MA206, Lesson 12 - Causation

What does it mean if two variables are **associated**?
The value of one variable gives information about the value of another variable

What is a **cause-and-effect** relationship? When can we conclude it?
A relationship where a change in the explanatory variable is causing the effect of the result variable. We can conclude this with a randomized experiment to mitigate confounding variables.

Define:
Explanatory Variable the "cause", the variable explaining the change

Response Variable the variable we think is being impacted, the "effect"

Confounding Variable a variable related to both the explanatory variable and response variable where its effect cannot be separated from the two

Observational Study comparing groups that are just there, observing what you see naturally

Experiment groups are created by what an experimenter chooses to do, assigning conditions to a control group and treatment group(s)

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		

Random Assignment → You can infer cause-and-effect conclusions; without random assignment, there is potential for confounding variables.
Random Sampling → You can make inferences to the population; without random sampling, there is potential for sampling bias.

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.

Whether or not the mother smoked while pregnant (Y/N); it is categorical.

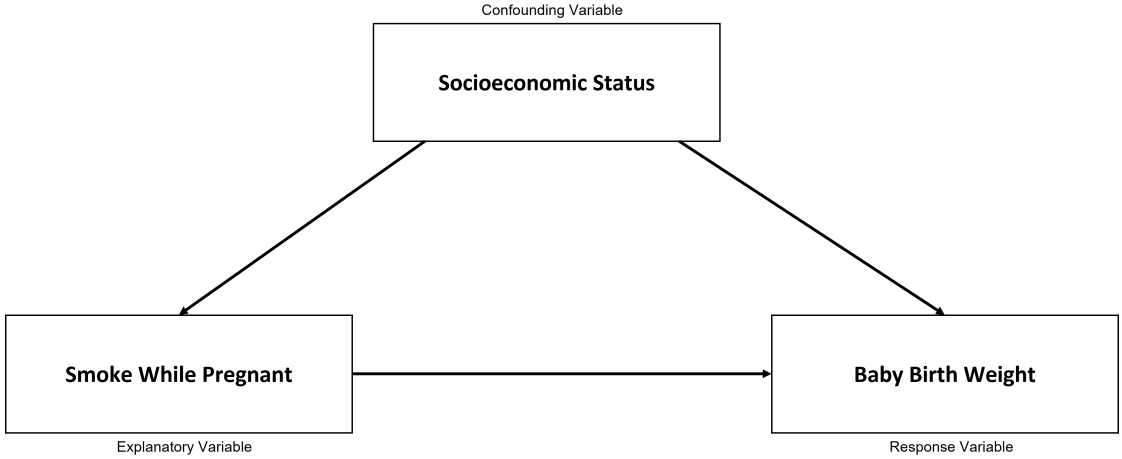
b) Identify the response variable in these studies. Classify it as categorical or quantitative.

Weight of the baby at birth, a quantitative variable.

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

Socioeconomic status may be a confounder because it may impact the rate at which people smoke (lower socioeconomic status correlates with a higher rate of smoking) and also have an impact on the birth weight of a newborn baby (lower socioeconomic status may result in correlation with a poorer diet, resulting in babies with lower birth weights).

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?

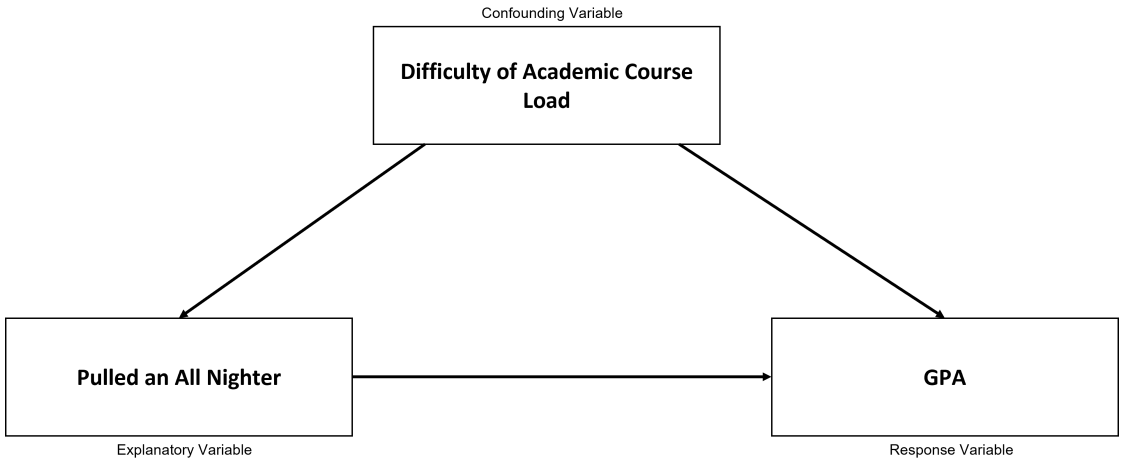
Do cadets who pull all-nighters have a lower GPA than cadets who do not pull all-nighters

b) What are the observational units?

