MA206, Lesson 12 - Causation

What does it mean if two variables are **associated**?

What is a **cause-and-effect** relationship? When can we conclude it?

Define:
Explanatory Variable

Response Variable

Confounding Variable

Observational Study

Experiment

Understand what can and cannot be inferred regarding random assignment and random sampling.

	By Random Assignment	No Random Assignment
By Random Sampling		
No Random Sampling		

1) Many studies have shown that women who smoke while pregnant tend to have babies who weigh significantly less at birth, on average, than women who do not smoke while pregnant.

a) Identify the explanatory variable in these studies. Classify it as categorical or quantitative.

b) Identify the response variable in these studies. Classify it as categorical or quantitative.

c) Socioeconomic status is a potential confounding variable. Explain what that means.

d) Draw the causal diagram with the above variables.

2) A group of 120 cadets from Thayer Hall were surveyed one afternoon about if they have ever pulled an all-nighter and what their current GPA is. It was found that cadets who claimed to never have an all-nighter had an average GPA of 3.1 while the average GPA for cadets who did claim to pull all-nighters was a 2.9.

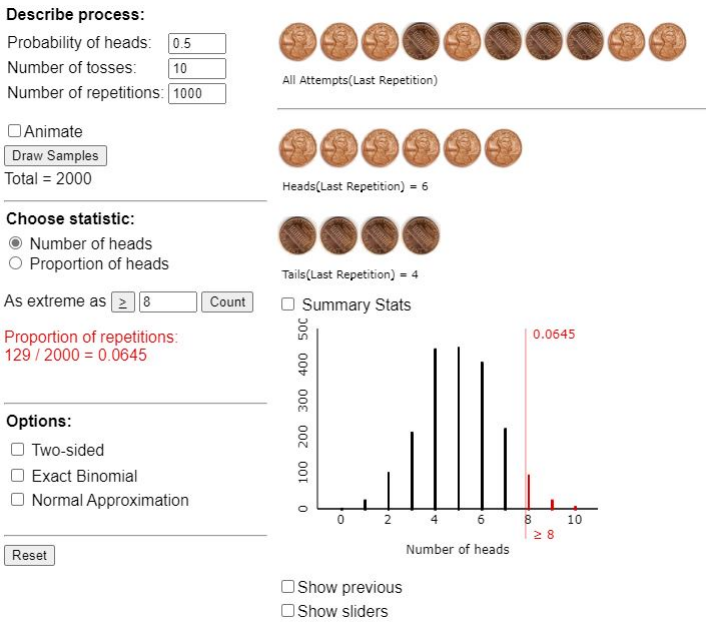
a) What do you believe the researcher's question was?

b) What are the observational units?

c) Give an example of a possible confounding variable and draw a causal diagram. Label your explanatory, response, and confounding variables.

d) Can you make a cause-and-effect conclusion from this data? Can you make inference to the population? Explain why or why not.

3) Examine the simulation results from the applet below to answer the following questions.



a) This analysis is for a categorical response.

b) Using symbols, what Null and Alternate Hypotheses are represented?

c) What is the observed statistic being tested?

d) Identify and interpret the resulting p-value.

e) If we set a significance level of 0.05, would we reject or fail to reject the null hypothesis?

f) What do the mathematical results tell us about generalizeability or cause-and-effect relationships?

4) Olympic games take place every two years and see competitors from all over the world compete in feats of strength and athleticism. Given their training and peak performance of the human body, one might wonder how they compare to the rest of us. The Olympics2016.csv file contains the results of a random sample of 2,014 Olympic athletes from the 2016 Summer Olympics (Rio De Janeiro). We want to compare the body composition of these athletes and see if they weigh less than the average North Americans, cited as 80.7kg.[1]

a) Can we use theoretical methods to run our analysis?

b) List the null and alternate hypotheses, as given in the problem.

c) List the standardized statistic and p-value for the hypothesis test from **b)** above. Interpret your results.

d) If we wanted a 90% confidence interval, what would our significance level (α) be?

e) What is the Margin of Error (MoE) for this data and a 90% confidence interval?

f) What is the 90% Confidence Interval for the average weight of 2016 Olympic athletes?

g) Who can these results be generalized to?

h) Can we infer a cause-and-effect relationship between being an Olympic athlete and weight loss?

