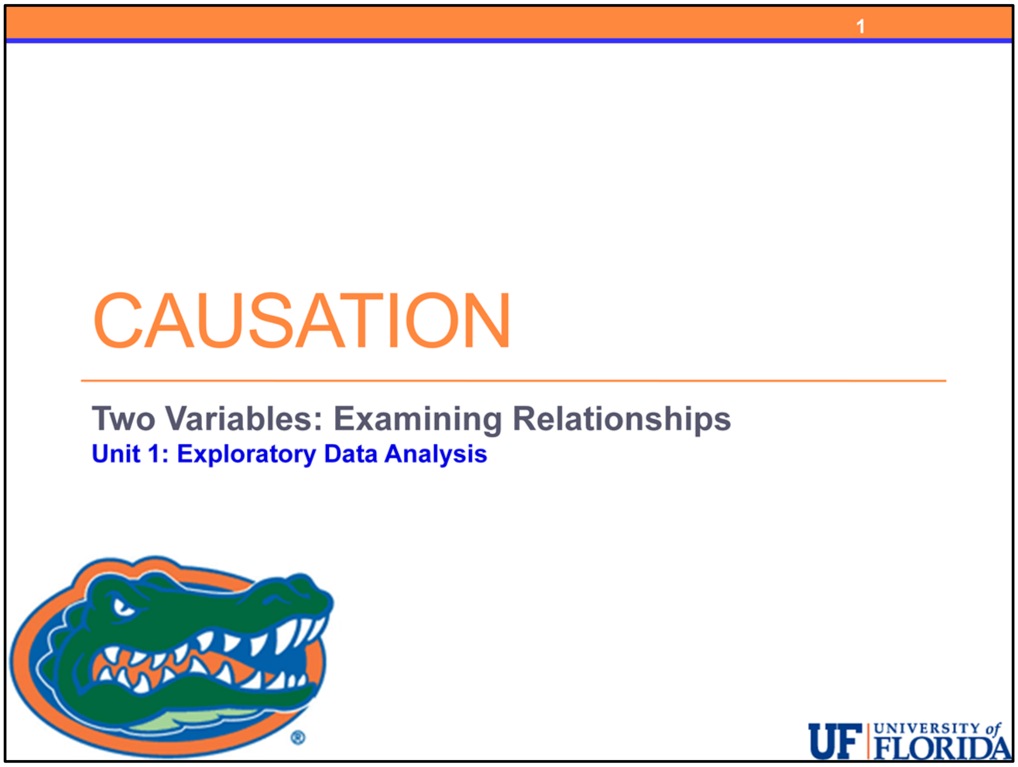


CAUSATION

Two Variables: Examining Relationships

Unit 1: Exploratory Data Analysis

We have talked about the Role-Type classification and discussed exploratory data analysis in each case. When we wish to investigate the effect of an explanatory variable (or even multiple explanatory variables) on a particular response variable, our goal is often to be able to say that changes in the explanatory variable CAUSED the observed changes in the response variable. For example,

