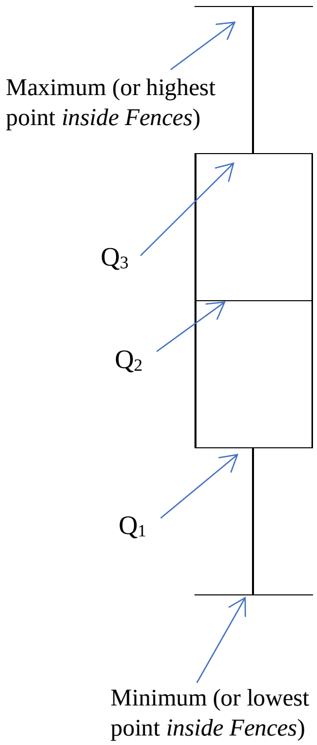
Chapter 2 – Comparing Groups

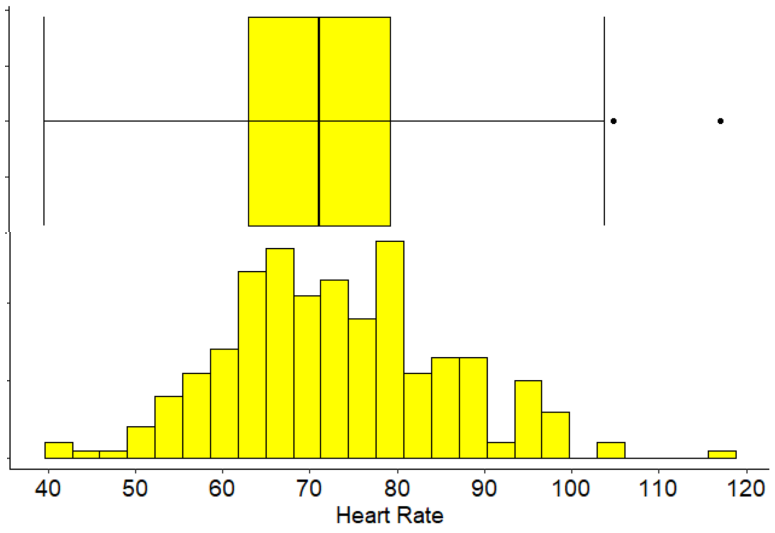
Boxplots


Maximum (or highest point *inside Fences*)
Q3
Q2
Q1
Minimum (or lowest point *inside Fences*)

- A graphical representation of the 5-number summary of a numeric variable.

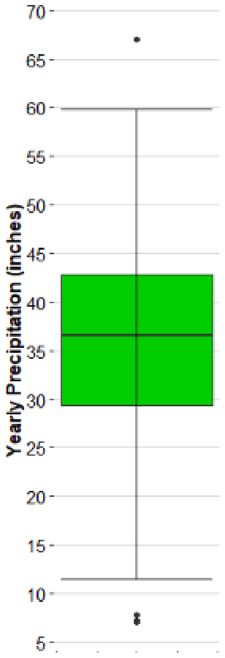
- The “whiskers” (outside lines) are the minimum and maximum values still inside the Upper/Lower fences.
    - Lower Fence = $Q_1 - 1.5(Q_3 - Q_1)$
    - Upper Fence = $Q_3 + 1.5(Q_3 - Q_1)$
- Outliers are denoted by a tiny dots past the first or last whisker—data values that fall outside these fences.

Why Boxplots?



Practice: A meteorologist records the yearly precipitation in 70 large U.S. cities. Between what 2 precipitation amounts do the middle 50% of cities fall in?

What was the highest yearly precipitation amount recorded?