[1] Walpole, Sarah C; Prieto-Merino, David; Edwards, Phil; Cleland, John; Stevens, Gretchen; Roberts, Ian; et al. (18 June 2012). "The weight of nations: an estimation of adult human biomass". BMC Public Health. BMC Public Health 2012, 12:439. 12 (1): 439. doi:10.1186/1471-2458-12-439. PMC 3408371. PMID 22709383

5) Sports teams prefer to play in front of their own fans rather than at the opposing team's site. Having a sell-out crowd should provide even more excitement and lead to an even better performance, right? Well, consider the Oklahoma City Thunder, a National Basketball Association (NBA) team, in its second season (2008–2009) after moving from Seattle. Using R, import the Basketball.csv dataset to conduct this analysis, which lists the home games of the Thunder during the season. (These data were noted in the April 20, 2009, issue of Sports Illustrated in the Go Figure column.)

a) What are the observational units of this dataset?

b) Identify the variables in this study and identify them as categorical or quantitative.

First, we want to investigate if the Thunder performed better at home. Overall for the 2008-2009 Season, the Thunder won 23 of their 82 games, or 28%. Using the provided data of home games, is the Thunder's at home win rate greater than their overall win rate?

c) State, in words and symbols, the null and alternate hypothesis.

d) What is the statistic and sample size for this data?

e) Does the data meet the validity conditions?

f) Using the appropriate methodology based on **e)** above, report your standardized statistic and p-value. Interpret your results using a 5% significance level.

g) Report a 95% confidence interval for the Thunder's at-home win rate. Does the null hypothesis fall within this range? Explain why this is or is not surprising, given the p-value.

We can visualize the values in R as a table to compute our observed statistics and begin our analysis for any differences with sold out games. For convenience, the win rate for sold out games and the win rate for games that did not sell out are presented below. Does there appear to be an association?

Win rate for not sold-out games is $\frac{12}{23} = 52.17\%$

Win rate for sold-out games is $\frac{3}{18} = 16.67\%$

i) It appears that whether a game sold out impacts win rate. Draw the causal diagram, labeling the explanatory and response variables.

j) What is the 95% confidence interval for the Thunder win rate for home games that are not sold out? Does it include the season average of 0.28?

k) Do you feel comfortable generalizing these results to all teams in the NBA?

l) Do you feel comfortable stating that the larger crowd of a sold-out game caused a lower win-rate for the Thunder? Why or why not?

m) Name a potential explanation for the association between if a game sold out and if the Thunder won or lost.

n) Update your causal diagram from 9 above to include this confounding variable and label it.

6) According to a 2018 report by the U.S. Department of Labor, civilian Americans spend 2.84 hours per day watching television. A faculty researcher, Dr. Sameer, at California Polytechnic State University (Cal Poly) conducts a study to see whether a different average applies to Cal Poly students. Suppose that for a random sample of 100 Cal Poly students, the mean and standard deviation of hours per day spent watching TV turns out to be 3.01 and 1.97 hours, respectively. There is not strong skew.

a) Is our statistic quantitative or categorical?

b) What is the value of our statistic (hint: $\hat{p}$ *and/or* $\bar{x}$ *and/or* s)?

c) Do we meet our validity conditions?

d) What is our 95% Confidence Interval for the true mean hours that Cal Poly students spend watching television per day?

e) Given our confidence interval above, what do we know about the results of a strength of evidence test with a null hypothesis of $\mu = 2.84$ and an alternate hypothesis of $\mu \neq 2.84$?

f) Report your standardized statistic (t or z) and p-value given the above data and a null hypothesis of $\mu = 2.84$ and an alternate hypothesis of $\mu \neq 2.84$.