DATA TO INSIGHT: AN INTRODUCTION TO DATA ANALYSIS

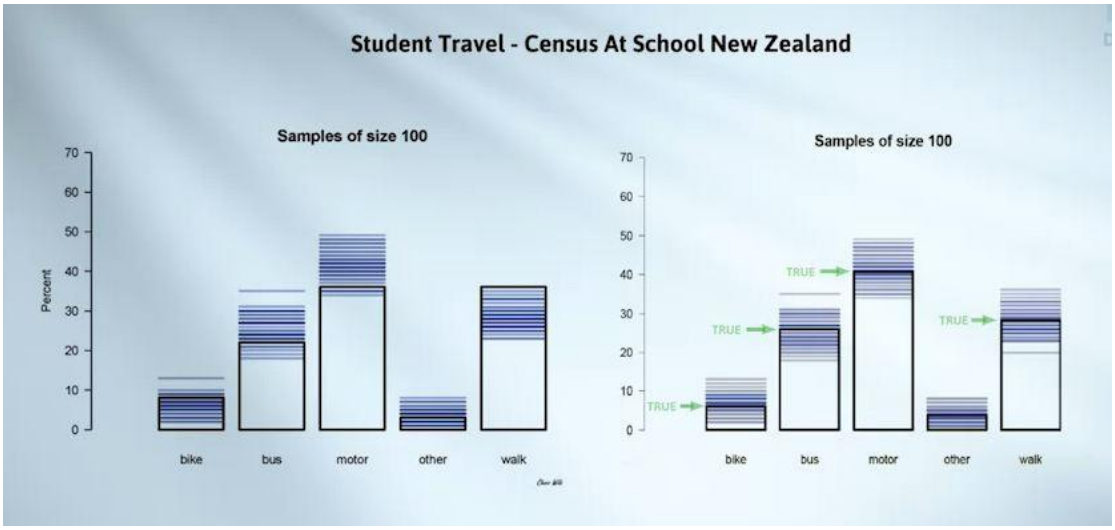
THE UNIVERSITY OF AUCKLAND

WEEK 6

ESTIMATION WITH CONFIDENCE by Chris Wild

Last week, we learned that random sampling is the most reliable way we know of obtaining data about populations without misleading biases. We also learned that it's far from perfect. Whenever we use data from a sample to estimate a population quantity, there will always be an error due to sampling. This week, we'll learn how to put an interval around our estimates to allow for sampling error. These intervals are called confidence intervals.

Do you remember this animation of samples from the CensusAtSchool database?

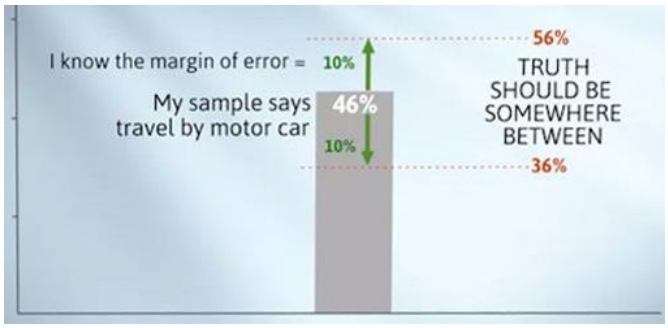


The question was about mode of travel to school. We can see considerable movement in the percentages from sample to sample.

In the static image, on the right the tops of the black bars give us the true percentages for everyone in the database. The blue lines show us the corresponding percentages for all the different samples we've seen.

In the real world, we only take one sample, and we only get one estimate of how high each bar is, one of the blue lines. The samples generate bands of blue around the truth, showing how wrong our estimates can be.

Let's focus on just one response. We'll choose motor, going to school by motor car. The black line is the truth (about 40%). From the blue band, it looks like using the value from a sample could make us wrong by up to about 10 percentage points in either direction. That's the basic idea of a margin of error.