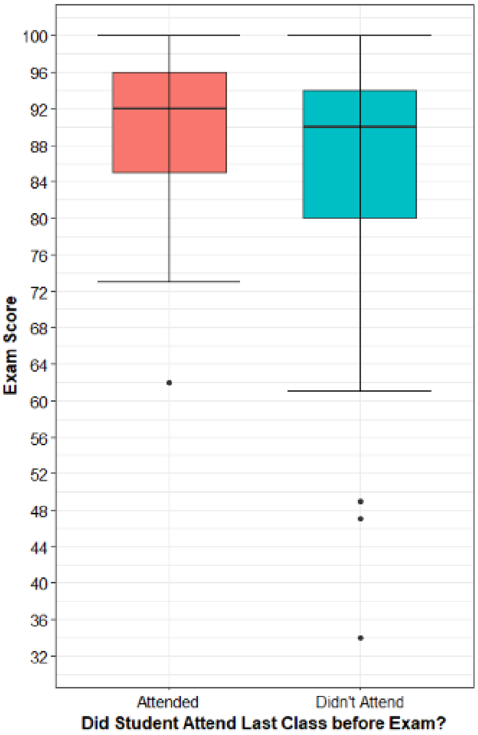


Multivariate Comparisons

- **Univariate vs. Multivariate Questions**
  - **Univariate Questions:** Ask about characteristics of one variable in isolation:
    - *What temperatures does Chicago typically see during the year?* What type of data would we collect to answer this question?
    - *What percent of residents are in favor of Proposition 5?* What type of data would we collect to answer this question?
  - **Multivariate Questions:** Ask about the relationship between two variables
    - *On average, is the mortality rate in Chicago higher on days with higher temperatures?* What two types of data would we collect to answer this question?
    - *Are homeowners who support Proposition 5 more likely to rent than those who don't support that proposition?* What two types of data would we collect to answer this question?
    - *Is amount of time spent studying different on average between Freshmen, Sophomores, Juniors, and Seniors?* What two types of data would we collect to answer this question?

- **Comparing Means, Medians, and Quartiles**
  - **Boxplots** can be very helpful for comparing the position of certain key points in a distribution. Most obviously, we can compare the medians of multiple groups, or other quartiles.

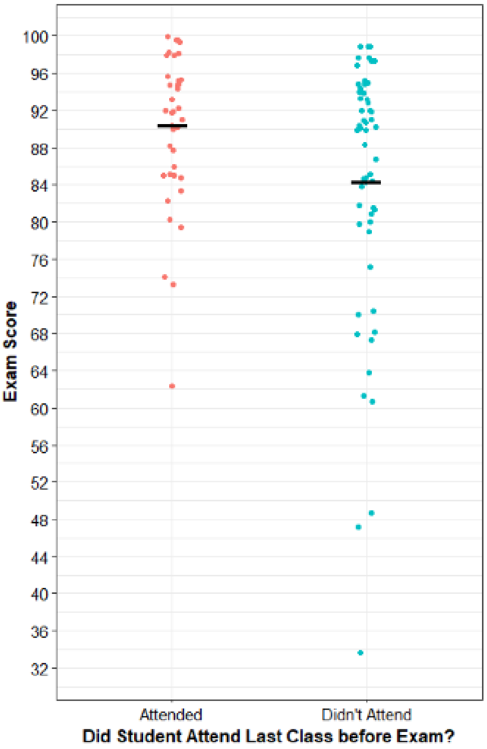
Example. Consider the following plot representing the following two variables:



What is the difference in median score of students who attended class as compared to those who didn't?

Consider the claim: "Some student who didn't attend class scored much higher than students who did." Is this true?

- **Jitter Plots** are an alternative that shows the position of all data points *(with some random jittering to make it easier to see all of the data and avoid excessive overlap)*
- In addition, there is a **black bar** at the position of the **sample mean** for each group.



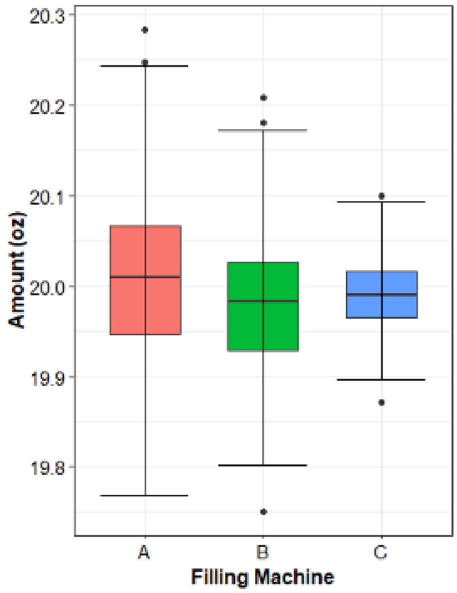
What is the approximate difference in mean scores of students who attended class as compared to those who didn't?