Cadets

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

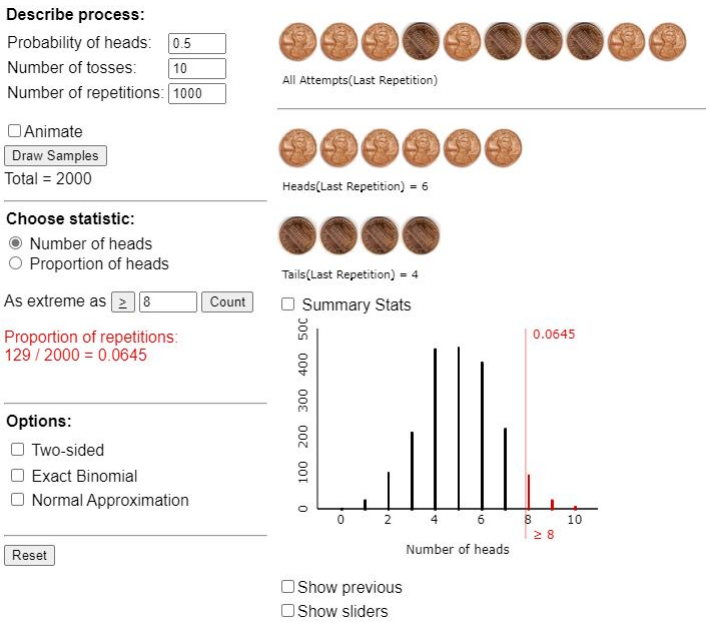
e.g. difficulty of classes/academic major, poor time management skills, prioritize Xbox over coursework



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

You cannot make cause-and-effect conclusions because it was not a randomized experiment, it was a single observational study. Additionally, you cannot make inference to the population because random sampling was not conducted; This was a convenience sample of cadets in Thayer Hall.

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?

$H_0 : \pi = 0.5$

$H_a : \pi > 0.5$

c) What is the observed statistic being tested?

$\hat{p} = \frac{8}{10} = 0.8$

d) Identify and interpret the resulting p-value.

The p-value is 0.0645, which provides moderate evidence against the null hypothesis that the true proportion of this process is 0.5.

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

Fail to reject the null hypothesis because 0.0645 is greater than 0.05, so we consider it plausible.

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

Nothing. The math results don't indicate either case.
Generalization occurs if we use random sampling to mitigate sampling bias.
Cause-and-effect can be inferred through random assignment to mitigate confounding variables.

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?

Yes, this is quantitative data so we need at least 20 observations and the data should not be strongly skewed. We have 2,014 observations and the data is not strongly skewed, therefore we meet our validity conditions.

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

$H_0 : \mu = 80.7$. The true mean weight of Olympic athletes from the 2016 Olympic Games is 80.7kg.
$H_a : \mu < 80.7$. The true mean weight of Olympic athletes from the 2016 Olympic Games is less than 80.7kg.

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

$\bar{x} = 74.0$
$s = 16.2$
$n = 2014$
$t = \frac{\bar{x}-\mu}{\sqrt{\frac{s^2}{n}}} = \frac{74-80.7}{\sqrt{\frac{16.2^2}{2014}}} = -18.56049$
p-value = pnorm(t) = pnorm(-18.56049) = $2.055603e^{-71}$.
With a p-value of $2.0556e^{-71}$, we have very strong evidence against the null hypothesis that olympic athletes have an average weight of 80.7. We conclude that they have a lower long-run average weight.

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

$\alpha = 0.1$.

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

Note that the Margin of Error is the Multiplier (M) $\times$ Standard Error (SE).
M = qt(1 - $\frac{\alpha}{2}$, n-1)
SE = $\sqrt{\frac{s^2}{n}}$
MoE = $M \times SE = qt(1 - \frac{0.1}{2}, 2014 - 1) \times \sqrt{\frac{16.2^2}{2015}} = 0.594035566$

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

(73.40596, 74.59404) kg.
CI = $\bar{x} \pm ME$ = 74 ± 0.594035566

[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

g) Who can these results be generalized to?

These results can be generalized to 2016 Olympic Athletes, as this is who the random sample was sampled from. We cannot generalize beyond this population.

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

No, we cannot infer Cause and Effect between these two because we did not mitigate confounding variables through a randomized experiment.

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?

Each individual home game

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

If the game was sold out or not, which is Categorical. Also, if the Thunder won or not, which is also Categorical

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.

H_0: The win rate of Oklahoma City Thunder home games (π) is equal to 0.28.
H_a: The win rate of Oklahoma City Thunder home games (π) is greater than 0.28

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

n = 41
$\hat{p} = \frac{15}{41} = 0.3659$

e) Does the data meet the validity conditions?

Yes, we have 15 successes and 26 failures, both of which are greater than 10

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

Using theory because we met our validity conditions, we are able to substitute into our formula and obtain $z = 1.22435$. This corresponds to a p-value of 0.1104102, which provides weak evidence against the null hypothesis that the home game win rate is 0.28. With a significance level of 0.05, we fail to reject the null hypothesis and conclude that 0.28 is a feasible value for the true at-home win percentage.

