DATA TO INSIGHT: AN INTRODUCTION TO DATA ANALYSIS

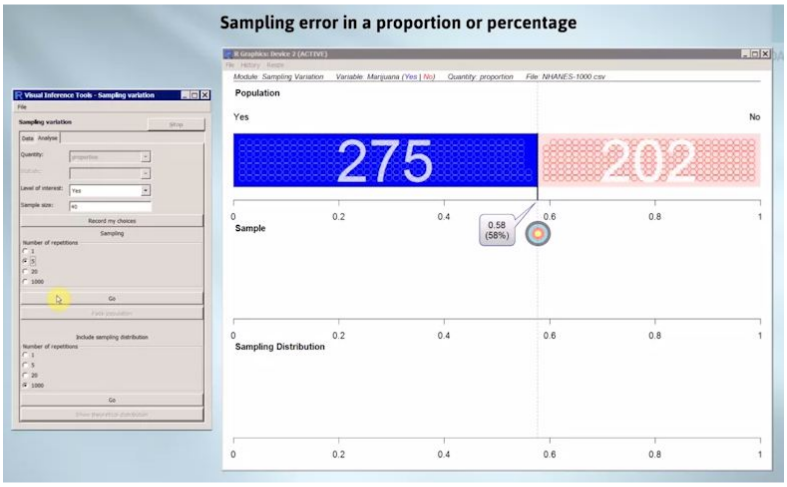
THE UNIVERSITY OF AUCKLAND

WEEK 5

RANDOM ERROR, PART II by Chris Wild

In Part One of this video, we only saw what sampling error behaviour looked like when we used a sample mean to estimate a population mean. If instead of a mean, we'd used a median or even a quartile, we'd get similar behaviour. But a single demonstration is insufficient to hang our discussion of sampling error zones. So let's now look at how sampling error behaves when we use the sample proportion to estimate a population proportion.

Let's use the marijuana question, which was answered by 477 people, of whom 275 (or 58%), said, "Yes, I've tried marijuana".



Sampling error in a proportion or percentage

Module: Sampling Variation   Variable: Marijuana (Yes | No)   Quantity: proportion   File: NHANES-1000.csv

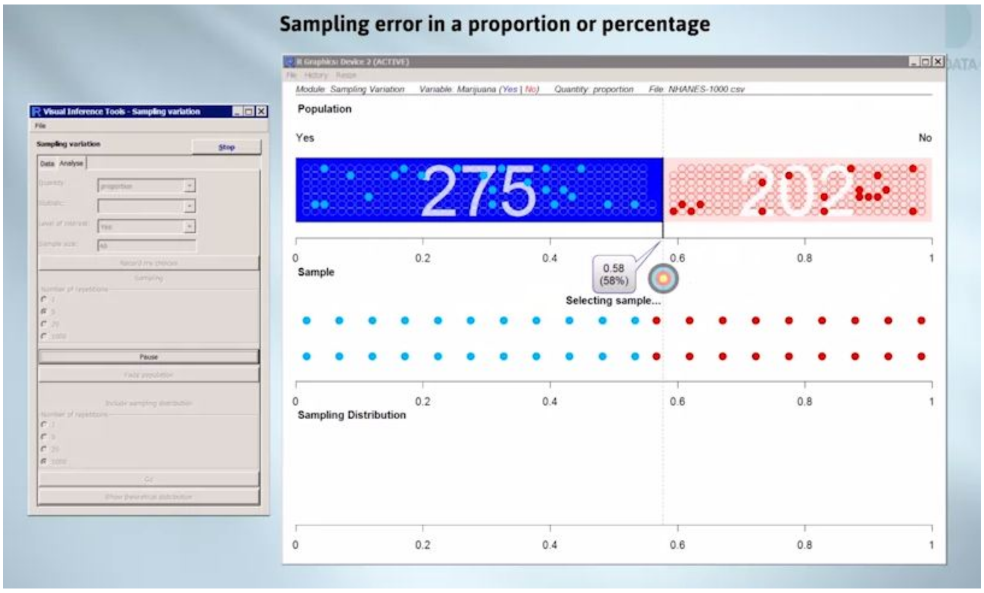
Quantity: proportion

Level of interest: Yes

Sample size: 40

First, to explain the upper display. Again, this is a representation of our population. We've drawn a segmented bar. Each of the 477 people is represented by a small circle.