Somebody who attended class scored on a 62 on the exam. Does this mean attending class is not potentially beneficial for students?

- **Comparing Variability**
  - We might also be interested in comparing other distributional characteristics, such as how variable the data is in different groups.

Example. Consider a Coca-Cola Production Plant that is building machinery to fill a 20 ounce Coke Bottle. As with any machine, there is a small amount of variability in the content poured into each bottle. The research team is comparing three filling machines, in which each machine filled 500 bottles.

Which machine results in the smallest amount of variability in weight across the bottles it fills?

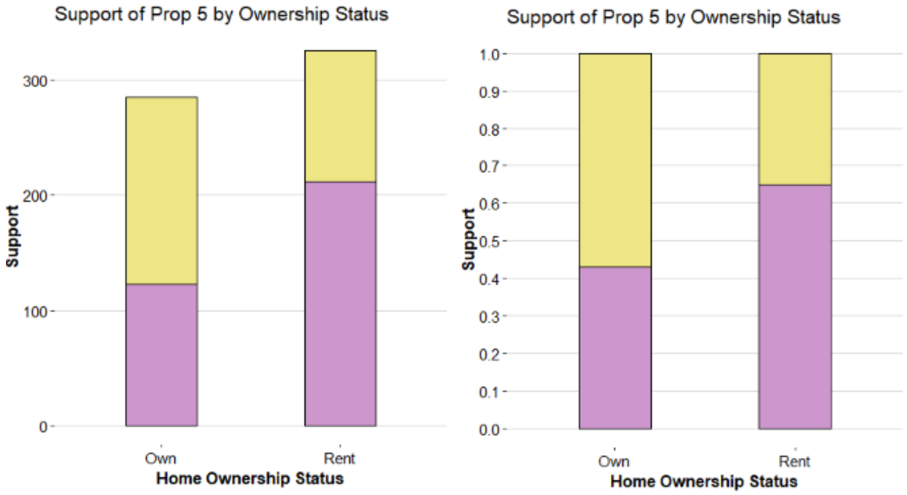


- **Comparing Proportions**
  - In cases where the two variables we examining associations with are non-numeric, we can instead compare proportions between each group.

Example. Consider this data examining the question we posed earlier: Support for Proposition 5 based on Owning/Renting status

Notice there are two plots here: One is a normal stacked barplot (showing counts on the y axis). The second is a **100% stacked barplot**, which shows the proportion who answered each response within each bar.

What is the approximate difference in proportions supporting Proposition 5 by Own/Rent status?



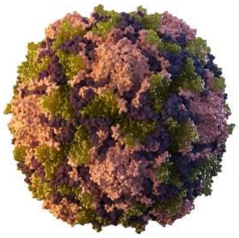
Comparing Proportions through "Risk"

- In biostatistics, the proportions we compare are often framed in terms of "risk."
    - Consider different questions we might ask regarding risk
        - What is the risk for infection if I'm vaccinated? **(Absolute Risk)**
        - What is the difference in infection risk between vaccinated and unvaccinated individuals? **(Absolute Risk Reduction)**
        - What is the infection ratio between vaccinated and unvaccinated individuals? **(Relative Risk)**
        - How **Effective** is the vaccine?—derived from relative risk by reporting the percentage of people who would likely *avoid* the infection by taking the vaccine.

    - We can make comparisons to risk in the context of both helpful interventions and hazardous interventions.