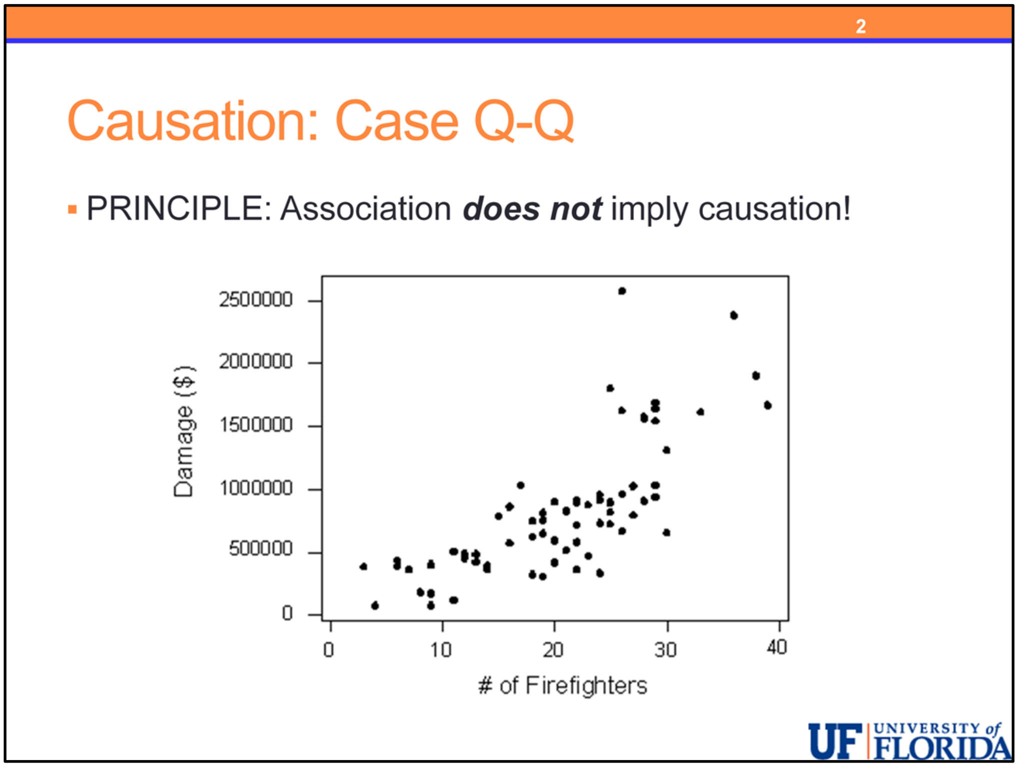
A new treatment results in a better outcome for patients.
An intervention works to reduce a risky behavior or promote a healthy one.
Following a particular diet will result in weight loss.

Unfortunately proving causality – beyond any doubt - is very difficult to accomplish. And, although designed experiments are our best hope of proving cause, even when we take as much control over our study as possible, there is always the possibility that there is some unmeasured variable or set of variables which actually were the true cause of the result that we observe.

Showing an association or relationship between two variables is easy, determining the causal nature of the observed association is much more difficult.

We hope to convince you that your initial reaction should be that just because you see a relationship between two variables does NOT mean that X had any causal effect on Y.

To illustrate we being with a simple example.



2

Causation: Case Q-Q

- PRINCIPLE: Association ***does not*** imply causation!

Here we have the number of firefighters vs. the amount of damage done.

If we were going to be naïve we would say something like “sending more firefighters causes more damage” and then we might conclude that to reduce damage we should send less firefighters!