The 275 people in the population who answered "Yes" are shown in the blue segment. They make up a proportion 0.58 (or 58%) of the population. That's the target we want to hit with our estimate. How well can we estimate this population proportion using just a sample of 40 people? When a person is selected their circle would get coloured in.



Sampling error in a proportion or percentage

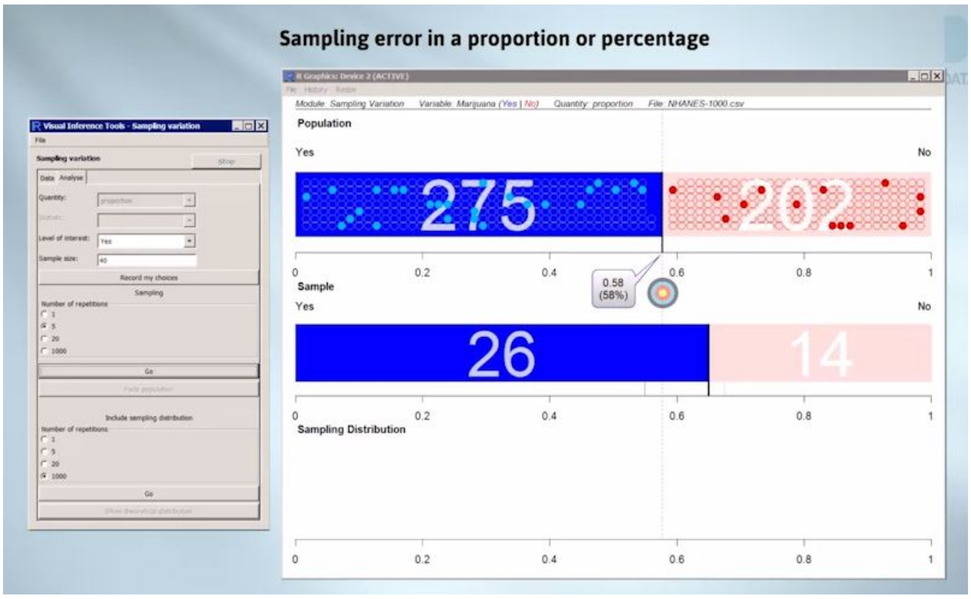
Module: Sampling Variation Variable: Marijuana (Yes | No) Quantity: proportion File: NHANES-1000.csv

Quantity: proportion

Level of interest: Yes

Sample size: 40

Let's now take samples of 40 people from these 477 people and, for each sample taken, we'll see how close we get to the true-value target of 58%. The selected people drop down into the middle panel, and we get the corresponding plot for the people sampled. The proportion of "Yesses" is about 55%, which is not too far from the target value of 58%.



Sampling error in a proportion or percentage

Module: Sampling Variation Variable: Marijuana (Yes | No) Quantity: proportion File: NHANES-1000.csv