- **Poliovirus Example**
    - In a 1954 experiment, children were randomly assigned to either receive the experimental polio vaccine or a placebo vaccine (a saline injection that would do nothing).
    - The children were monitored for about one year, and the results are presented below:

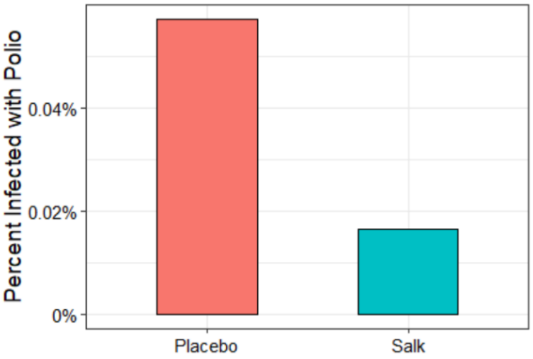
	Polio	**No Polio**	**Total**
Salk Vaccine	33	200,712	200,745
Placebo	115	201,114	201,229



    - **Absolute Risk** is just a proportion. It is the proportion of cases out of the total.
    - We can calculate the absolute risk under each intervention.

Absolute Risk for Polio with Vaccine	**Absolute Risk for Polio with Placebo**

Absolute Risk *Reduction* reports the absolute value difference between risk for two different groups. It is typically reported as an absolute value and framed in terms of which condition had the lower risk.



ABR = $|Risk_A - Risk_B|$.

This can also be reported as a percentage: $|Risk_A - Risk_B|*100\%$

Practice: What is the absolute risk reduction for polio when taking the vaccine?

Relative Risk (RR) *(sometimes referred to as a "Risk Ratio")* represents the ratio of risk under one condition to another condition.

RR = $\frac{Risk_A}{Risk_B}$ **.** *This can also be reported as a percentage*: $\left(\frac{Risk_A}{Risk_B}\right)*100\%$

Practice: What is the relative risk for polio when taking the vaccine?

The estimated risk of contracting polio after taking the Salk Vaccine is ______ times the risk for polio after taking the placebo.

- If relative risk is **below 1**, that means risk is **reduced**.
- If relative risk is **above 1**, that means risk is **increased**.

Effectiveness: Represents the percent of individuals that would avoid the infection by taking part in the intervention. This is often reported for vaccines, but could be reported for other treatments as well.

Effectiveness = 1 – RR. *This can also be reported as a percentage:* (1 – RR)*100%

Practice: How effective is the Salk Vaccine at preventing polio?

Practice: Consider a study in which 40 participants are asked to complete a task that required mental focus, where half were assigned to drink a caffeinated beverage beforehand, and half drank water. Each participant was given a score based on how well they completed the task, as judged by an expert group, with a maximum possible score of 50 and minimum of 0.

- Caffeine Group: The mean score was 37.2, the median was 43, and the standard deviation was 5.1.
- Water Group: The mean score was 38.7, the median was 42, and the standard deviation was 4.3.

Let's identify the Caffeine group as Group 1, and the Water group as Group 2.

On average, the caffeine group scored ______ points higher/lower. Or symbolically: $\bar{x}_C$ - $\bar{x}_W$ = ______

Which group's scores was more variable?

The lowest score came from a student in the caffeine group. Would that be good evidence to use in arguing that water is better than caffeine in readying students for this task?

Each of these score distributions is skewed. Based on the summary statistics and provided information, which direction does each distribution appear to skew?