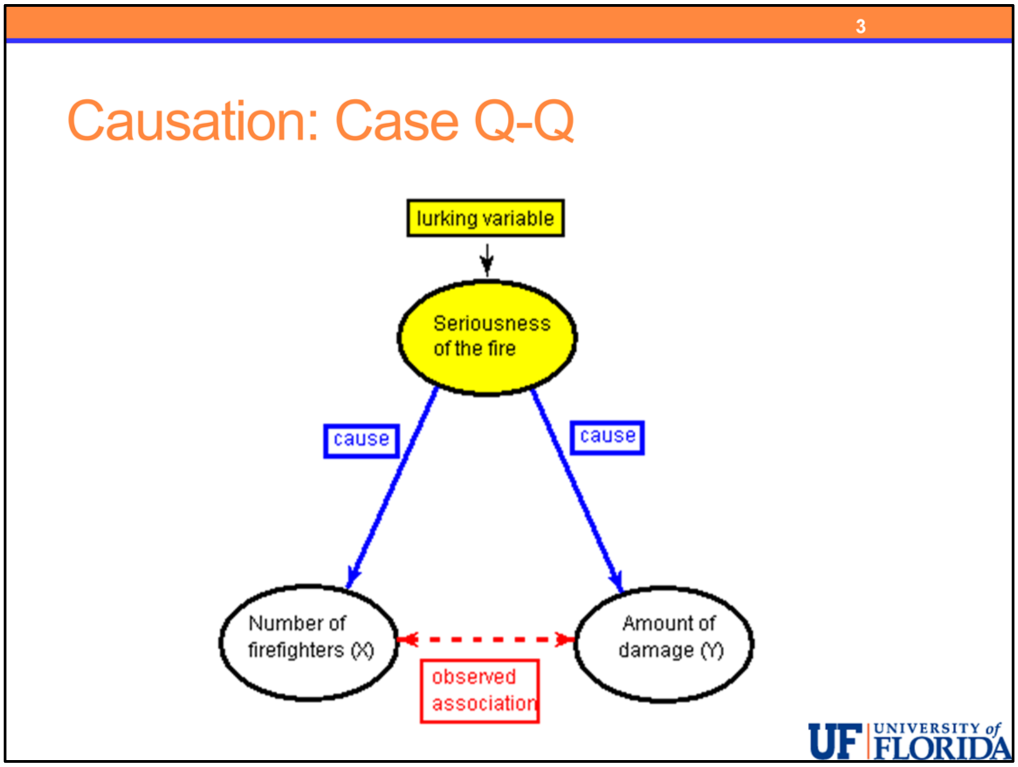
Hopefully it is clear that this is not a reasonable conclusion to the data displayed in this scatterplot.

The questions is – what other variable is the likely CAUSE of BOTH of these variables to be smaller or larger?

In this case, the likely underlying cause would be the severity of the fire. A larger fire would cause more damage and require more firefighters on the scene. A smaller fire would cause less damage and require less firefighters.

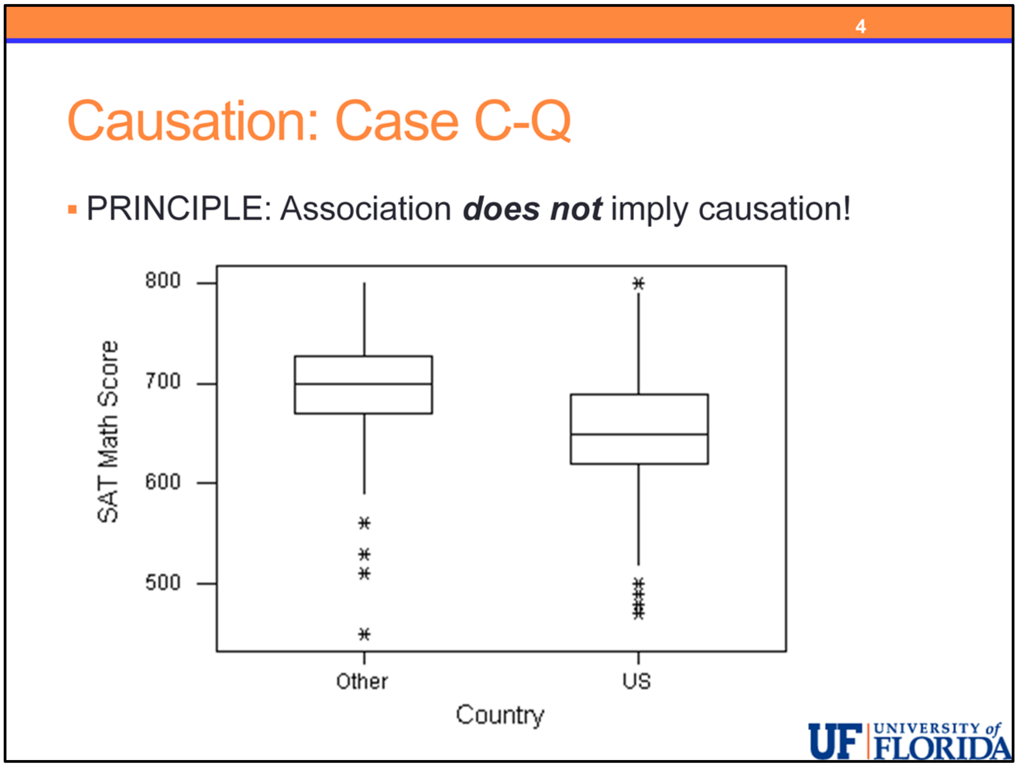
This variable – severity of fire – is called a lurking variable.

