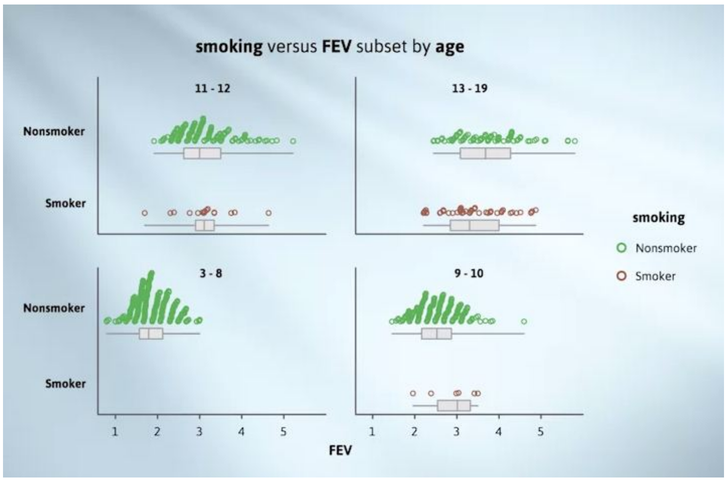
DATA TO INSIGHT: AN INTRODUCTION TO DATA ANALYSIS

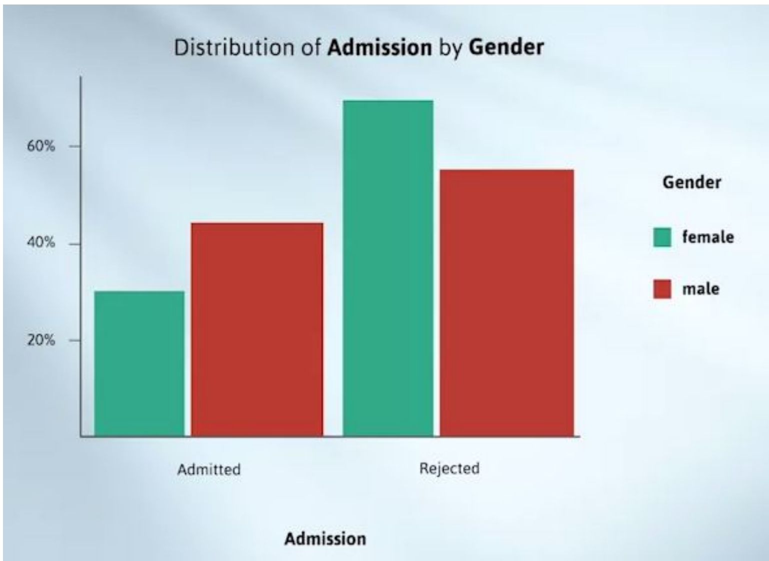
THE UNIVERSITY OF AUCKLAND

WEEK 5

CAUSATION AND CONFOUNDING, PART II by Chris Wild

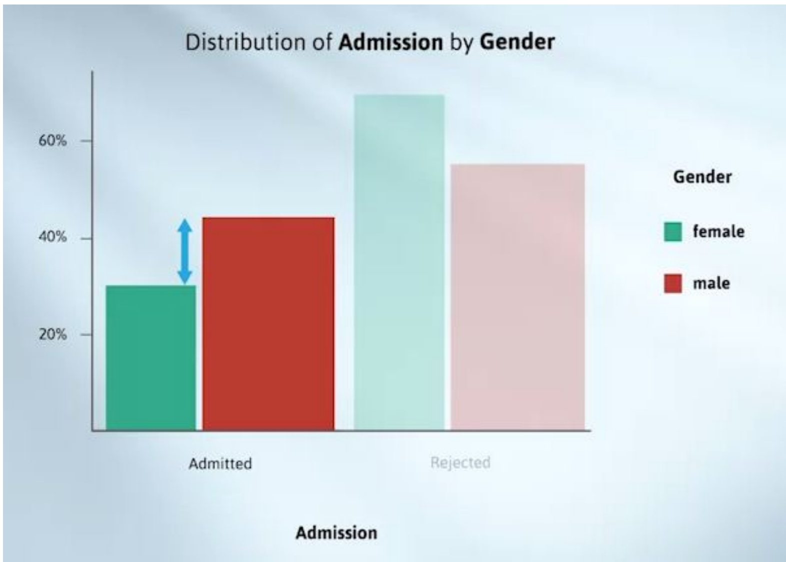


In our previous example, plotting smoking against lung function, we've seen how a confounder or lurking variable can give you a false impression.