Explaining Variability

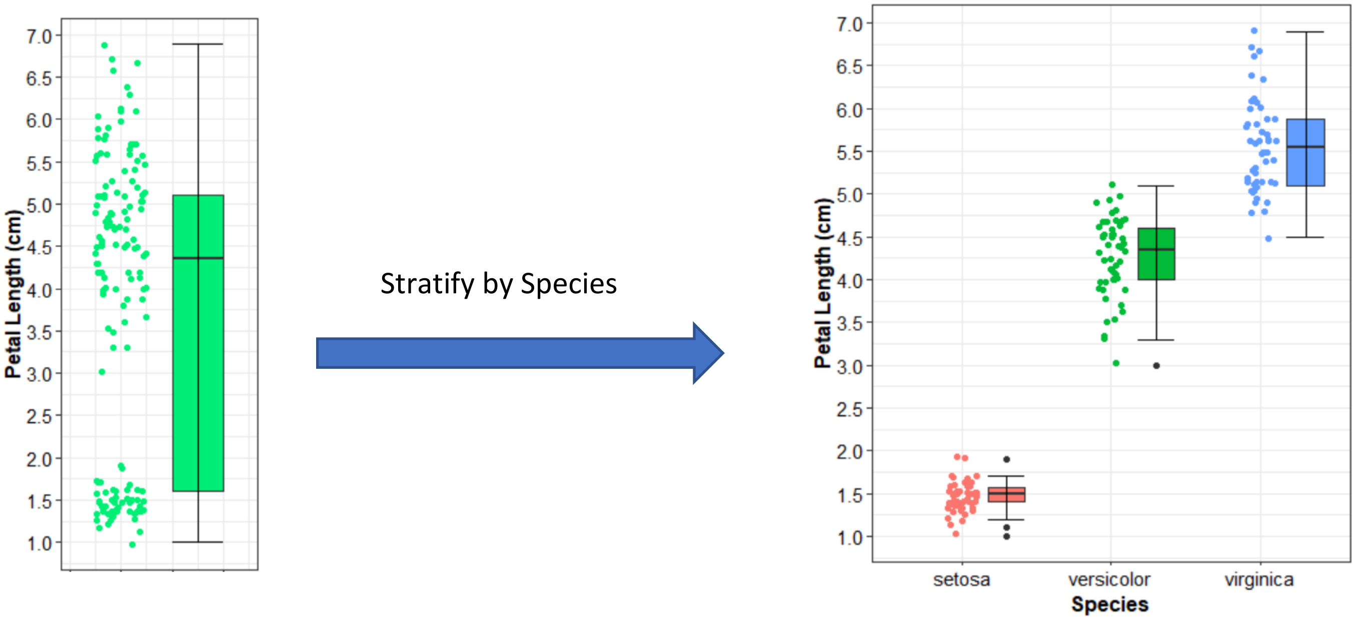
- Identifying a response variable
  - A **response variable** is a variable that we have an interest in explaining or understanding better.
  - Often when posing question, we want to use information from one or more variables to better explain or predict the response variable.

Example. A professor wants to understand how they can best support their students to succeed, using exam scores as the indicator of success. They collect information such as class attendance, homework scores, and time reported working outside of class as indicators to see if these things might possibly predict exam scores.

The response variable in this example is: ____________________________

- Explaining Variability through Stratification
  - **Stratification** is the analytical process of breaking up one variable into subgroups based on the value of another variable.

Example. Consider a garden of iris plants. A botanist measures petal length of each iris that has blossomed. The distribution of petal lengths is represented below. He also notices there are 3 distinct species of iris plants in this garden, so he stratifies by species as well.