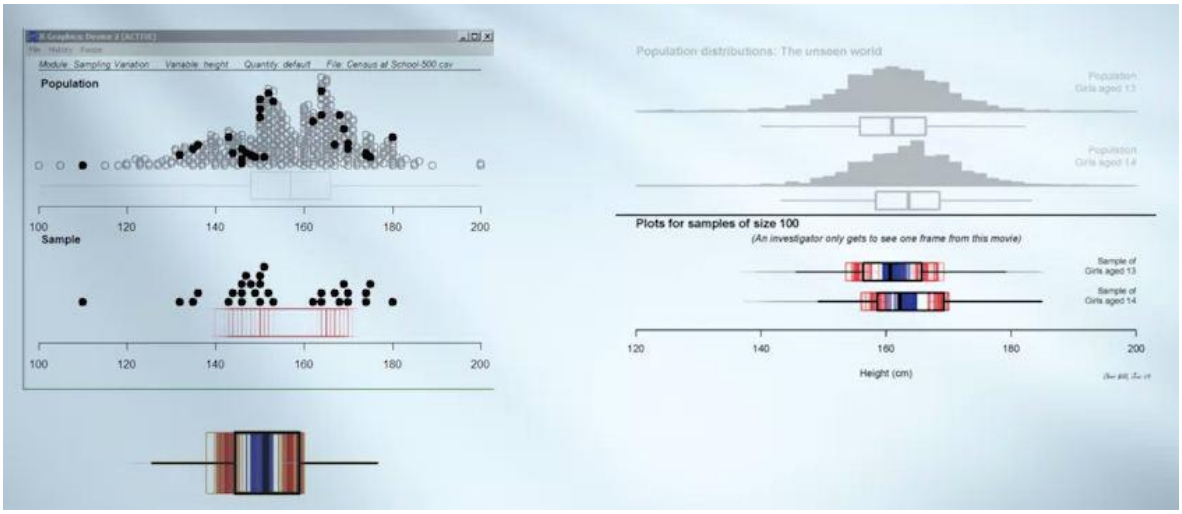
Suppose my sample gave 46% for travelling by motor car. If I all I had was my sample value (46%) but I knew sample values in this situation had a margin of error of about 10%, I'd conclude that the true population value was somewhere between 36% and 56%.

That's the basic idea behind what statisticians call a confidence interval, a range we're pretty sure covers the true value. It's a common sense way of thinking. But to do it, we need to know "What's the margin of error we need to allow for".

We used this slide in the last video of week five. The margins of error we need to allow for get smaller as the samples get larger. But it shouldn't be surprising that more information tends to get us closer to the truth. Percentages taken from samples as small as 30 can have enormous errors. The errors are much smaller if we

have larger samples, like 1,000 (roughly 3 percentage points in the pictures we're seeing here).

Here is sampling error displayed in a few more situations.



Still from animation

They all reinforce the basic idea that whenever we get an estimate from a sample, we want to allow a margin around that estimate. And that margin needs to be big enough to allow for the usual extent of the sampling errors.

We've seen the first statement here before ("All estimates from data are wrong"), and it's true for all practical purposes.

Of course, a wrong answer isn't very useful unless you're fairly sure that it's not too wrong -- close enough to the truth to be useful for whatever decisions you're making. (We always need a "margin of error" that tells us how wrong our estimate could be.)

But here we run into a problem. We want to allow a margin around the estimate big enough to allow for the usual extent of sampling errors, but we never get to see sampling variation movies like these. All we get is one sample. So we can't see the truth, we can't see the other samples, and so we can't see the blue bands, which tell us how wrong we're likely to be.

So how can we get an estimate of the level of sampling error-- the answer to, "How wrong could I be?" -- from just the single sample we have? It looks like an impossible problem, but it isn't.

Historically, statisticians approached the problem of estimating the likely extent of sampling error using mathematical theory.

More recently, computers have enabled us to use a very simple idea that's much easier to understand than the mathematical theories, much more powerful, and much more generally applicable. That idea is called Bootstrap resampling, and that's the subject of our next video.

Other videos will show the construction of confidence intervals in various situations, how confidence intervals work, and we'll add this sort of information onto graphs so that, when we read graphs, we can take account of the levels of uncertainty. So all the best for Week Six.