Review of Module 8

Vocabulary Review

Point estimate: this is another word for the observed statistic

- For two categorical variables this will be the observed difference in sample proportion
- Since this value is calculated from the sample the notation used is $\hat{p}_1 - \hat{p}_2$

Example Myopia:

- $\hat{p}_{full} - \hat{p}_{no} = 0.447 - 0.901 = -0.454$

Example Good Samaritan Study:

- $\hat{p}_{hurry} - \hat{p}_{no\ hurry} = \frac{2}{20} - \frac{11}{20} = -0.45$

Interpreting the sample difference in proportion:

Include:

- Summary measure (difference in proportion)
  - What numerical value are we calculating?
  - This is dependent on the type of variable(s) in our study
  - Give the value of the statistic and the order of subtraction
- Context
  - Observational units/cases – what or whom are we collecting data on
  - Variable of interest – what success are we focusing on in the research question.
  - ***For comparison studies we also need the explanatory variable groups.***
- *Note: the statistic is calculated from the sample so DO NOT use the word true in the interpretation.*

Example Good Samaritan Study:

- **The proportion of Princeton Theological Seminary students who helped the actor in the hurry condition is 0.45 lower than those in the no hurry condition.**

Example Myopia:

- **The proportion of children with no myopia who slept in full light is 0.454 lower than the proportion who slept with no light.**

Type of plots

- **Segmented bar plot**
  - The explanatory variable is on the x-axis and the plot is segmented by the response variable
- **Mosaic plot**
  - The width of bars depend on the sample size in each group

Review of Module 8

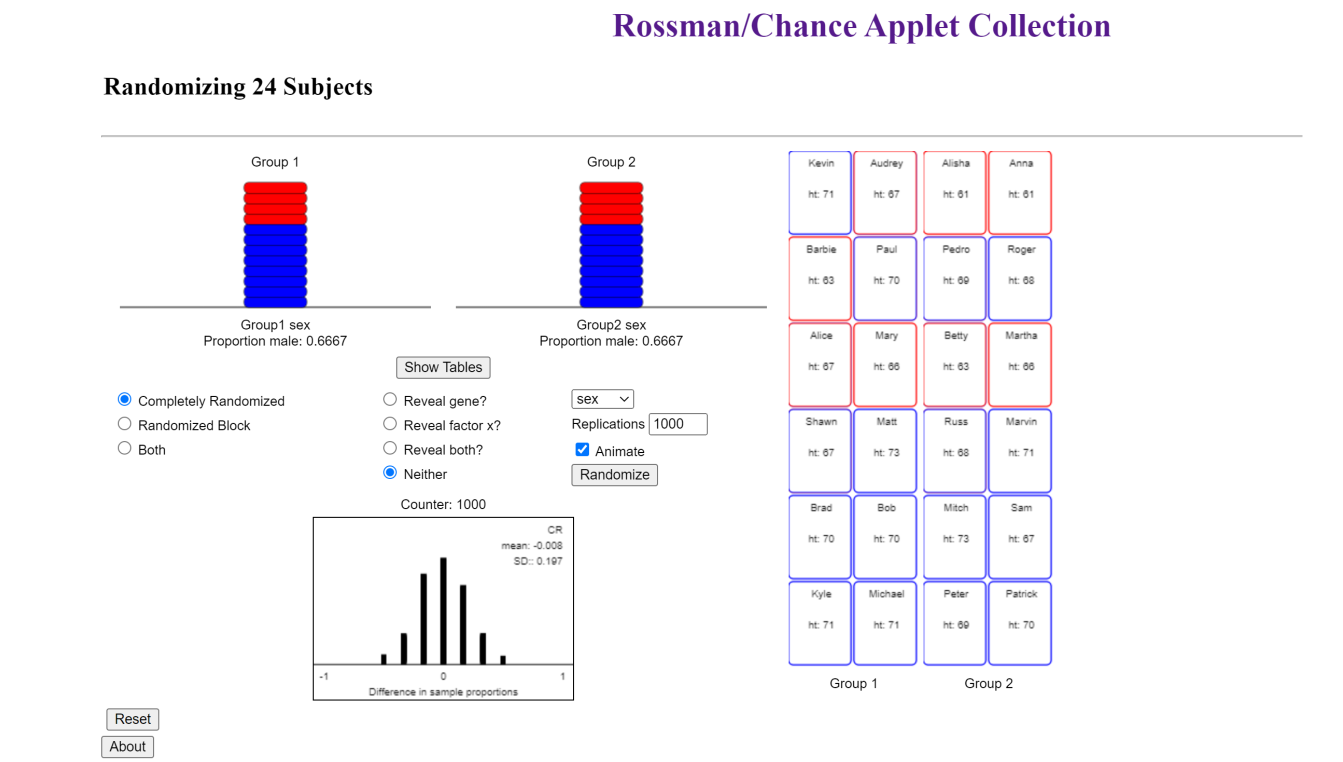
- **Roles of Variables**:
  - **Explanatory variable:** the variable that *potentially* "explains" the change in the response variable
    - The groups we are comparing
    - In an experiment this is the variable that the researcher manipulates
  - **Response variable:** the variable that is *potentially* impacted or changed by the explanatory variable
    - The data collected
    - This variable cannot be manipulated by the researcher
  - **Confounding variable**: a variable related **both** to the ***explanatory variable*** and the ***response variable*** so that the effects on the response variable cannot be separated from the effects of the explanatory variable
- **Random sampling:** each observational unit in the population has the same chance of being selected for the sample
- **Random assignment:** observational units are assigned at random to the explanatory variable groups; also called **randomization**
- **Study Design:**
  - **Observational Study –** observational units are categorized into naturally formed explanatory variable groups
  - **Randomized Experiment –** actively create explanatory variable groups by randomization to compare
- **Scope of Inference:**
  - **Generalization**
    1. How was the sample selected?
       a. Randomly without sampling bias
          i. Can generalize the results of the study to the target population
       b. Non-random sample or a sample with sampling bias
          i. The results of the study only apply to observational units similar to those in the sample

