of Buyers by State of
      Origin for Gallatin County Home Sales",
      #Title your plot
      y = "Relative Frequency", #y-axis label
      x = "Age Group") + #x-axis label
  scale_fill_grey()
```


In this second plot, we are plotting the column proportions calculated conditional on the state of origin for the buyer.

```
moving %>%
  ggplot(aes(x = From , fill = Age_Group))+ #Enter variables to plot
  geom_bar(stat = "count", position = "fill") +
  labs(title = "Segmented bar plot of State of Origin of Buyers by Age
      Group for Gallatin County Home Sales",
      #Title your plot
      y = "Relative Frequency", #y-axis label
      x = "State of Origin") + #x-axis label
  scale_fill_grey()
```

Segmented bar plot of State of Origin of Buyers by Age
Group for Gallatin County Home Sales
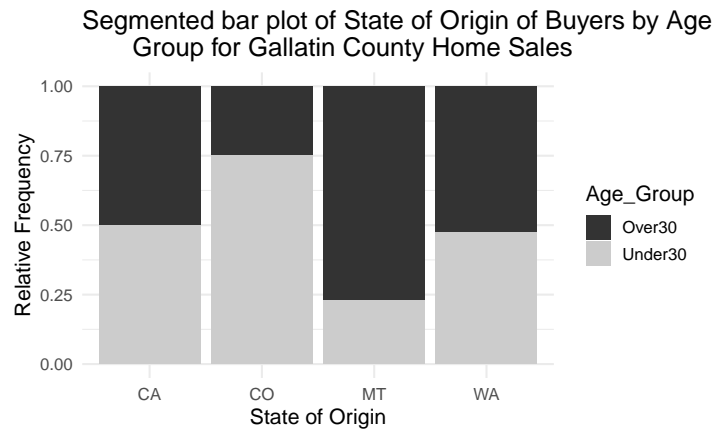
Mosaic plot:

```
moving$Age_Group <- factor(moving$Age_Group, levels = c("Under30", "Over30"))
moving %>% # Data set piped into...
  ggplot() +    # This specifies the variables
  geom_mosaic(aes(x=product(From), fill = Age_Group)) +
    # Tell it to make a mosaic plot
  labs(title = "Mosaic plot of State of Origin Segmented by
  Age Group for Gallatin County Home Sales",
      # Title your plot
      x = "State of Origin",   # Label the x axis
      y = "") +  # Remove y axis label
    scale_fill_grey(guide = guide_legend(reverse = TRUE)) # Make figure color
```

Mosaic plot of State of Origin Segmented by
Age Group for Gallatin County Home Sales
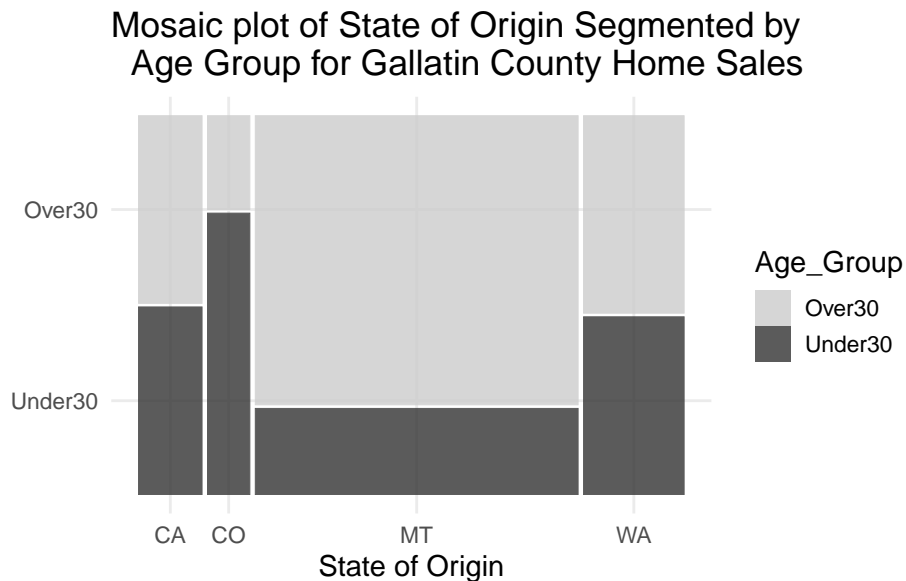
- Why is the bar for MT the widest on the mosaic plot?

Simpson's paradox - Video 4.4

- When an apparent _____ between explanatory and response variables reverses when accounting for _____ variable.

Example: The "Berkeley Dataset" contains all 12,763 applicants to UC-Berkeley's graduate programs in Fall 1973. This dataset was published by UC Berkeley researchers in an analysis to understand the possible gender bias in admissions and has now become a classic example of Simpson's Paradox.