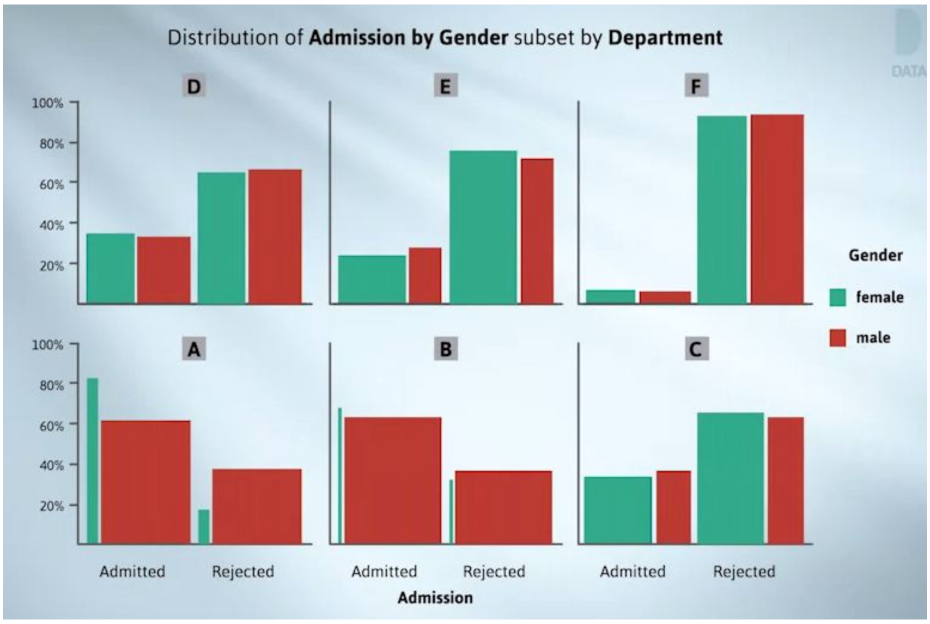
Here's a similar problem with categorical variables. We've using data that is extremely famous in this context. Our data concerns the admissions to graduate study at the University of California, Berkeley, in the fall quarter of 1973.

We are using the four-and-a-half-thousand- odd admissions where the department being applied to is known. We'll concentrate on "being admitted" rather than "being rejected", so we'll push that into the visual background.

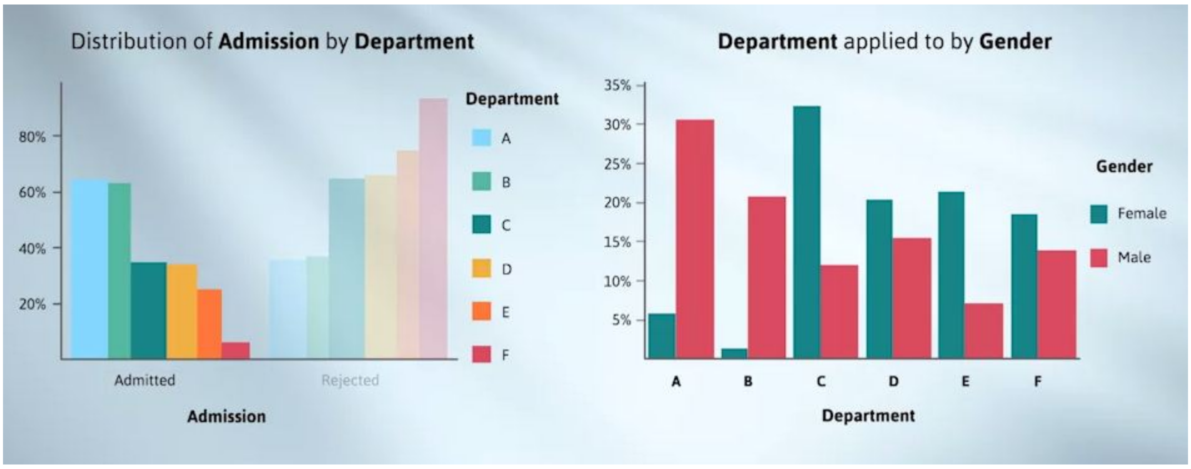


When we look at admission decisions by gender, we see that nearly 45% of men were admitted to graduate study versus only about 30% of women. This looks a lot like gender discrimination.