Causation

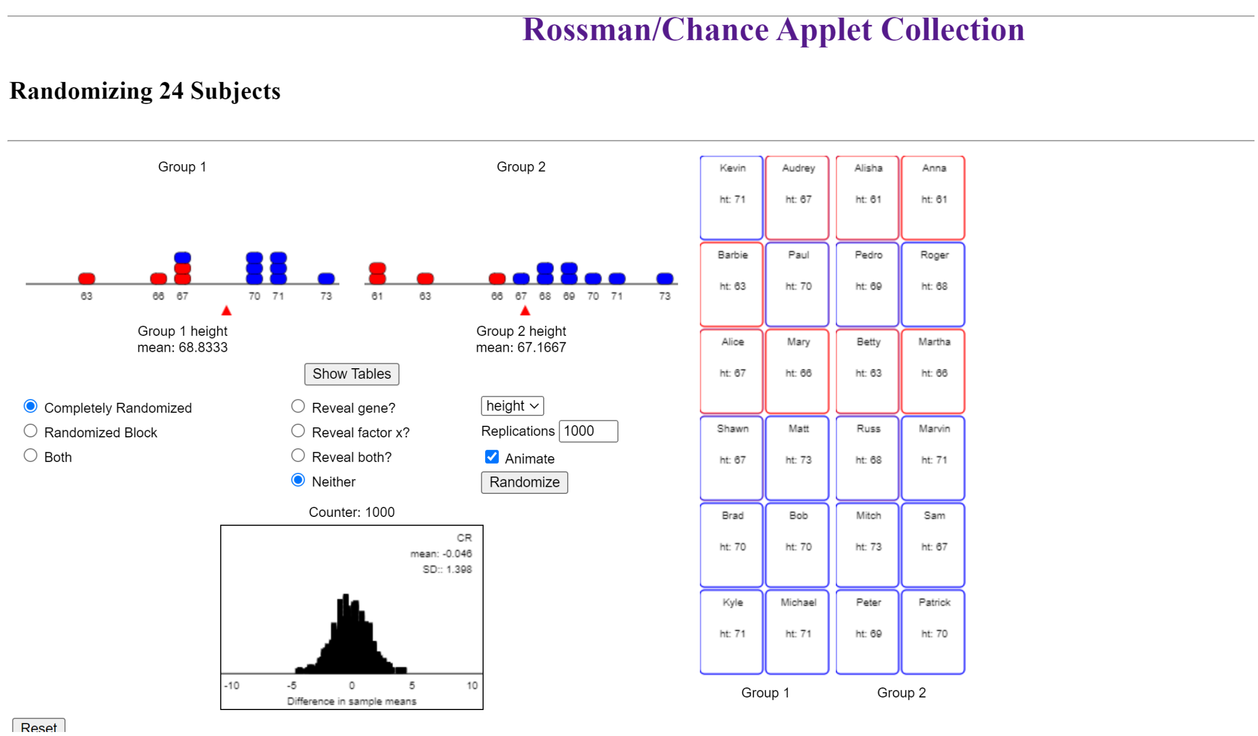
    2. Study Design
       a. Randomized Experiment
          i. Can discuss a causal relationship between the explanatory and response variables since the potential confounding variables are evened out between explanatory variable groups
       b. Observational Study
          i. Can only discuss an association between variables since there are potential confounding variables

