THE UNIVERSITY OF AUCKLAND
NEW ZEALAND
Te Whare Wānanga o Tāmaki Makaurau

WEEK 5

WHY WHAT I SEE IS NOT QUITE THE WAY IT REALLY IS by Chris Wild

Welcome back. Well, you've made it past the halfway point, and you're still standing. Congratulations and welcome to week five.

So far, we have simply been concentrating on exploring our data. We didn't pause to worry about the reliability or representativeness of the data we were analysing. Our priority was to give you an appreciation of the potential of data analysis.

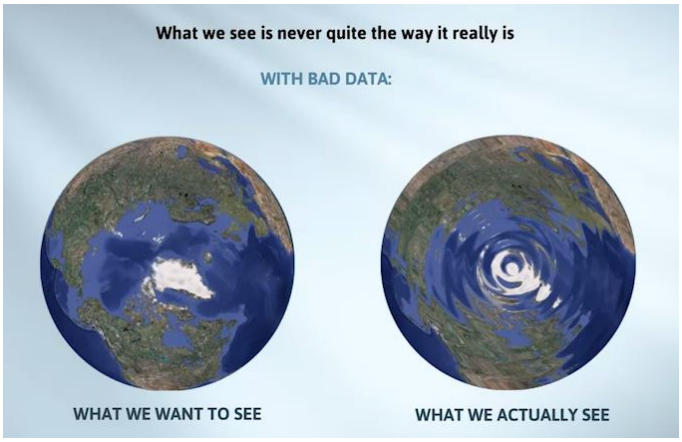
This week, we'll discuss how data can mislead us. International comparisons are a great example of why data sets have to be approached with extreme care.

Different countries often measure things in different ways, and the raw data they collect is of varying quality. So the apparent differences between countries that we see in graphics may not be telling us about the true differences between those countries.

The Gapminder and NHANES websites have extensive documentation on all their data series. They clearly define the meaning of the variables and, for Gapminder in particular, comment about the quality of different pieces of data.

Generally, when we are looking at data, we want to see some aspect of our world, but what we actually see is a somewhat fuzzy approximation where only the biggest features can be reliably made out.

There are statistical ways of assessing and allowing for uncertainties, the fuzziness.



When we have bad data, however, it can be much worse. We still want to see this, but what we end up looking at is more like this, a total distortion of reality. And we may never know that this has happened.

The first law of data analysis is garbage in, garbage out. No amount of sophisticated data analysis can turn bad data into reliable conclusions. If data is really bad, you should just walk away from it.
We'll start the week by talking about how data gets to be bad. Then we will go on to problems and determining whether changes in a predictive factor actually cause a change in outcome. And we'll finish the week talking about random error, with a particular focus on sampling errors.

In real-world data analysis, the analyst is always engaged in a struggle. A struggle to distinguish between fact, patterns that reflect the way things really are in the world, and artefact, artificial patterns caused by inadequacies of the data collection.

Best wishes for week five. This is a pivotal week. It lays out a set of problems and sets up weeks six and seven, where we work towards solutions. The content is much more philosophical. It is a week where you'll spend most of your time thinking about issues rather than doing data analysis. I hope you'll find it stimulating.