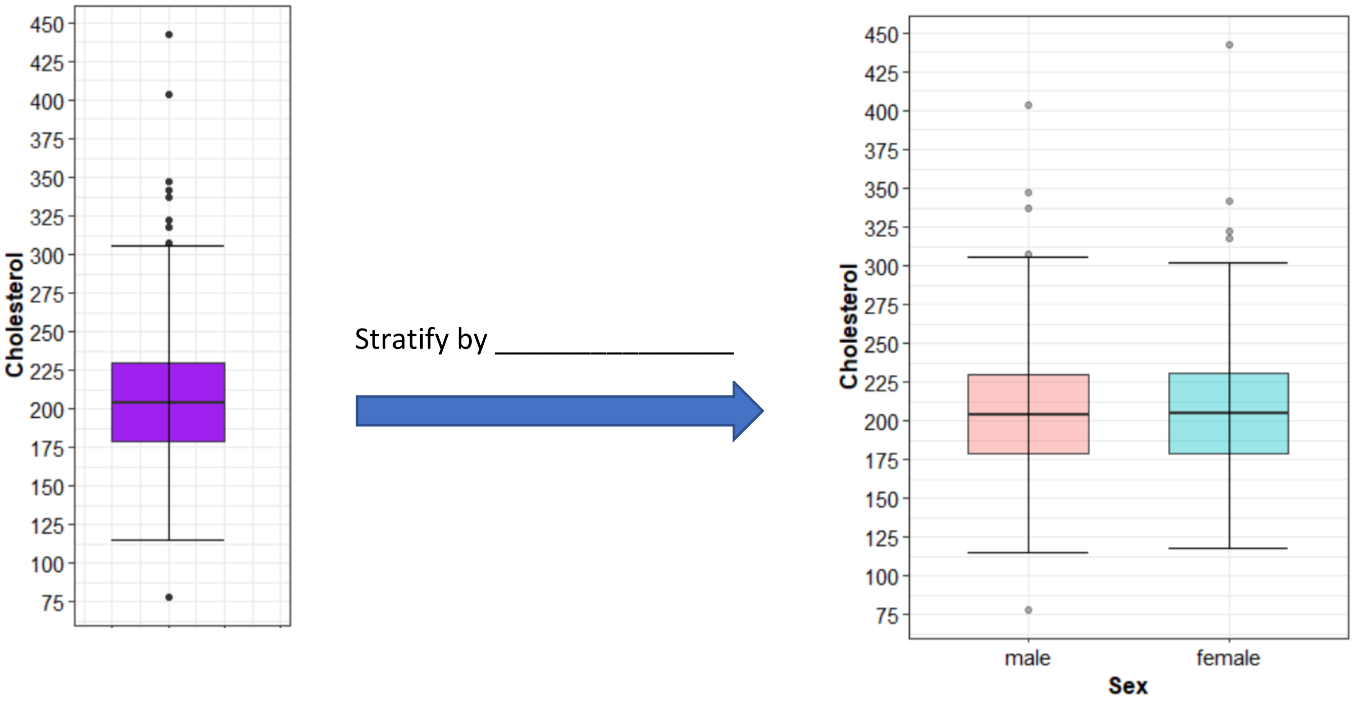


	All
Mean	3.76
SD	1.765

	Setosa	Versicolor	Virginica
Mean	1.46	4.26	5.55
SD	0.174	0.470	0.552

Can we make more accurate predictions for an iris flower's petal length by stratifying by species? Just using the graphs and summary statistics here, do you think we're explaining a lot of variability, a little, or basically none?

Example. Are cholesterol levels different by biological sex? Consider the following data representing approximately 403 adults.



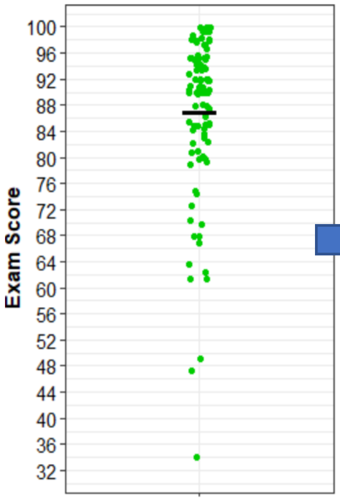
Stratify by ________________

	All
Mean	207.8
SD	44.4

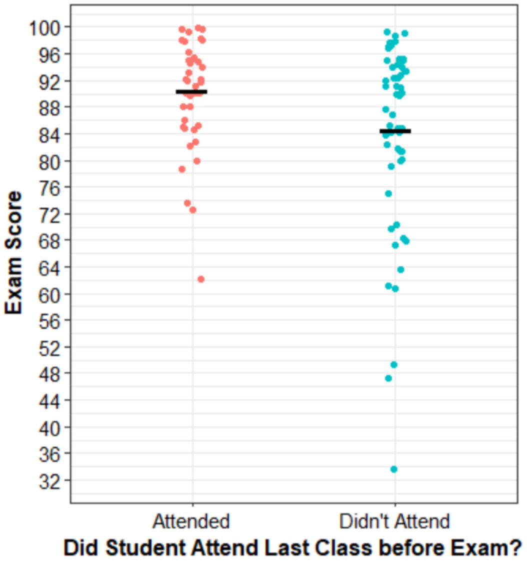
	Male	**Female**
Mean	207.5	208.3
SD	45.5	43.7

Can we make more accurate predictions for cholesterol levels by knowing someone's biological sex? Just using the graphs and summary statistics here, do you think we're explaining a lot of variability, a little, or basically none?