Activity 13:

- In this activity we investigated study design (observational study vs. randomized experiment) and how this impacts scope of inference.
- In the 1st part of the activity, we studied several different potential confounding variables in the Atrial Fibrillation study. In this study, the explanatory variable is whether the patient received the drug or the placebo and the response variable is whether the patient's heart rate stabilized.
- When we randomized (randomly assigned) the participants in the Rossman Chance applet we saw that, on average, the difference in proportion of males is zero between the two groups.



- Not every sample will result in a difference in proportion equal to zero but as we can see by the plot, on average, the difference in proportion of males is equal to zero.

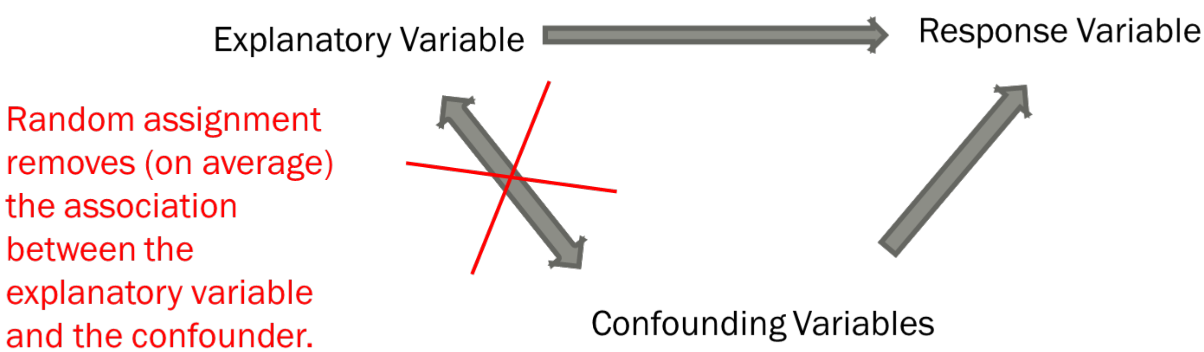


- The same is seen with height. If we randomly assign the subjects to groups, on average, the difference in mean height is zero.
- When we randomly assign observational units to the explanatory variable groups, the confounding variables are evened out, on average, between the two explanatory variable groups eliminating the

relationship between the potentially confounding variable and the explanatory variable. When random assignment is used it can be said that the explanatory variable *caused* the change in the response variable.

