1.09 Tests for two groups: Controlling for other variables

In this video I'll discuss how we can take into account the effect of a control variable on the relation between the independent and dependent variable. We'll look at experimental and statistical control of possibly confounding variables and see how to check if a control variable is in fact a confounder and provides an alternative explanation for an effect on the dependent variable.

Suppose I want to test whether a raw meat diet is healthier for cats than canned food. I ask cat owners to indicate whether they feed their cat raw meat or canned food. Cat health is measured on a scale between zero and ten by a veterinarian.
What variables should be included as control variables? Well a control variable can only affect the relation between the primary variables of interest when it is related to both the independent and the dependent variable.

A possible confounder here is the owners concern for their pet's health. Suppose owners who are more concerned with their pet's health, more often choose a raw meat diet. Suppose they also take better care of their cat (providing more stimulation, more regular visits to the vet and higher quality food regardless of type). If health concern is related to the type of diet and the cat's health, it might entirely explain the relation between diet and health. If this is the case, the observed relation between diet and health is called **spurious**, it can be explained by a common cause.

One way to eliminate health concern as a confounder is to control it **experimentally**. We could make sure we only let owners participate that have the same, medium concern for their cat's health. We have turned health concern into a constant. Since health concern is the same, it cannot be related to diet and cannot affect the relation between diet and health.
Another solution is to assign cats to the diets randomly, relying on randomization to eliminate any relation between diet and health concern.
One of the goals in **experiments**, especially lab experiments, is to eliminate as many potential confounders as possible by using randomization and keeping extraneous variables constant. Of course we're not always able to control variables experimentally.

Another way to take potential confounders into account is by controlling them **statistically**. Statistical control means measuring the **control variable,** and checking whether the relation between the independent and dependent variables holds at each level of the control variable.

Suppose we are unable to keep health concern constant or randomly assign cats to the diets. We could measure health concern in owners and categorize their health concern as low, medium or high.
We can now statistically control for health concern by seeing if the mean health scores differ in the low, medium and high concern levels. Of course this requires that we have enough observations in each three concern levels.

If the raw meat group has a similarly higher mean health rating for all three health concern levels, then health concern does *not* provide an alternative explanation for the observed relation between diet and health. If the difference in mean health rating between the raw meat and canned food group changes or disappears for one or more of the health concern levels, then there are three possibilities.

First, health concern is a true confounder, a common cause of both the independent and dependent variable, resulting in a spurious relation between diet and health that disappears once we take health concern into account.

Secondly, health concern can **moderate** the relation between diet and health, meaning the relation becomes weaker or stronger, depending on the level of health concern. For example a raw meat diet is indeed healthier, but more so when the owner is more concerned with health, for example because more concerned owners make sure the cat eats the right portions which makes the health benefits of raw meat more effective.

Thirdly, health concern can **mediate** the relation between diet and health, meaning diet does not influence health directly, but only indirectly through health concern. The example is a bit far-fetched here, but feeding a raw meat diet requires owners to educate themselves about food preparation and portioning. This might lead owners to become more concerned with their cats health, leading to other changes (such as portion control and more visits to the vet), which are in fact responsible for an increase in cat health. Diet influences health indirectly *through* the intermediate variable health concern.

The relation between the variables of interest can change drastically when taking a control variable into account. It's even possible for the relation between the variables of interest to reverse completely. We refer to this phenomenon as Simpson's paradox.

For example, if we look at the relation between health and weight in a group of four-year-old cats we might find this scatterplot, suggesting a positive relation, heavier cats are healthier.
However, if we take into account the cats sex as a control variable, we see that if we consider female and male cats separately we find a negative relation; for *both* sexes heavier, more obese cats are less healthy. Since the females are less healthy than the males and they also weigh less, we find a spurious positive relation between health and weight when we ignore sex.

Here's another example. Suppose we approach two hundred owners of cats with urinary problems. A hundred were diagnosed and prescribed a raw meat diet; a hundred were diagnosed and prescribed special canned food. We assess if the cat's health improved after feeding the prescribed diet. Suppose we find that 24% of cats on a raw diet improved, compared to 31% on a canned food diet. This would suggest advising canned food in the future.

Now suppose that the prescription of raw versus canned food is related to the 'seriousness' of the urinary problems and that seriousness is also related to health improvement. We see that for both serious and normal health issues the raw diet outperforms the canned diet.
The overall percentage of improvement for the raw meat diet is greatly lowered because there are so many serious cases - obviously with a smaller chance of improvement - put on this diet.
This is a typical example of Simpson's paradox. The relation reverses - the raw diet suddenly outperforms the canned food diet for not just one, but *both* levels of the control variable - once the control variable is taken into account.

Here we've looked at categorical control variables, but control variables can also be quantitative. For example, we could have used a health concern rating between 1 and 10, measured using a questionnaire. Of course with a quantitative control variable it's no longer feasible to check the relation between the variables of interest at each individual value of the control variable. However, quantitative control variables can be added as a covariate in regression analysis or analysis of variance to check their influence. These techniques will be discussed in later chapters, although we won't go explicitly into covariates there.