Example. Return to this example we saw earlier. Is there a difference in Exam score on average based on whether or not students attended the last day of class before the exam?



Stratify by ______________



	All
Mean	86.8
Standard Deviation	12.4

	Attend	**Didn't Attend**
Mean	90.3	84.3
Standard Deviation	8.4	14.2

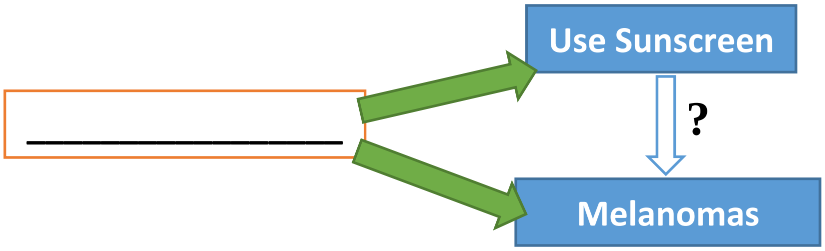
Can we make more accurate predictions for someone's exam score by stratifying by attendance on the last class period before the exam? Just using the graphs and summary statistics here, do you think we're explaining a lot of variability, a little, or basically none?

Making Causal Claims

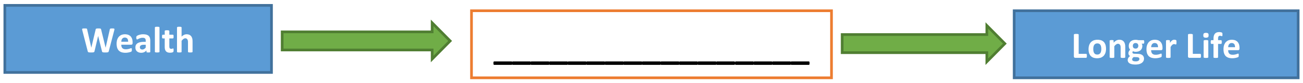
- Introducing Causality
  - Stratifying is an analytical technique that helps us decide if there may be an association between two variables.
  - But…we need context of how the data was collected or additional analysis before we can decide if changes in one _____________ changes in the other.
- Confounders and Mediators
  - When finding an association between two variables, there are a few possibilities.
  - It's also possible that *more than one* of these could be true at the same time!

Causality	**No Causality**
• The predictor directly causes changes in the response • The predictor indirectly affects the response variable by beginning a causal chain that will affect the response (there is a ______________ variable)	• The predictor is merely associated with something that causes changes in the response (there is likely a __________________ variable)

Example of Confounding Variable. Consider a medical study to examine factors that might lead to melanomas (skin cancer). One researcher notes that people with melanomas were much more likely to have reported using sunscreen in the last year. Does that mean that sunscreen is causing skin cancer? Can you think of any possible confounders in this relationship?



Example of Mediating Variable: People who earn more income tend to have longer lives. Does that mean that money itself is directly increasing lifespan?



Remember that confounders have to have a direct relation with both variables/factors in the association! If we notice something that directly relates to one factor, but not the other, then it doesn't explain the relationship.

Practice: A study finds that people who carry lighters have a higher rate of lung cancer. Consider the following explanations and whether it is a mediator, a confounder, or neither. Consider drawing a diagram of each to show what is affecting what.

A. Genetics—some people are more genetically prone to lung cancer than others.
   a. Mediator
   b. Confounder
   c. Neither

B. Smoking cigarettes—people who smoke cigarettes have a higher rate of lung cancer and are also more likely to carry lighters
   a. Mediator
   b. Confounder
   c. Neither

C. Lighter fluid—inhaling the fumes from lighters causes lung damage that leads to cancer
   a. Mediator
   b. Confounder
   c. Neither