A lurking variable is a variable with an important effect on the outcome which is not included among explanatory variables under consideration. It may be known or unknown! It is the unknown lurking variables that are of the most concern in any statistical analysis.



Causation: Case Q-Q

In this case we observed the association between the number of firefighters and the amount of damage but we did not observe the cause.



4

Causation: Case C-Q

- PRINCIPLE: Association ***does not*** imply causation!

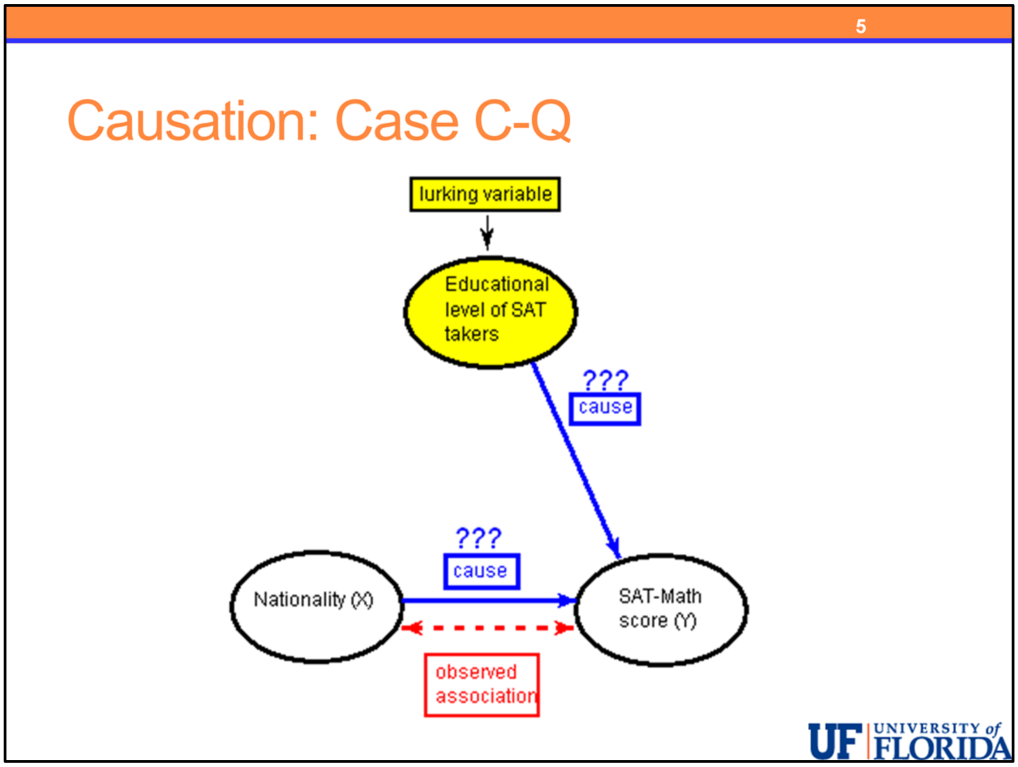
Here’s another example. We have SAT Math scores from US students and those from other countries.

In the boxplot, we observe that on average, students from other countries score higher on the SAT Math.

Is there a causal effect of country on SAT math scores?

Maybe! It is entirely possible that US students are less prepared in mathematics, although it is doubtful that is the case in a comparison to all other countries.

One issue could be that US students are forced to take the test in high school if they are planning to go to college. However, students from other countries would only take the SAT if they were applying to colleges in the US and thus it may be a more select group of students who are taking the test.