		Assignment of Explanatory Variable		
		Random allocation of explanatory variable	Individual decides explanatory variable (non-random)	
Selection of Observational Units from the Population	Random sample	The observational units are randomly selected from the population; then the explanatory variable (treatment) is randomly assigned.	The observational units are randomly selected from the population, but the value of the explanatory variable is not randomly assigned by the researcher.	Conclusions generalize directly to the population.
	Other sampling method (non-random)	The observational units are observed (somehow!) and then randomly allocated to the levels of the explanatory variable.	The observational units are observed (somehow!) and the value of the explanatory variable is not randomly assigned by the researcher.	Conclusions might not be generalizable because of volunteer bias.
		Significant conclusions are considered to be cause and effect.	Significant conclusions must be framed with possible confounding variables.	

- Scope of inference has two parts: generalization (to which group of observational units can the results be generalized to) and causation (can a causal relationship exist between the explanatory and response variable)



- In *observational studies* because of the presence of confounding variables, we *cannot conclude a causal relationship between the explanatory variable and the response variable*. We are unable to distinguish which variable (the explanatory or the confounding variables) is affecting the response variable.
- In *randomized experiments* we *can conclude a causal relationship between the explanatory variable and the response variable* because the confounding variables are evened out between explanatory variable groups, on average.
- In the atrial fibrillation study, the participants are volunteers but they do use random assignment. Since random assignment is used this is a randomized experiment and the researchers can conclude that the new drug *caused* the increased chance of stabilization. However, since the participants are volunteers, selection bias exists in this study so the results only apply to patients with atrial fibrillation similar to those in this study.

Review of Module 8

Questions 12/13

- The last two questions are very similar as they have the same explanatory variable and response variable. In both cases we are trying to see if there is a difference in proportion of COVID cases for those that are immunized and those that are not.
- The study designs however are different. The study in question 12 is a randomized experiment (volunteers are randomly assigned to either receive the vaccine or the placebo) while question 13 is an observational study (participants were not assigned to the explanatory variable groups).
- The sampling method also differs between the two studies. In question 12, the sample consists of volunteers while in question 13, the researchers took a random sample.
- Based on the described sampling method and study design, assuming statistically significant results are found, the scope of inference for question 12 is researchers could conclude a causal relationship exists between the variables for US adults without COVID-19 similar to those in the study.
- Based on the described sampling method and study design, assuming statistically significant results are found, the scope of inference for question 13 is researchers could conclude a non-causal relationship exists between the variables for all US adults without COVID-19.

Parameter of Interest: What information do we want to know about the population?:

- The parameter of interest is used in the hypotheses statements, in conclusions, and in many interpretations!

Include:

- Reference of the population (true, long-run, population, all)
  - Clearly refer to the population
- Summary measure (difference in proportion)
  - What numerical value are we calculating?
  - This is dependent on the type of variable(s) in our study
- Context
  - Observational units/cases – what or whom are we collecting data on
  - Variable of interest – what success are we focusing on in the research question.
    - ***For comparison studies we also need the explanatory variable groups.***

Example Good Samaritan Study:

- ***The difference in true proportion of Princeton Theological Seminary students who would help the actor in the hurry condition and in the no hurry condition (hurry – no hurry).***

Hypothesis Test (test of significance/inference): test to show evidence based on the sample statistic against the null hypothesis

Hypotheses:

- Null Hypothesis: This is the known claim that we are trying to disprove; may be based on random chance
  - For comparison studies we assume there is no difference between groups
  - H_0: $\pi_1 = \pi_2$ or H_0: $\pi_1 - \pi_2 = 0$

subscripts represent explanatory variable groups

- Alternative: this is the claim we are testing that is based on the research question
  - $H_a: \pi_1 \begin{Bmatrix} > \\ \neq \\ < \end{Bmatrix} \pi_2$ or $H_a: \pi_1 - \pi_2 \begin{Bmatrix} > \\ \neq \\ < \end{Bmatrix} 0$
    - The direction of the alternative (the sign) is determined by the research question

