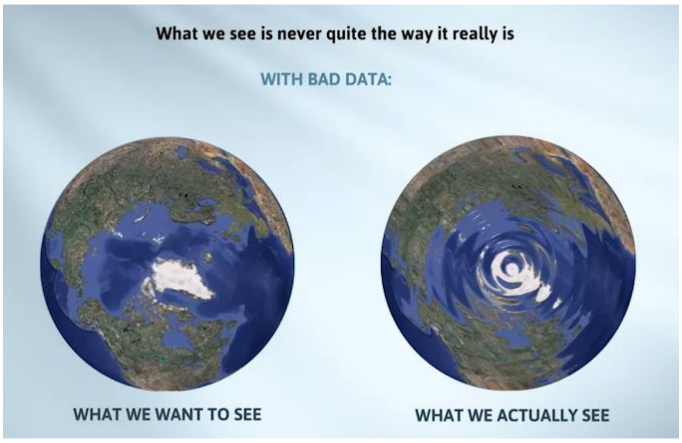
DATA TO INSIGHT: AN INTRODUCTION TO DATA ANALYSIS

THE UNIVERSITY OF AUCKLAND

WEEK 5

RANDOM ERROR, PART I by Chris Wild

Hi. In our earlier videos this week, we looked at ways data can conspire to mislead us. We looked at problems caused by bad measurement systems and biased selection mechanisms. These are data quality issues. They result in the data we're looking at giving a distorted picture of the reality we're trying to understand.

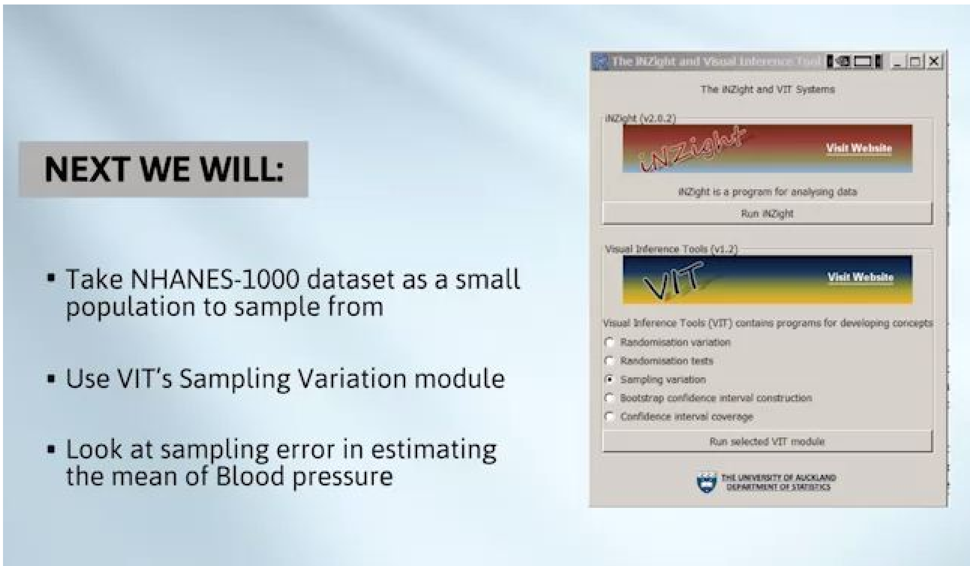


We want to look at this, but the data is showing us this.

Then we looked at the problems of confounding that restrict our ability to draw cause and effect conclusions from observational data. It simply can't be done reliably.

What's common about all these problems is that none of them go away as we get more data. They're big data problems, as well as small data problems.

Random error, the subject of this video, is a small data problem, by which I mean it's a problem in data sets that are not huge. The most reliable ways people know for getting representative data use some sort of random selection mechanism to avoid bias. I'll show you the simplest version of the problem. Later, we'll be able to extrapolate.



NEXT WE WILL:

- Take NHANES-1000 dataset as a small population to sample from
- Use VIT's Sampling Variation module
- Look at sampling error in estimating the mean of Blood pressure

The iNZight and Visual Inference Tool

The iNZight and VIT Systems

iNZight (v2.0.2)