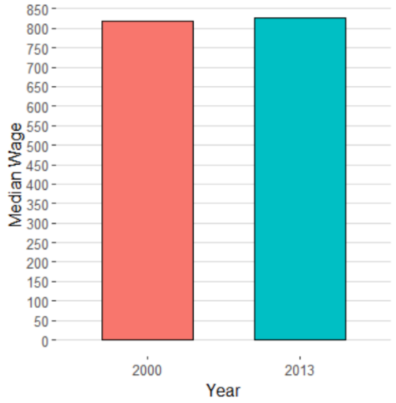
D. Radon—radon exposure raises one's risk for lung cancer
   a. Mediator
   b. Confounder
   c. Neither

Practice. Consider an instructor who teaches two sections of the same course with the same curriculum. She finds that students have higher final exam scores in the 2pm section as compared to the 9am section. There are several possible explanations. Identify which explanation is a causal relationship and which is an association explained by a confounder.

A. Students who enroll in the 2pm are more likely to be Juniors/Seniors, while the 9am section is more likely to be Freshmen/Sophomores. Juniors/Seniors have stronger prior knowledge and stronger studying skills, so they perform better.

B. The professor teaches better at 2pm than at 9am because she has a chance to see what works and make changes before the 2pm class.

- **Identifying Simpson's Paradox through Stratification**
  - In 2013, the Labor Department reported that the median weekly income in the United States had risen by about 1%, from $819 to $827 (after adjusting 2000 dollars for inflation)



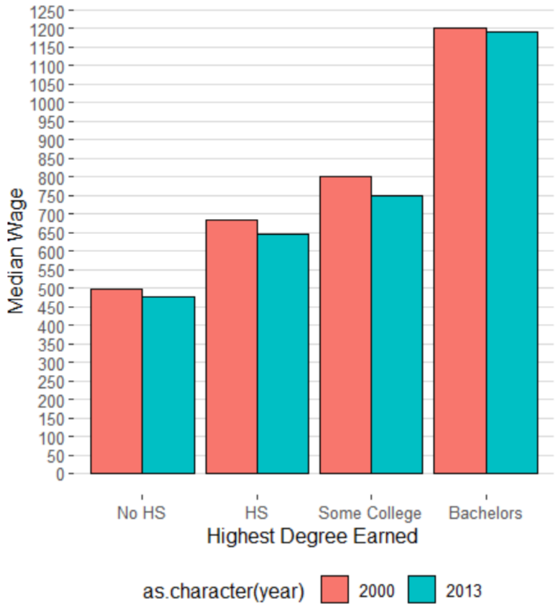
	2000	2013
Median Weekly Income	$819	$827

  - Accompanying that figure was the breakdown for wage change, stratified by education level.

	2000	2013
Less than High School	$498	$477
High School Diploma	$685	$647
Some College	$801	$751
Bachelors or higher	$1201	$1193

* Figures from: https://www.bls.gov/charts/usual-weekly-earnings/usual-weekly-earnings-over-time-by-education.htm

*2000 figures adjusted for inflation using this calculator: https://www.bls.gov/data/inflation_calculator.htm



  - **This is an Example of Simpson's Paradox:** After stratifying by a variable, we find that a relationship we observed actually flips directions.
    - Median wages rose on average from 2000 to 2013
    - After stratifying by ________________________, income levels actually _______________.
  - **How?**

  - As an example, the following proportion adjustments to each subgroup could create a higher overall median wage, even as each subgroup is dropping!

	2000	2013
Less than High School	22% at $498	21% at $477
High School Diploma	24% at $685	23% at $647
Some College	26% at $801	27% at $751
Bachelors or higher	28% at $1201	29% at $1193

Simpson's Paradox can occur when we compare groups that are not equivalent. The workforce in 2013 is not equivalent to the workforce of 2000, so generic comparisons of the two without accounting for underlying differences may create a misleading comparison about the state of wage growth.

Simpson's paradox happens in many situations, including understanding the mortality rate of COVID-19. See this video: https://www.youtube.com/watch?v=t-Ci3FosqZs

Practice: To determine how effective masks are in preventing the spread of COVID-19, researchers identified cities that had a mask mandate and cities that did not. They then tracked the percentage of residents who contracted COVID-19 over the following 4-month period.

What other variables should the researchers consider collecting and stratifying by?