5

Causation: Case C-Q

lurking variable

Educational level of SAT takers

???
cause

???
cause

Nationality (X)

SAT-Math score (Y)

observed association

UF UNIVERSITY of FLORIDA

What we don’t know is the educational level of the test takers. Are we comparing apples to apples or apples to oranges?

Here we observed an association between nationality and SAT Math score and nationality may have a causal relationship with SAT Math score. But there may also be one or more lurking variables which are contributing factors.

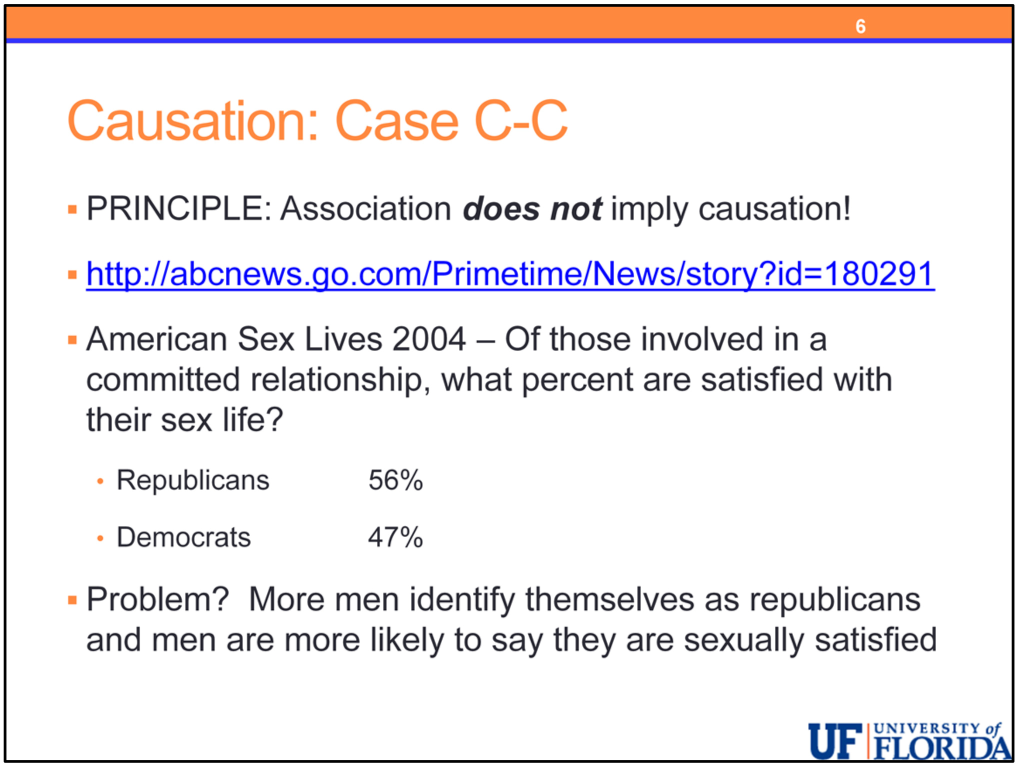
Since no potential lurking variables were measured, we cannot isolate the effect of nationality alone on SAT Math score. When this happens we say that the lurking variable is counfounded with the explanatory variable, nationality.

We might say: The educational level of the test takers is a potential counfounder of the association between nationality and SAT Math score.

The conclusion: An observed association between two variables is note enough evidence that there is a causal relationship between them.

The Important Principle: Association does not imply causation!

Maybe you have previously heard “Correlation does not imply causation” and although this is fine as correlation is another general term for an association or relationship, however, often the term correlation is used to refer to the specific case of linear relationship between two quantitative variables as in Pearson’s correlation coefficient.



6

Causation: Case C-C

- PRINCIPLE: Association ***does not*** imply causation!
- http://abcnews.go.com/Primetime/News/story?id=180291
- American Sex Lives 2004 – Of those involved in a committed relationship, what percent are satisfied with their sex life?