Quantity: proportion

Level of interest: Yes

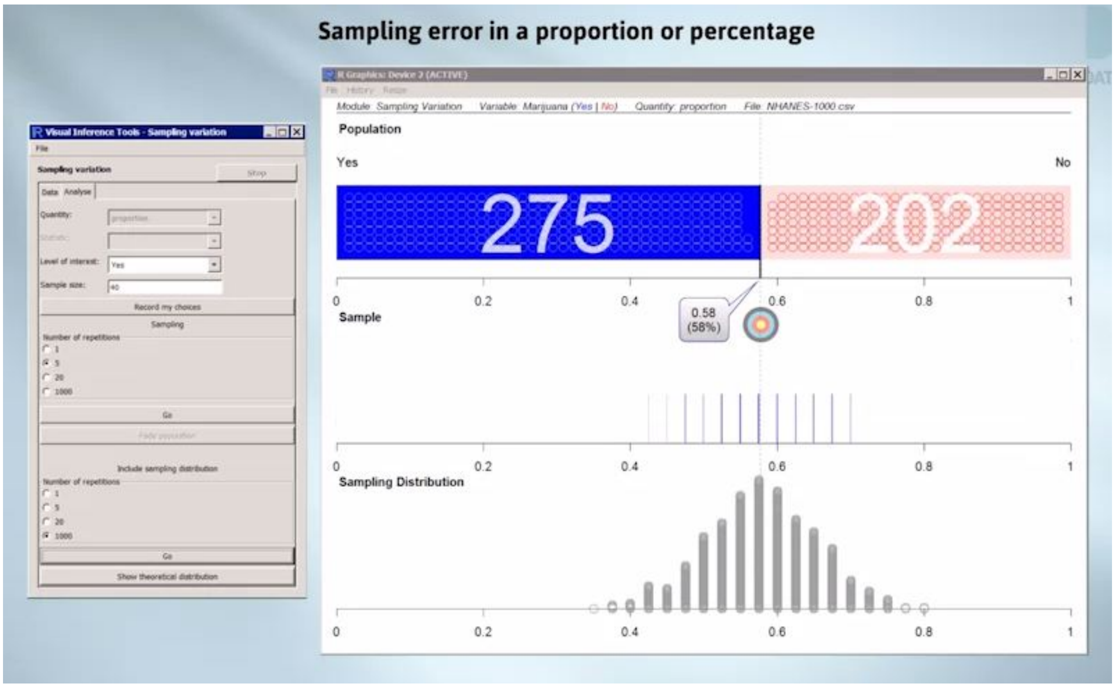
Sample size: 40

Sampling — Number of repetitions: ○ 1 ● 5 ○ 20 ○ 1000

Include sampling distribution — Number of repetitions: ○ 1 ○ 5 ○ 20 ● 1000

We'll do this two more times, leaving footprints from the past samples to give you some time to get familiar with the display. This next one is far too big. A bit better, but still far too big.

Now we'll do it 1,000 times.



Sampling error in a proportion or percentage

Visual Inference Tools - Sampling variation

Sampling variation

Quantity: proportion

Level of interest: Yes

Sample size: 40

Sampling

Number of repetitions

○ 1
● 5
○ 20
○ 1000

Include sampling distribution

Number of repetitions

○ 1
○ 5
○ 20
● 1000

R Graphics: Device 2 (ACTIVE)

Module: Sampling Variation Variable: Marijuana (Yes | No) Quantity: proportion File: NHANES-1000.csv

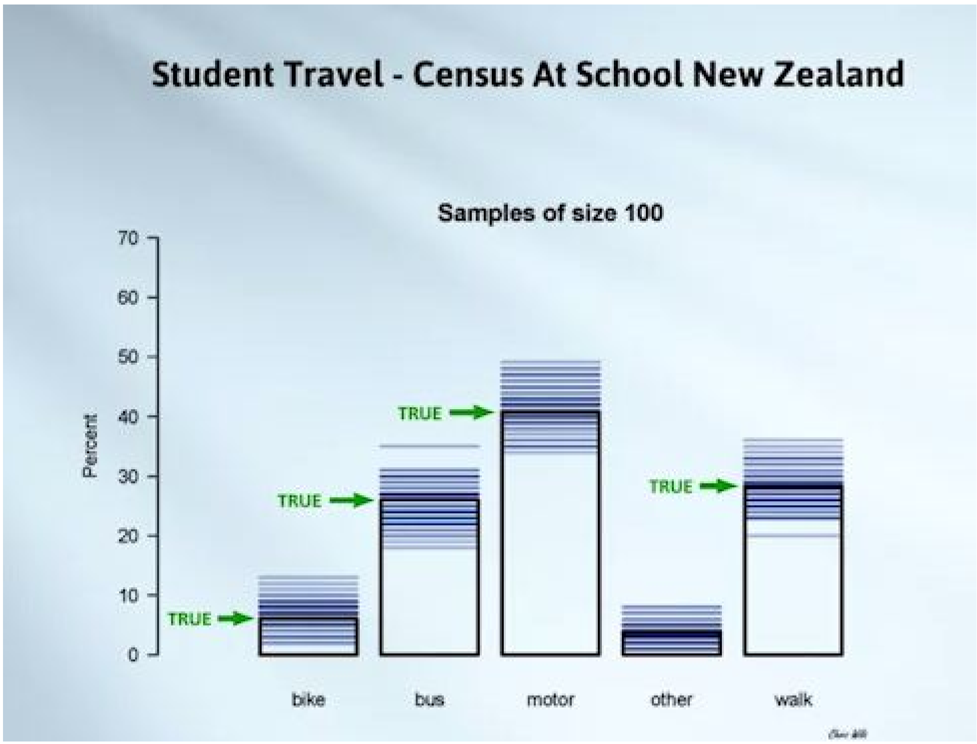
Again, each dot in the bottom panel is the estimate or answer given by one of our thousand samples. We can see the variability and the estimates of the percentage who said "Yes" to the marijuana question.