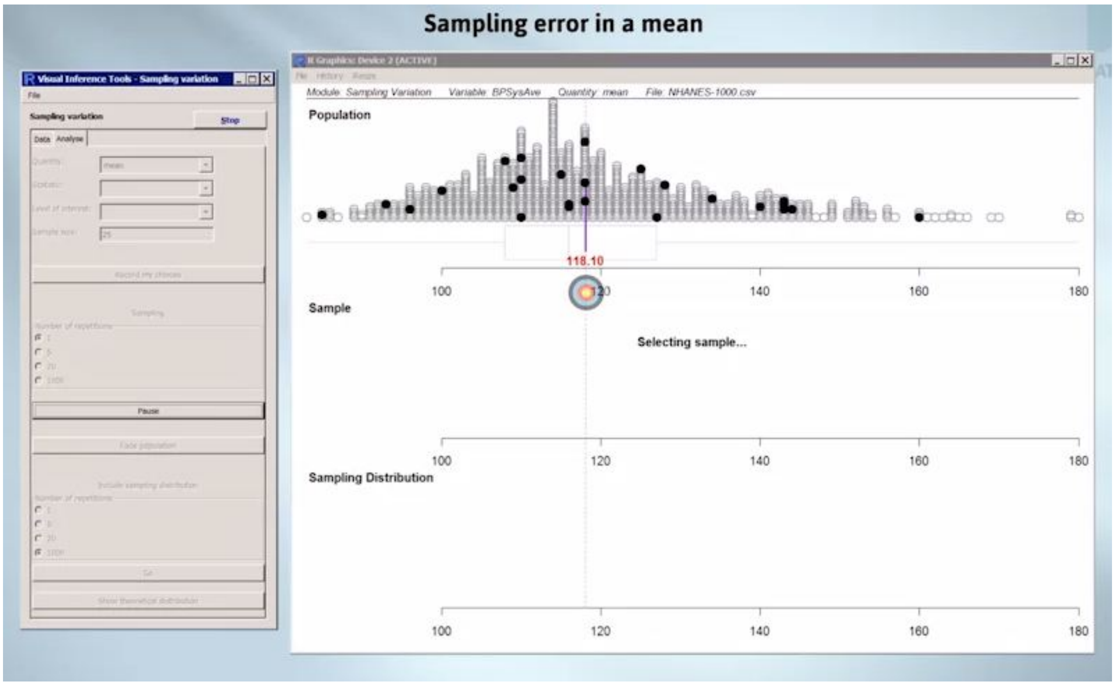
iNZight is a program for analysing data

Visual Inference Tools (v1.2)

Visual Inference Tools (VIT) contains programs for developing concepts

○ Randomisation variation
○ Randomisation tests
● Sampling variation
○ Bootstrap confidence interval construction
○ Confidence interval coverage

We're going to treat the 1,000 people in our NHANES-1000 data set as a small population and we'll sample from it. We want to see how well we can estimate features of this population by using just a small sample from it. I want to keep the number's fairly small so you can see everything that goes on. The animations you'll see come from VIT's Sampling Variation module, activated using the bottom part of the familiar iNZight start-up window.



Sampling error in a mean

Visual Inference Tools - Sampling variation

Sampling variation

Quantity: mean

Sample size: 25

Sampling

Number of repetitions
● 1
○ 5
○ 20
○ 1000

Include sampling distribution

Number of repetitions
○ 1
○ 5
○ 20
● 1000

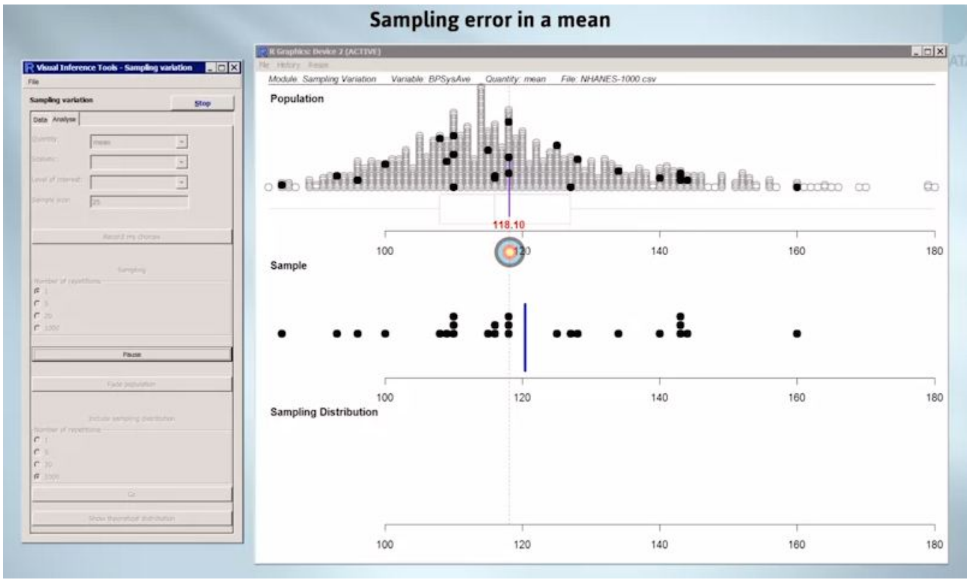
R Graphics: Device 2 (ACTIVE)

Module: Sampling Variation   Variable: BPSysAve   Quantity: mean   File: NHANES-1000.csv

Selecting sample...