```r
discrim <- read.csv ("data/berkeley.csv")

discrim %>%
  ggplot(aes(x =Gender, fill = Admission))+
  geom_bar(stat = "count", position = "fill") +
  labs(title = "Segmented bar plot of Sex of Berkley Applicants by
       Admission Status",
       y = "Relative Frequency",
       x = "Sex") +
  scale_fill_grey()
```
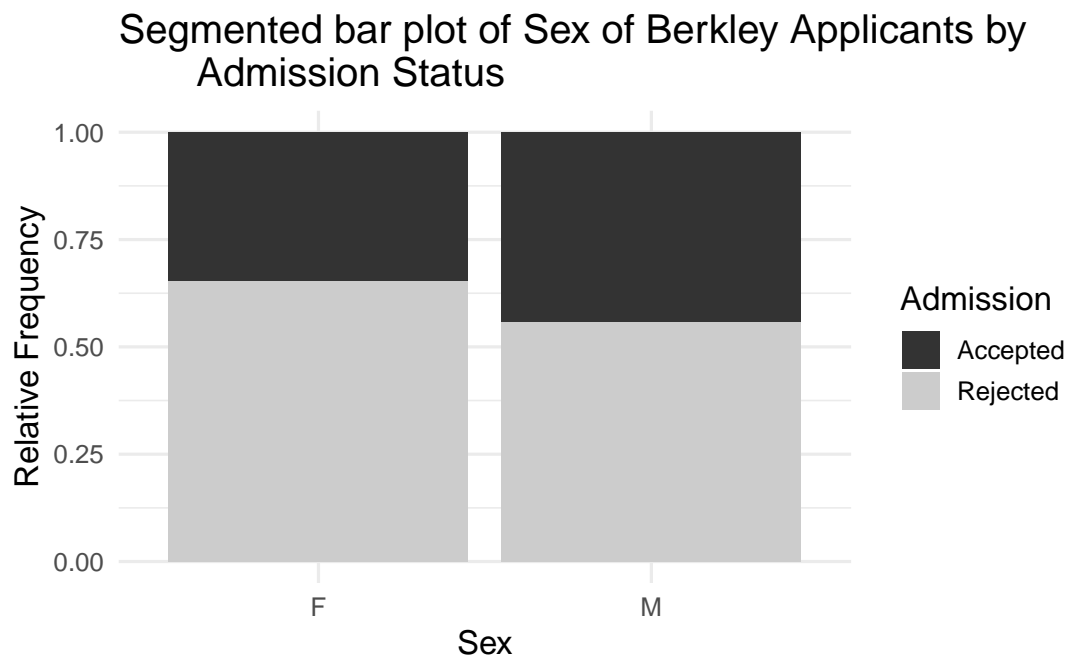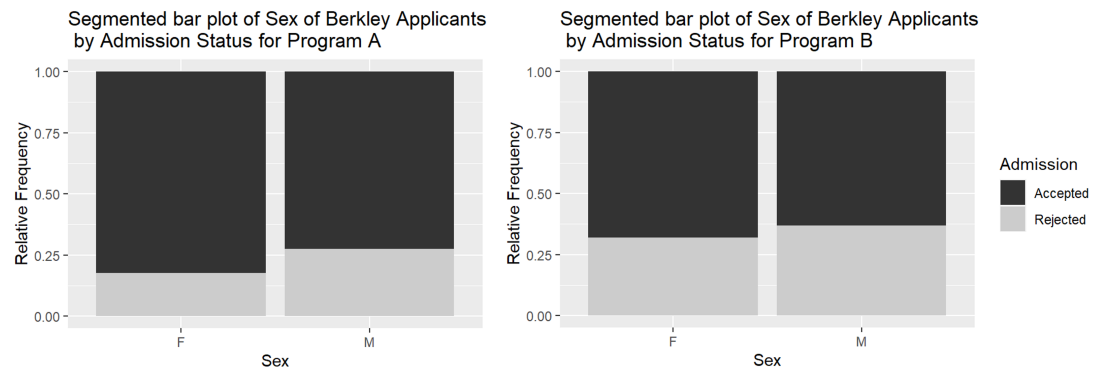

The data showed that 44% of male applicants were accepted and 35% of female applicants were accepted. Does it appear that the female students are discriminated against?

We can break down the data by major. A major code (either A, B, C, D, E, F, or Other) was used.