Some samples have given us values below 0.4, or 40%. Some have given us values up near 80%. They're quite often a long way from the true value of 58%.

What we can see from this example is that when we take one sample of size 40, the sampling error could be huge, probably big enough to make the estimate essentially useless for many practical applications.

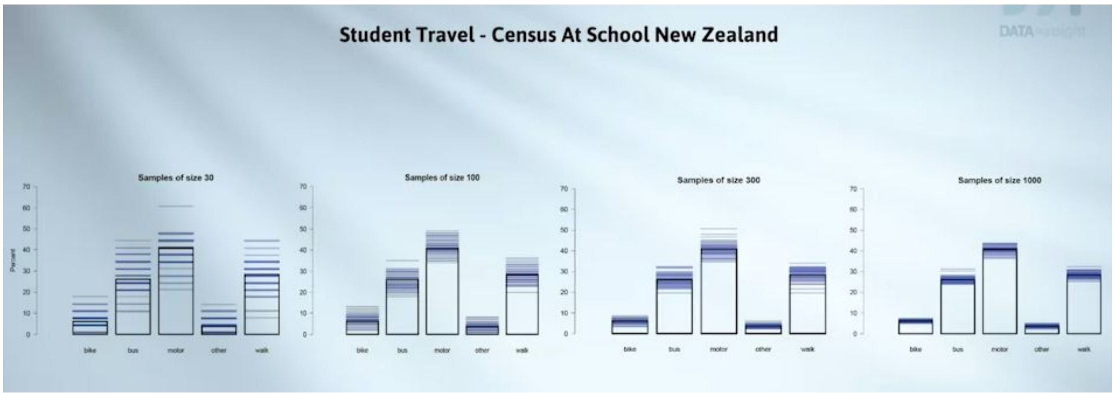
Remember, in reality,

- we don't know the true value of the target
- we only ever take one sample and
- we can never know how big the sampling error is. (We'll learn later how we can estimate how big it could be.)



But sampling errors aren't always as bad as the one we've just seen. Here we are sampling from the CensusAtSchoolNewZealand database. The question asked how students travelled to school. We're seeing bar charts of the percentages using the various modes of travel from repeatedly taking samples of size 100.

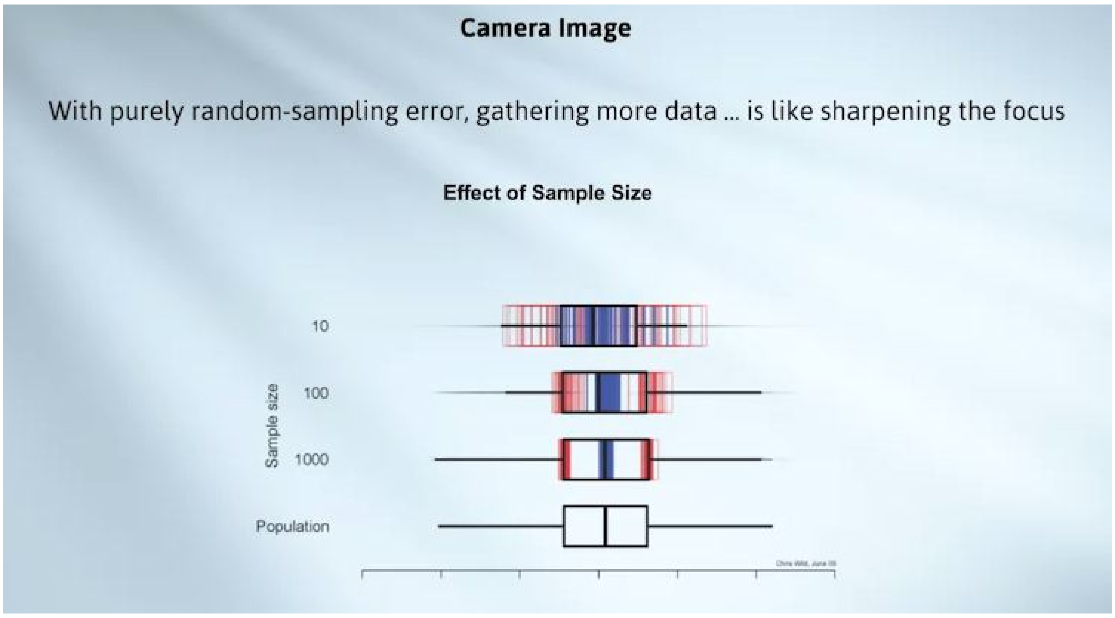
Again, footprints are being left from the results of previous samples so we can see how different the answers these samples give us can be.