At the top of this screen is a dot plot of the systolic blood pressures of everyone in our 1,000 person population. Marked up on the graph is the position of the mean of all these blood pressures. It's 118.1. That's the target we want to hit with our estimate. How well can we estimate the mean blood pressure for this whole population using just a sample of 25 people? How close will it get us to our target of 118.1? Each dot in the dot plot represents a person. We're going to randomly select people, one at a time, until we get 25 of them. When a person gets selected their dot turns black. Here they are being selected now.

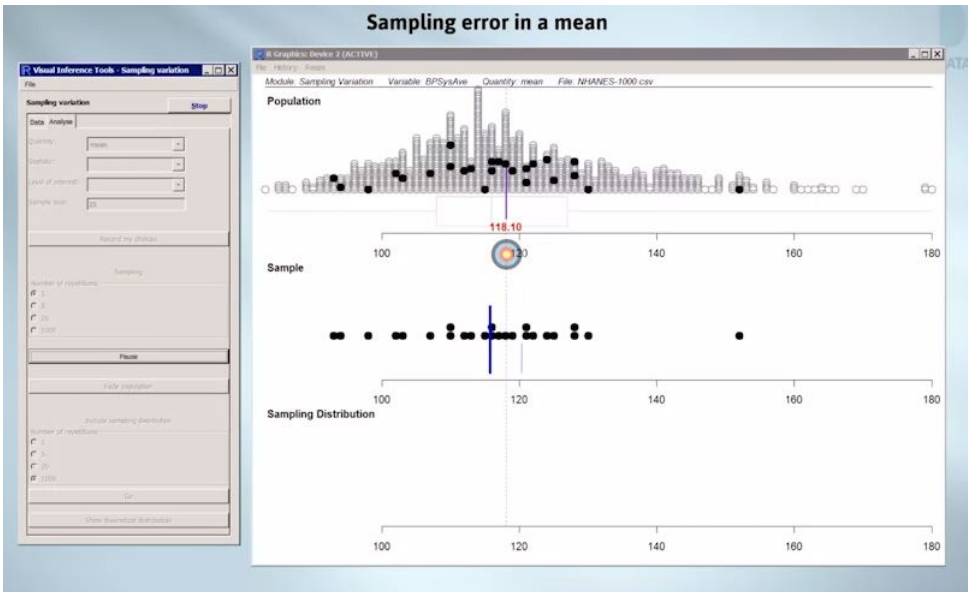
Now we're going to make a plot using just those people selected and mark the position of their mean.



Sampling error in a mean

Module: Sampling Variation   Variable: BPSysAve   Quantity: mean   File: NHANES-1000.csv

Well, we got that wrong. The real mean is 118.1, but our sample gave about 121. We'll try again and see if we can do better.