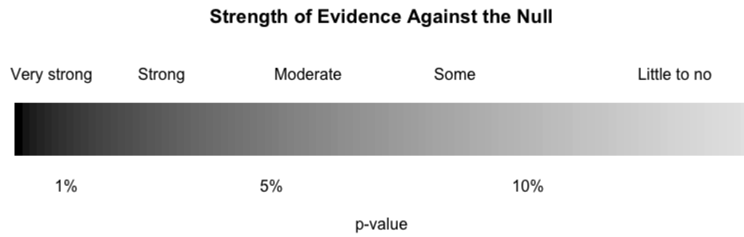
- **Example Good Samaritan Study:** Do these data provide evidence that those in a hurry will be less likely to help people in need in this situation?
  - *H₀: **There is no difference in true proportion of Princeton Theological Seminary students who helped the actor in the hurry condition and in the no hurry condition***
    - *Note: we are assuming that the proportion of seminary students who help in the hurry condition is equal to the proportion in the no hurry condition*
  - $H_o: \pi_{hurry} - \pi_{no\ hurry} = 0\ or\ H_o: \pi_{hurry} = \pi_{no\ hurry}$
  - *H_A: **The true proportion of Princeton Theological Seminary students who helped the actor in the hurry condition is less than in the no hurry condition***
    - *The direction of the alternative is less than because the research question asks for evidence that those in a hurry will be less likely to help people in need in this situation!*
  - $H_A: \pi_{hurry} - \pi_{no\ hurry} < 0\ or\ H_o: \pi_{hurry} < \pi_{no\ hurry}$

Null Distribution: simulation distribution created based on the assumption that the null hypothesis is true; centered at the null value

- **How can we use cards to simulate one sample for the null distribution?**
  - Label cards with the total number of successes and the total number of failures in the sample.
  - Mix the cards together and shuffle into two piles
    - 1 to represent group 1
    - 1 to represent group 2
  - Plot the difference in proportion of success between the groups from one simulation
- **Example Good Samaritan Study**: Label 13 cards with helped and 27 with did not help. Mix together and shuffle into two piles: 20 in hurry and 20 in no hurry. Plot the difference in proportion of seminary students that would help the actor in the hurry condition and the no hurry condition (hurry – no hurry) in one simulated sample.

Strength of Evidence: How much evidence does the p-value provide **against** the null?

- Use the guidelines for the strength of evidence



- The smaller the p-value the MORE evidence there is against the null hypothesis

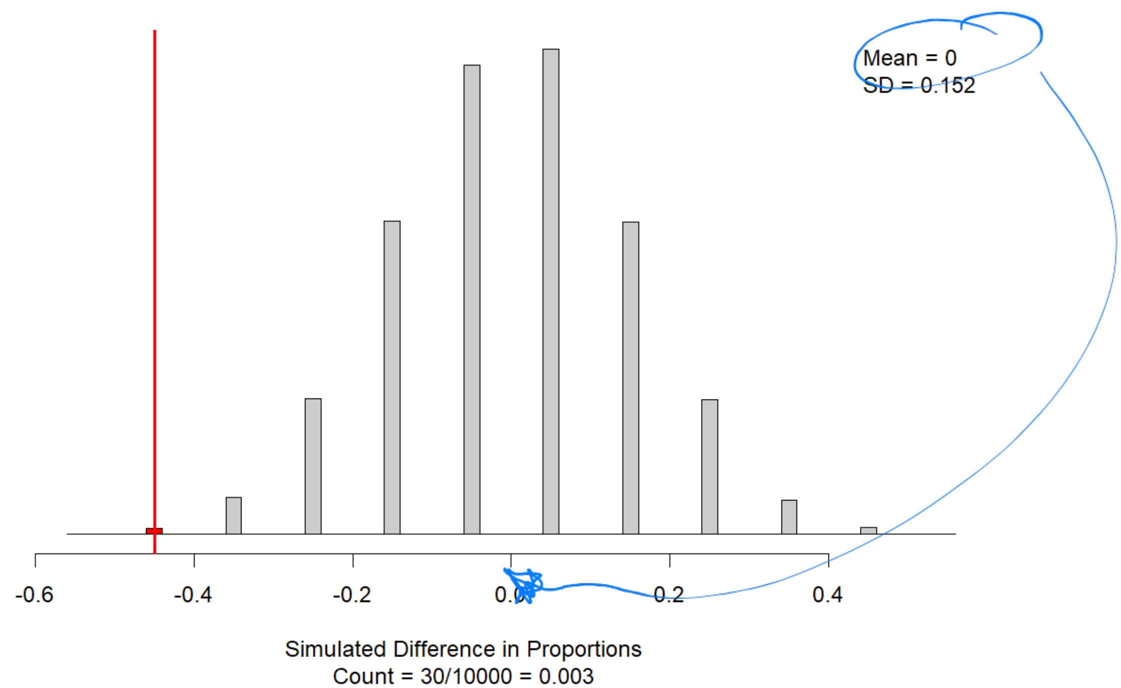
Finding the p-value: Count the number of simulated statistics as or more extreme than the statistic divided by the total number of simulated statistics