What's going on? Why are female admission rates lower overall?

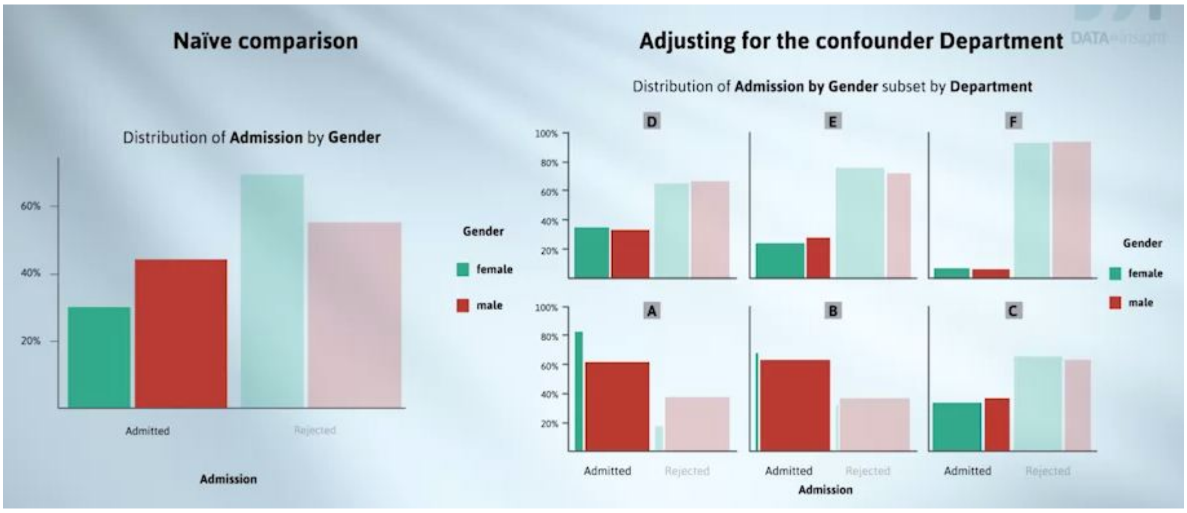


The keys are in the graph. Departments A and B have much higher admission rates than the rest, and very few females apply to them. See how narrow the bars are?



Here we're looking at admission percentages by department. We're still concentrating on being admitted, not being rejected. Most females are applying to the departments C to F, which have low admission rates, whereas many more males are applying to departments A and B, which have very high admission rates. So that's why females had lower admission rates overall.

The basic message in this example is the same as with the lung function and smoking example. We got a false impression when we just looked at the relationship between the two variables on their own.



We needed to adjust for the effect of an important confounder-- in this case, department. In the context of categorical data, the problem we are looking at is called Simpson's paradox.

The basic idea behind all confounder adjustment methods is to do our comparisons within subsets chosen to have similar values of the confounder. That's what we've just done with departments and did previously with age groups. In this video, we've used the simplest version of this idea in that we actually formed the subsets. But where there are several confounders, we can only subdivide the data into subsets so many times before we run out of data, unless we're working with enormous data sets.

Statisticians have developed sophisticated ways of trying to accomplish the same thing for smaller data sets. Tools include matched sampling and regression modelling, but they all involve making some fairly strong assumptions.

Adjustment is helpful in trying to estimate the real effects of an exposure like smoking. However, we can only adjust for confounders we have thought of and collected data about. Consequently, we can never completely solve the causation problem. There is always the chance that effects we think we are seeing is the result of a confounder we have never even considered.

The alternative name for a confounder is a lurking variable. This is a great name, especially with confounders that we didn't think to allow for. It conjures up images of walking home on a dark night with a mugger lurking in the shadows to attack from behind. But with unknown lurking variables, we'll probably never even know that we've been mugged. That brings us to the end of this video. Next time, Chris will be talking about random error.