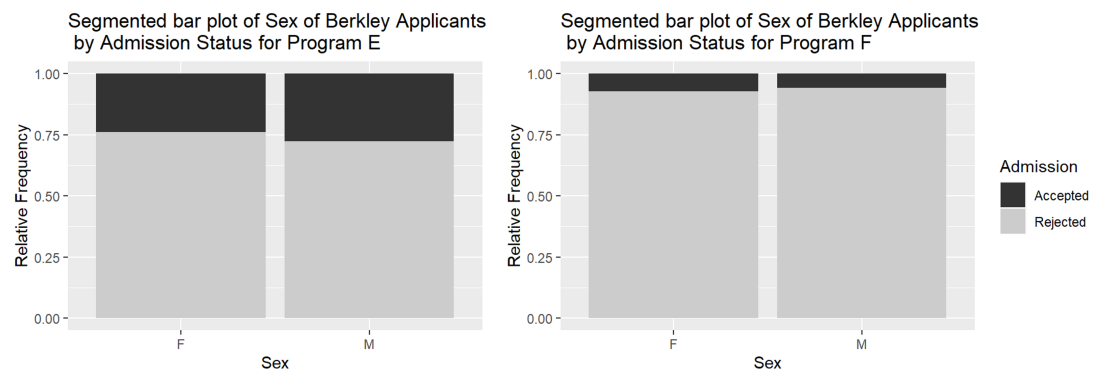
Here we look at the relationship between admission status and sex for Program A and for Program B.

Showing Program C and Program D.
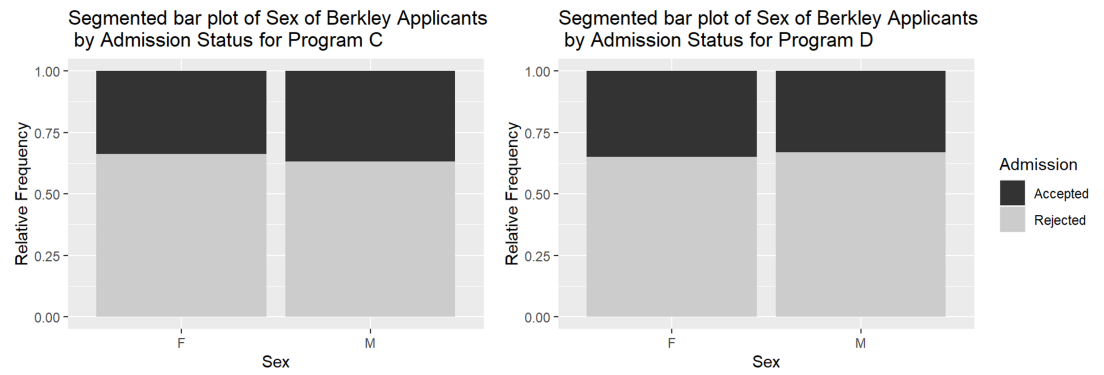
And finally, Program E and F.

We can see in several programs the acceptance rate is higher for females than for males.

Two categorical variables - Video 15.1

- In this module, we will study inference for a _____ explanatory variable and a _____ response.

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()
```
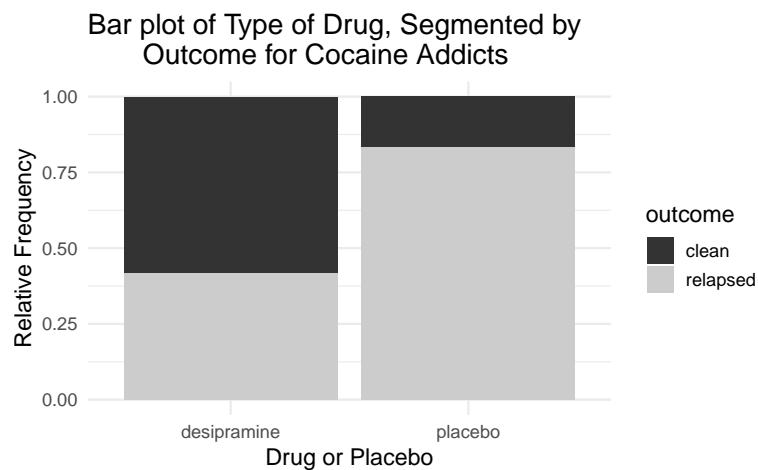
Is the independence condition met for simulation inference?