In this static image, the black bars show the true percentages in the database. The distances between the blue footprints and the black true values are the sampling errors being made for each sample.

We can see that increasing the sample size reduces how big the sampling error can be. It suggests the following metaphor from photography. When random sampling error is the only source of error, increasing the sample size is like sharpening the focus.

Here's the same thing using box plots of a numeric variable.



Camera Image

With purely random-sampling error, gathering more data ... is like sharpening the focus

Or looking at it another way.

With only random errors

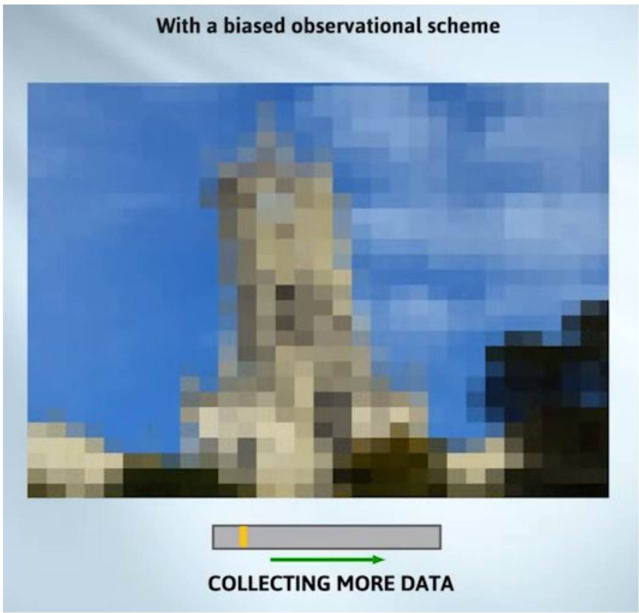
COLLECTING MORE DATA

With only random errors

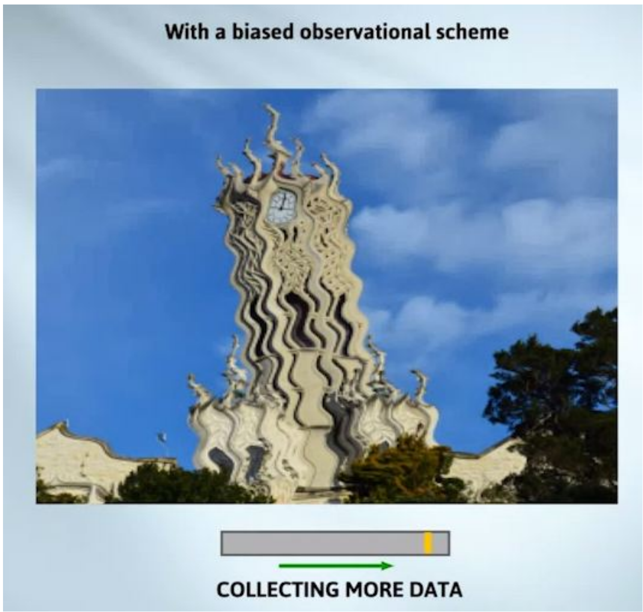
COLLECTING MORE DATA

But if there are biases affecting what data we get to see, gathering more data just gives us a clearer view of a distorted image.

We're not seeing this.



We're seeing this.



I don't want to leave you thinking, "Random sampling's terrible. Look at all the errors we get". Random sampling is still the best way we know of getting data that's not plagued by biases. Everything else tends to be much worse. And with random sampling, we can get a pretty good idea of how reliable our estimates are, whereas non-random selections tend to lead to unknown biases of unknown size. So you can have no idea of how wrong you could be.

Some take-home lessons

- all estimates are wrong.
  Well, we might occasionally get something right completely by accident, but we'll never know when this has happened, so the bullet stands.
- Bad estimates lead to bad decisions.
- An estimate is not particularly useful if we have no idea how wrong it could be.
- Sampling errors get smaller when we take bigger samples, whereas systematic biases don't.

As a rule of thumb for how these things work. We can halve the sizes of sampling errors by taking four times as many observations. But we clearly need ways of determining how wrong our estimates are likely to be. And that's where we're going next week.