$$p - value = \frac{\#\ of\ simulations\ equal\ to\ \hat{p}_1 - \hat{p}_2\ and\ more\ extreme}{total\ \#\ of\ simulations}$$

See the code below for the two-proportion test for Good Samaritan Study:

```
two_proportion_test(formula = Behavior~Condition, # response ~ explanatory
          data = good, # Name of data set
          first_in_subtraction = "Hurry", # Order of subtraction: enter the name of Group 1
          number_repetitions = 10000, # Always use a minimum of 1000 repetitions
          response_value_numerator = "Help", # Define which outcome is a success
          as_extreme_as = -0.45, # Calculated observed statistic (difference in sample proportions)
          direction="less") #  Alternative hypothesis direction ("greater","less","two-sided")
```

Review of Module 8



There are FOUR things we ask about the p-value

Evaluation of a p-value:

- How much evidence does the p-value provide AGAINST the null hypothesis?
- **Example Good Samaritan Study:** The simulation p-value for this study was found to be 0.003 (in my simulation!).
    - ***There is very strong evidence against the null hypothesis that there is no difference in true proportion of Princeton Theological Seminary students that would help the actor in the hurry group and the no hurry group.***

Interpretation of a p-value:

- What the p-value measures: the probability of observing the sample statistic or more extreme if the null hypothesis is true (Don't forget the context!)
- Include in the interpretation:
    - Statement about probability (in x% of simulated samples, in x out of 1000 simulated samples, with a probability of x%)
    - Statistic in context (give the value and in words what the statistic represents)
    - more extreme (direction of the alternative)
    - If the null hypothesis is true in context (give the null value and in words what the null represents)
        - Note: context only needs to be included in either the statistic OR the null

Review of Module 8

Example Good Samaritan: The simulation p-value for this study was found to be 0.003 (in my simulation!).

- ***We would observe a sample difference in proportion of -0.45 (hurry-no hurry) or less with a probability of 0.003 if we assume there is no difference in true proportion of Princeton Theological Seminary students that would help the actor in the hurry group and the no hurry group.***

  OR

- ***If there is no difference in true proportion of Princeton Theological Seminary students that would help the actor in the hurry group and the no hurry group, we would observe a sample difference in proportion of -0.45 (hurry-no hurry) or less in 0.2% of simulated samples.***

  OR

- **In 2 out 1000 simulated samples**, ***we would observe a sample difference in proportion of -0.45 (hurry-no hurry) or less if we assume there is no difference in true proportion of Princeton Theological Seminary students that would help the actor in the hurry group and the no hurry group.***

Conclusion: Answers the research question. Write a conclusion as the amount of evidence in support of the alternative.

- **Example Good Samaritan Study**: Write a conclusion in context of the study.
  - ***There is very strong evidence*** *that* ***the long-run proportion of seminary students that would help the actor in the hurry group is lower than in the no hurry group.***

Decision: compare the p-value to the set significance level

- If the p-value is less than the significance level (α), the decision will be to reject the null hypothesis
- If the p-value is greater than the significance level (α), the decision will be to fail to reject the null hypothesis
- **Example Good Samaritan Study**:
  - Since we have a very small p-value less than the significance level we will reject the null hypothesis.