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 = 10000, # 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")
```


Simulated Difference in Proportions
Count = 2/10000 = 2e−04
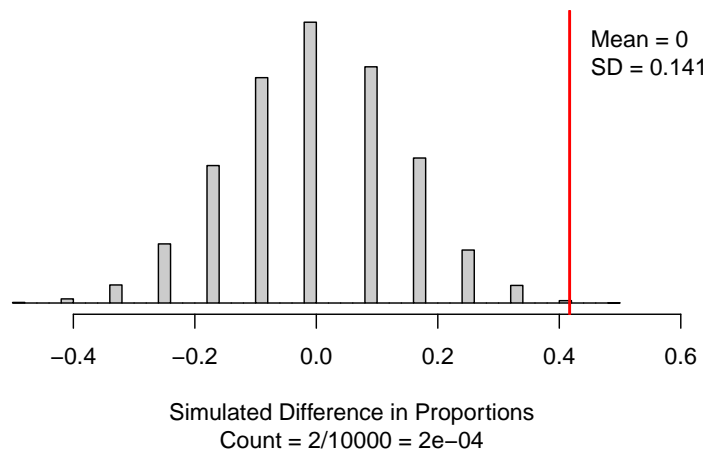
Explain why the null distribution is centered at the value of zero:

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 = 10000, # Always use a minimum of 1000 repetitions
        confidence_level = 0.99) # Enter the level of confidence as a decimal
```


Bootstrap Difference in Proportions
Mean: 0.414 SD: 0.125 99% CI: (0.083, 0.708)
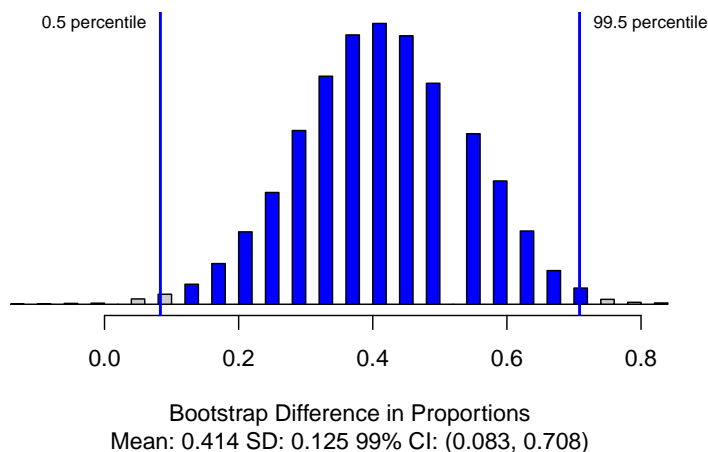
Confidence interval interpretation:

- 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 RelativeRisk

- 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 :

1.2.2 Concept Check

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

1. Explain why the null distribution is centered at the value of zero.

2. Does the confidence interval agree with the p-value?

3. What is the difference between a mosaic plot and a segmented bar plot?

4. What does relative risk measure?

1.3 Activity 16: Study Design

1.3.1 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.

1.3.2 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.

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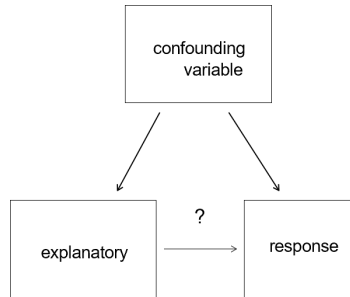
11. What is the purpose of random assignment of the subjects in a study to the explanatory variable groups? Cross out the arrow in the figure above that is eliminated by random assignment.