$z = \frac{0.36585 - 0.28}{\sqrt{\frac{(0.28)(1-0.28)}{41}}} = 1.22435$

p-value = 1 - pnorm(z) = 0.1104102

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.

$\hat{p} \pm qnorm(1 - \frac{\alpha}{2}) \times \sqrt{\frac{\hat{p}(1-\hat{p}}{n}} = 0.3659 \pm 1.959964 \times \frac{(0.3659)(1-0.3659)}{41} = (0.218, 0.513)$

Our null hypothesis of 0.28 falls within this range, which is not surprising as the p-value of 0.1104102 is larger than our significance level of 0.05.

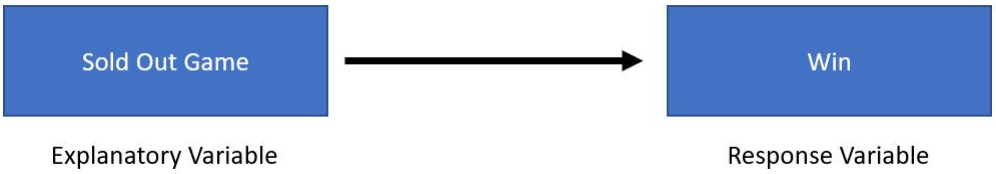
We can visualize the values in R as a table to compute our observed statistics and begin our analysis for any differences with sold out games. For convenience, the win rate for sold out games and the win rate for games that did not sell out are presented below. Does there appear to be an association?

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

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

There does seem to be an association because there is a difference in win rates depending on sold-out games versus not sold-out games.

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?

The 95% confidence interval is (0.318, 0.726), which does not include 0.28. This indicates that, with a significance level of 0.05, we would reject the null hypothesis that the true win rate for the Thunder for home games that are not sold out is 0.28.

$= \frac{12}{23} \pm qnorm(1 - \frac{0.05}{2}) \times \sqrt{\frac{(12/23)(1-12/23)}{23}} = (0.318, 0.726)$

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

No, there was not random sampling to mitigate bias, so we cannot generalize these results outside of the sample taken - that is, the Thunder.

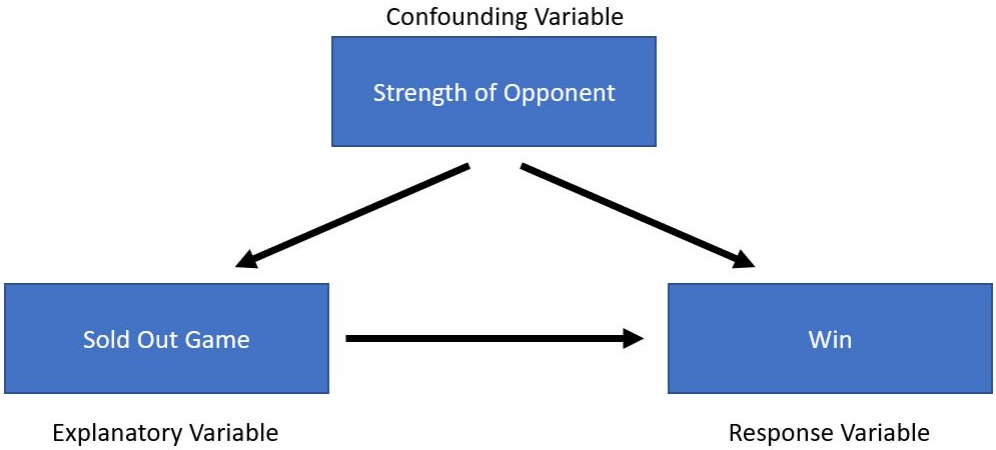
l) Do you feel comfortable stating that the larger crowd of a sold-out game caused a lower win-rate for the Thunder? Why or why not?

No, because there may be confounding variables not captured in this summarized data, which is an observational study and not a randomized experiment. Without randomized assignment, we cannot infer cause and effect.

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

Answers will vary, could include strength of opponent.

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


Confounding Variable
Strength of Opponent
Sold Out Game
Win
Explanatory Variable
Response Variable

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?

Quantitative

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

$\bar{x} = 3.01$, s=1.97

c) Do we meet our validity conditions?

Yes, we have at least 20 observations ($100 \geq 20$) and the data is not strongly skewed.

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

Confidence Interval $= \bar{x} \pm qt(1 - \frac{\alpha}{2}, n-1) \times \frac{s}{\sqrt{n}}$ $= 3.01 \pm qt(1 - \frac{.05}{2}, 99) \times \frac{1.97}{\sqrt{100}}$
$= (2.6191, 3.4009)$

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 $\mu \neq 2.84$?

We know that the p-value will be greater than 0.05, as 2.84 did "make the cut" and falls within our 95% confidence interval.

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 $\mu \neq 2.84$.

$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{3.01 - 2.84}{\frac{1.97}{\sqrt{100}}} = 0.8629$

p-value = 2*(1 - pt(abs(t), n-1)) = 2 * (1 - pt(abs(0.8629442), 99)) = 0.3903