Looking at Statistics with a Critical Eye

- There's a phrase you've probably heard before: "There are lies, damned lies, and statistics!" Many of the most egregious deceptions are made *using* data.
    - But how can that be? Wouldn't the presence of data suggest some merit behind one's claim??
- As we have discussed with validity in experimental design, there are many threats to trustworthy data collection. Add on top of that likelihood for misinterpretation of results, and even purposeful deception. Faulty claims based in data can be some of the most dangerous.
- We net consider some common fallacies in data science so that you can be an informed consumer and honest producer of the statistical work you will engage with one day.

Observational Studies – The Weaknesses you should know

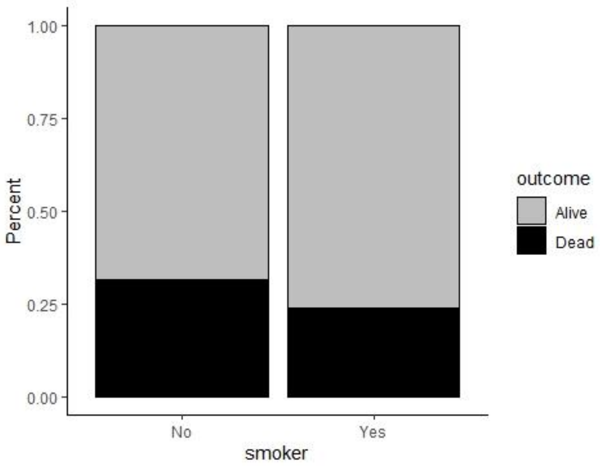
- As an alternative to Experiments, Observational studies can also assess the relationship between two variables, but without the luxury of a "controlled" interventions. They essentially make note of patterns "as life happens."
- For this reason, observational studies are very difficult to use as a basis for identifying causal links.
- Observational studies are necessary when we cannot reasonably assign people to groups or allow for any kind of planned intervention.
    - Studying the potential effects of alcohol consumption on fetal development. Is it ethical to assign pregnant women into alcohol and non-alcohol groups?
    - Studying whether brain size is for infants born with and without cleft pallet. The presence or absence of cleft pallet is naturally occurring rather than administered.

- **Stratification – Exploring observational study data for confounders**
    - When experimental designs are not possible, or simply not available, observational studies can still be used to explore the possibility of a causal link by adding known confounding variables to our collected data.
    - But since observational studies do not allow for equivalent groups and no controlled intervention, I can't be sure it is the supposed "treatment" factor that is really what's causing changes in my response variable.
    - With stratification, we essentially break our data into smaller subgroups that help us better compare "apples to apples" and "oranges to oranges."
    - Stratification is also a good analytical technique for experiments when I'm concerned that my groups may not be balanced!
    - The following simulation can show us what it might look like to stratify observational study by a possible confounding variable and see the relationship disappear or actually flip! https://istats.shinyapps.io/MultivariateRelationship/

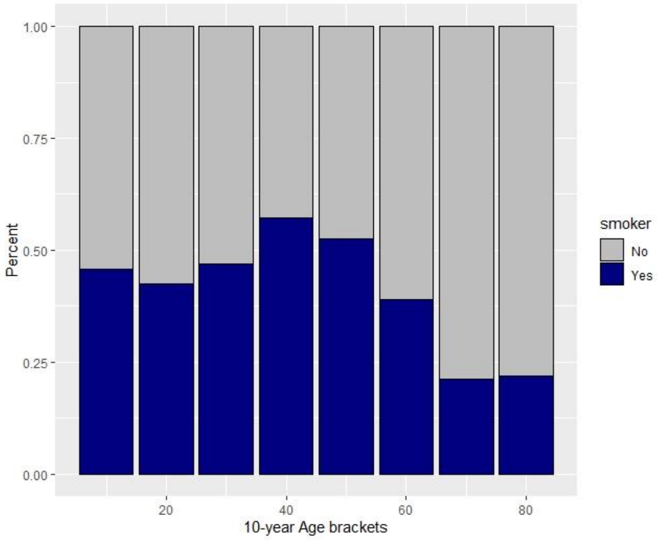
We can also see stratification in this observational study that we saw earlier:

- Consider this old study from the UK: *Data on age, smoking, and mortality from a one-in-six survey of the electoral roll in Whickham, a mixed urban and rural district in the UK. The survey was conducted in 1972-1974 to collect various information. A follow-up on those in the survey was conducted over twenty years later, noting whether participants had passed away or not.*

- A striking observation was made between survival and smoking status. As the 100% stacked bar chart below shows, non-smokers were more likely to have passed away than smokers. Does this suggest that smoking somehow increases length of life?



- As you might guess, this is an observational study, a prospective study to be specific. We didn't "assign" participants to be smokers and non-smokers randomly. We don't know from this graph if these two groups of people are truly equivalent in all respects except for smoking.
- But fortunately, several other variables were collected as part of this study, including the age of the participants at first data collection. Would smokers and non-smokers in 1972-1974 include the same general distribution of ages?



- This second graph shows the smoking status breakdown by age. In 1972-1974, it would appear that smoking was noticeably less prevalent among those in their 70s and 80s!
- We could stratify our data by age by looking at death rates for smokers vs. non-smokers, but separately by age bracket. This might reveal a different story!

	Alive	Dead	% Dead
Smoker	443	139	23.9%
Non-Smoker	502	230	31.4%

- This first table shows that, when we ignore the age of participants, the mortality rate of smokers 20 years later was 7.5% lower than that of non-smokers.

Famously known as an example of Simpson's Paradox

- But when we compare "apples to apples" by stratifying for age, that difference in mortality rate essentially disappears within each age bracket. Stratifying doesn't exactly enable us to make a causal link; rather, it allows us to test possible confounders in our investigation.

	Smokers		Non-Smokers		% Dead	
Ages	**Alive**	**Dead**	**Alive**	**Dead**	**Smokers**	**Non-Smokers**
18-29	109	3	146	3	2.7%	2.0%
30-39	119	5	134	6	4.0%	4.3%
40-49	96	22	78	10	18.6%	11.4%
50-59	78	34	76	25	30.4%	24.7%
60-69	38	44	57	72	53.7%	55.8%
70+	3	31	11	124	91.2%	91.8%

Practice: A survey is conducted to college students asks whether they eat dinner at approximately the same time every day. This survey also asks how many hours of sleep they get a night. The survey finds that students who answer "yes" to eating dinner at the same approximate time of day also get more sleep per night on average.

What other variables should the researchers consider collecting if they'd like to explore whether eating time is a potential causal link for sleep?

doing homework night / day time ?
taking nap ?
early morning class ?

Practice: A group of cardiologists identified patients with diagnosed heart disease. The researchers then looked back at medical records to determine which were prescribed a particular aspirin that the researchers suspected might have links to heart disease. They found a clear association

What other variables should the researchers consider collecting if they'd like to explore whether taking this aspirin might be causally linked to heart disease?

age ?
family medical history ?
gender ?