Republicans	56%
Democrats	47%

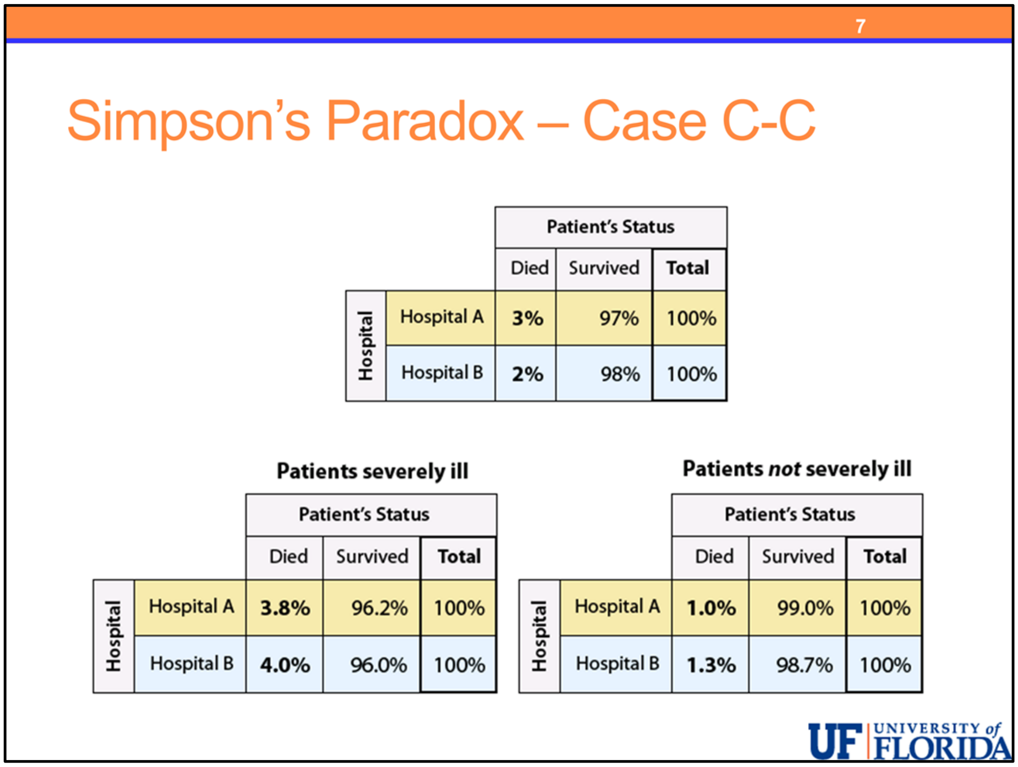
- Problem? More men identify themselves as republicans and men are more likely to say they are sexually satisfied

UF UNIVERSITY of FLORIDA

Weblink: {http://abcnews.go.com/Primetime/News/story?id=180291}

In an ABC news poll reported that among those involved in a committed relationship, republicans were more likely than democrats to be satisfied with their sex lives.

However, more men identify themselves as republicans and men are more likely to say they are sexually satisfied. Is political affiliation the cause or gender or some combination?



Simpson's Paradox – Case C-C

		Patient's Status		
		Died	Survived	Total
Hospital	Hospital A	3%	97%	100%
	Hospital B	2%	98%	100%

Patients severely ill

		Patient's Status		
		Died	Survived	Total
Hospital	Hospital A	3.8%	96.2%	100%
	Hospital B	4.0%	96.0%	100%

Patients *not* severely ill

		Patient's Status		
		Died	Survived	Total
Hospital	Hospital A	1.0%	99.0%	100%
	Hospital B	1.3%	98.7%	100%

In the previous examples we didn't discuss what would be revealed by adjusting for those lurking variables, only discussed the difficulties with claiming the results implied a causal effect.

In this example, we will illustrate adjusting for a lurking variable and the result will emphasize the reason to be concerned.