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.

24

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.
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.

Inferences to population can be made

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

1.3.4 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:

1.3.5   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.

1.3.6   Additional notes

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

1.4 Activity 17: Summarizing Two Categorical Variables

1.4.1 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

1.4.2 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.

1.4.3 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 the file to the RStudio server

- Open the RScript file

Highlight and run lines 1–2 to load the packages needed for today's activity and load the data set. 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 5 in the R script file) reads in the data set and names the data set `myopia`. Highlight and run line 5 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 9 in the R script file, highlight and run line 9 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 14. Highlight and run lines 13–20.

```
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_grey()  # Make figure black and weight
```

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.
```
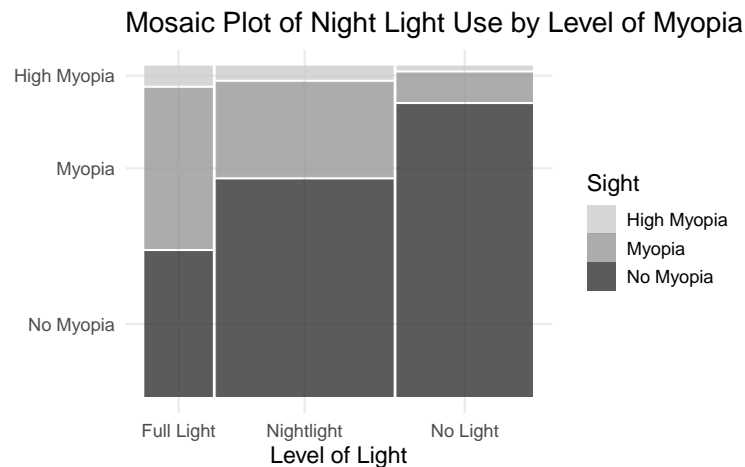

12. What is similar and what is different between the segmented bar chart and the mosaic bar chart?

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.

1.4.4 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.

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 18: The Good Samaritan

1.5.1 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

1.5.2 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.

1.5.3 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 1–7 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>
```

32

```
#> 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 proporiton 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 $\neq$):

$$H_A : \pi_{\text{hurry}} - \pi_{\text{no hurry}} \underline{\hspace{3cm}} 0$$

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 13–20

```
good %>%
  ggplot(aes(x = explanatory, fill = response))+ #Enter the variables to plot
  geom_bar(stat = "count", position = "fill") +
  labs(title = "Segmented Bar Plot of Princeton Seminary Students that Help the actor \nbetween those tha
       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.

7. 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**.

Segmented bar plot of Simulated Data
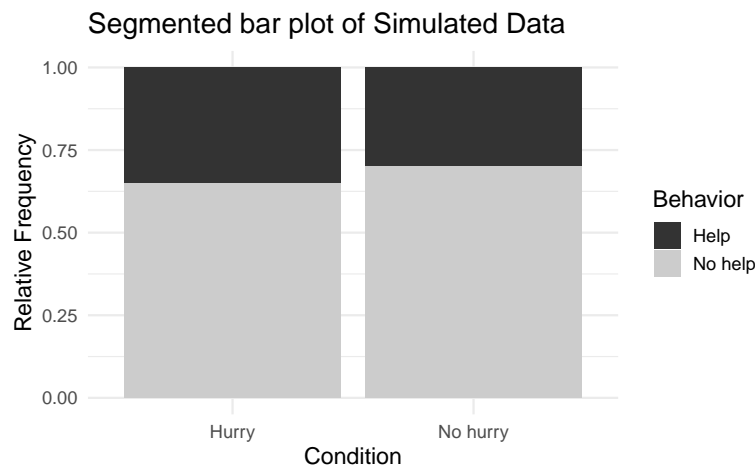
To create the null distribution of differences in sample proportions, we will use the `two_proportion_test()` function in R (in the `catstats` package). We will need to enter the response variable name and the explanatory variable name for the formula, the data set name (identified above as `good`), 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 8 in place of the `xx`'s to produce the null distribution with 1000 simulations; highlight and run lines 24–30.

```
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.

1.5.4   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.

1.5.5   Additional notes

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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