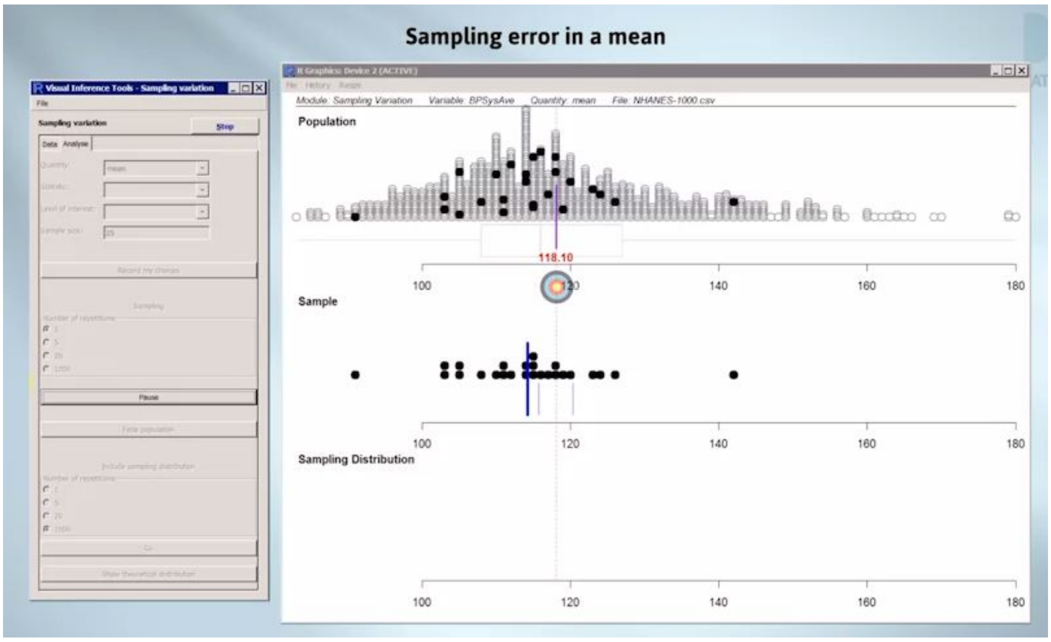


Sampling error in a mean

Module: Sampling Variation   Variable: BPSysAve   Quantity: mean   File: NHANES-1000.csv

Just as bad as last time, but on the other side.

Every time we do this we're going to leave a blue mark at the position of the new sample mean. This will build up a history of all the sample means we've seen. Let's do it a third time.

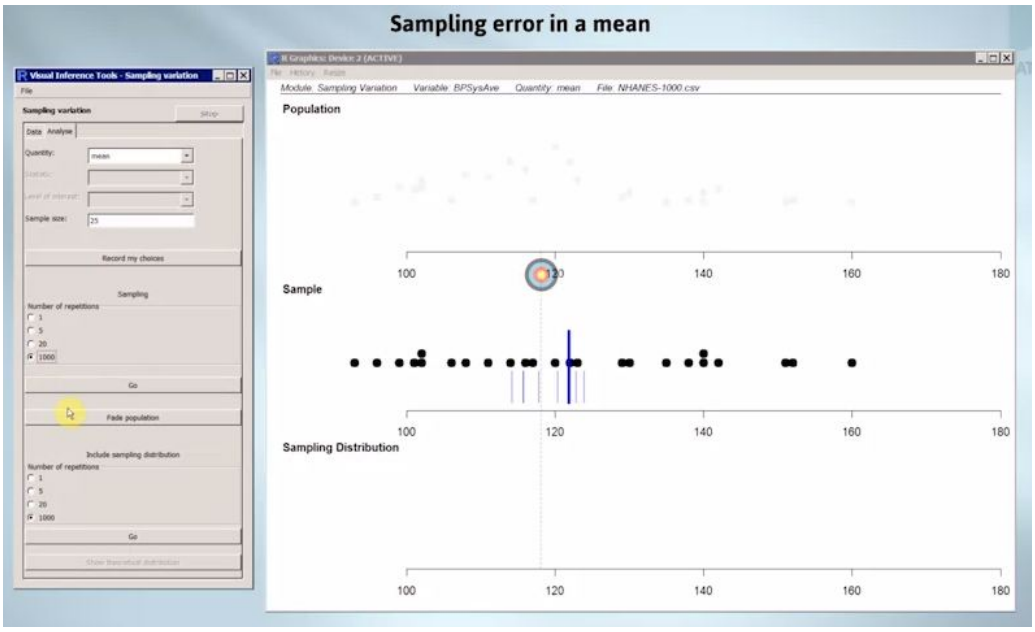


Sampling error in a mean

Module: Sampling Variation Variable: BPSysAve Quantity: mean File: NHANES-1000.csv

Even worse.

Of course, in the real world, we sample when we can't see our population and don't know the real answer. So we'll put a veil over the population to make it a little more realistic and do it five more times.



Sampling error in a mean

Module: Sampling Variation Variable: BPSysAve Quantity: mean File: NHANES-1000.csv