The overall results of this data show that among hospital A's patients, 3% died and among hospital B's patients, 2% died. This indicates that the survival rate for hospital B is slightly better than that for hospital B. This is certain a true statement. But if we were to jump to the conclusion that hospital B is a "better hospital," we would be incorrect.

In the breakdown below, we see that once we stratify our analysis by whether the patient is severely ill or not, we see that, in fact, the survival rate in hospital A is slightly better than that for hospital B for both severely ill and not severely ill patients.

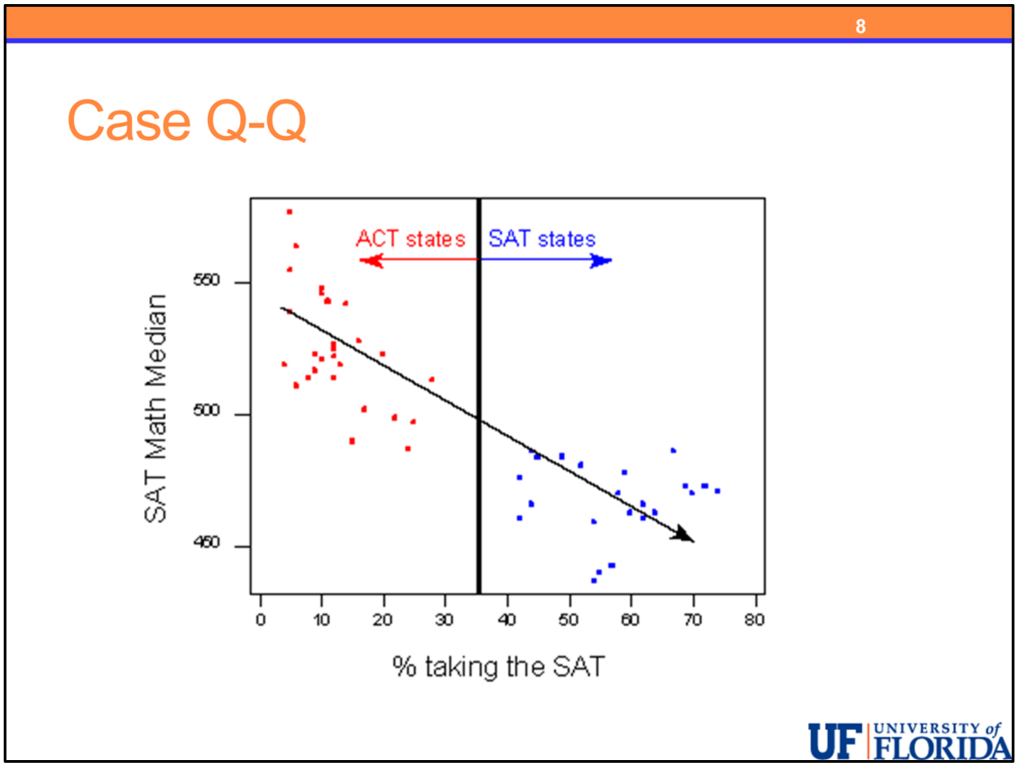
The reason the initial table shows that overall hospital A has a lower survival rate is that the change of death for severely ill patients is greater (we can see this in the table below) and that hospital A had more sevelerly ill patients (which is not illustrated on this slide).

In the materials you can find the raw totals and see that both hospitals had 600 patients who were not severely ill but hospital A had 1500 severely ill patients compared to 200 for hospital B.

Although this data is entirely fabricated, it illustrates, not only the problem of lurking variables in general but that it is possible that our conclusion can be entirely the reverse of what the unadjusted analysis suggests.

This is an example of Simpson’s Paradox, where after we consider the lurking variable, not only is our conclusion altered, but it is REVERSED!

This can occur in any of the cases we have studied.


8

Case Q-Q

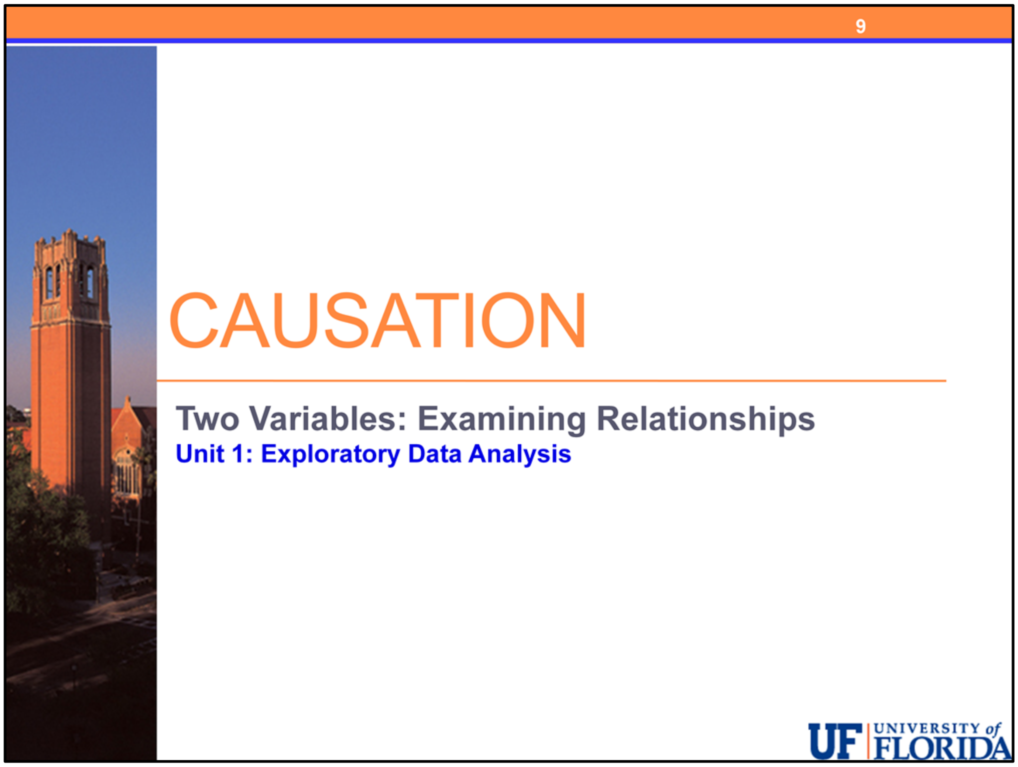
One final example. In this case the observed association is not reversed when we consider the lurking variable but considering the lurking variable helps us better understand the overall trend.

Although it does make sense that when a smaller percentage of students are taking the SAT, the scores would be higher due to the fact that it is a more select group of students taking the exam. We can see that in the SAT states there isn’t much of a negative trend.

In the SAT states, increasing the percentage doesn’t seem to decrease the scores.

In the ACT states, the negative trend does seem more pronounced.

The combination of these two groups, forms the overall decreasing trends we see in the graph.



9

CAUSATION

Two Variables: Examining Relationships

Unit 1: Exploratory Data Analysis

UF UNIVERSITY of FLORIDA

Hopefully we have convinced you that we must be careful when we claim a causal effect and be aware that it may be difficult to account for all possible counfounders in our statistical analysis. We will see that a well designed experiment is our best option to solve these problems.

We will also learn that there are two main ways to account for lurking variables.

In experiments, we control for lurking variables by randomizing subjects into our treatments.

In observational studies, we can control for lurking variables by measuring them and stratifying our analysis (or adjusting for them in other ways in more advanced methods).

We will elaborate more on these issues in the next unit on Producing Data.

In Unit 1 we covered the concepts and methods of exploratory data analysis for one variable and for relationships between two variables. We defined the role-type classification and concluded with this discussion on the pitfalls of lurking variables in proving causal relationships between variables.