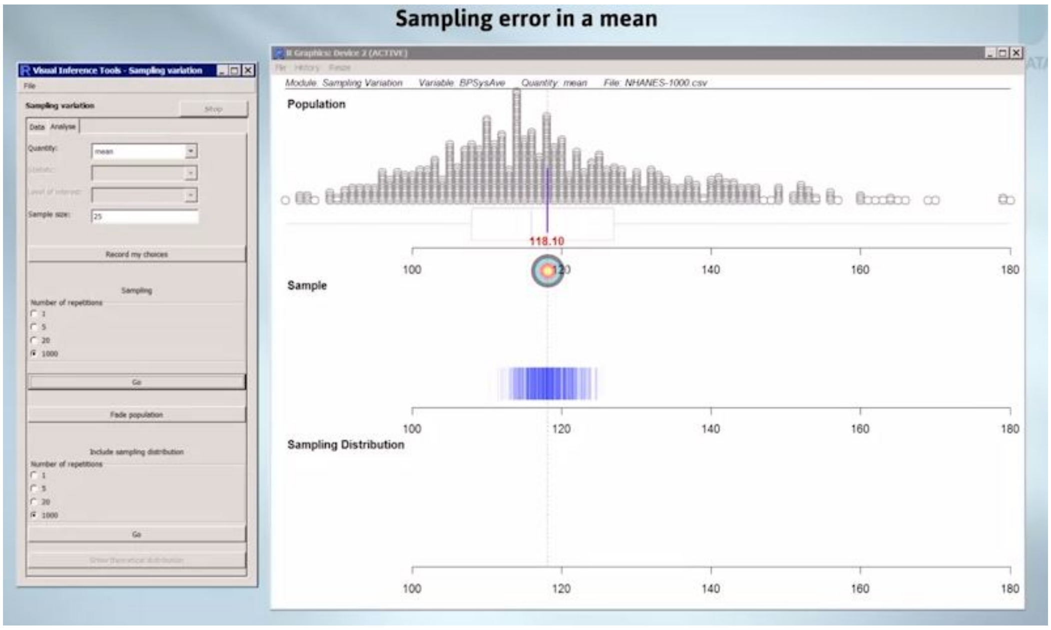
Quantity: mean

Sample size: 25

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

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

Every answer we've got from a sample is wrong. What happens if we do it a thousand times?



Quantity: mean

Sample size: 25

Number of repetitions: ● 1000

Number of repetitions: ● 1000

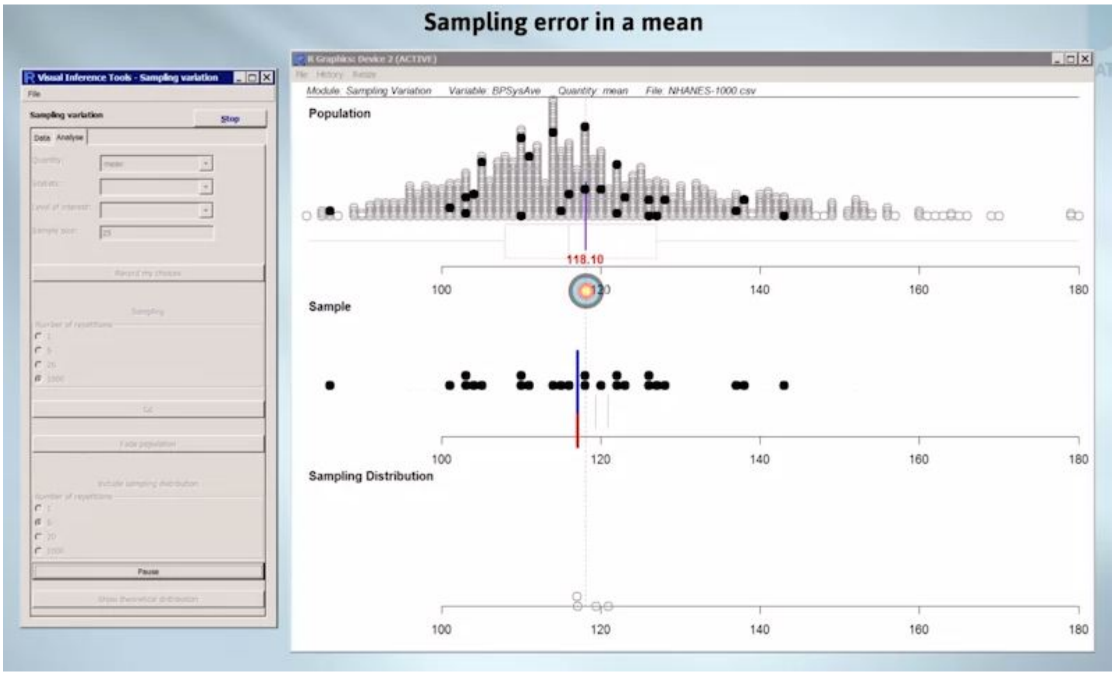
Module: Sampling Variation Variable: BPSysAve Quantity: mean File: NHANES-1000.csv

We can see that almost all of the answers we've gotten are between about 110 and 130. What does this tell us about the process of taking a sample and using it to estimate a population mean?

The first thing is that the answer from our sample is generally going to be wrong. For a wrong answer to be useful to us, we need to have some idea about how wrong it could be.

If I take a sample and calculate its mean, we'll get one of these blue marks. In the real world, we wouldn't see any of the other blue marks, or the true value, so we unlikely to be bigger than 10, say", or put it another way, "The likely error is less than 10". It may be that the likely error is small enough not to matter to us. If the likely error is very big, we'd probably think, "This estimate is too untrustworthy to be useful".

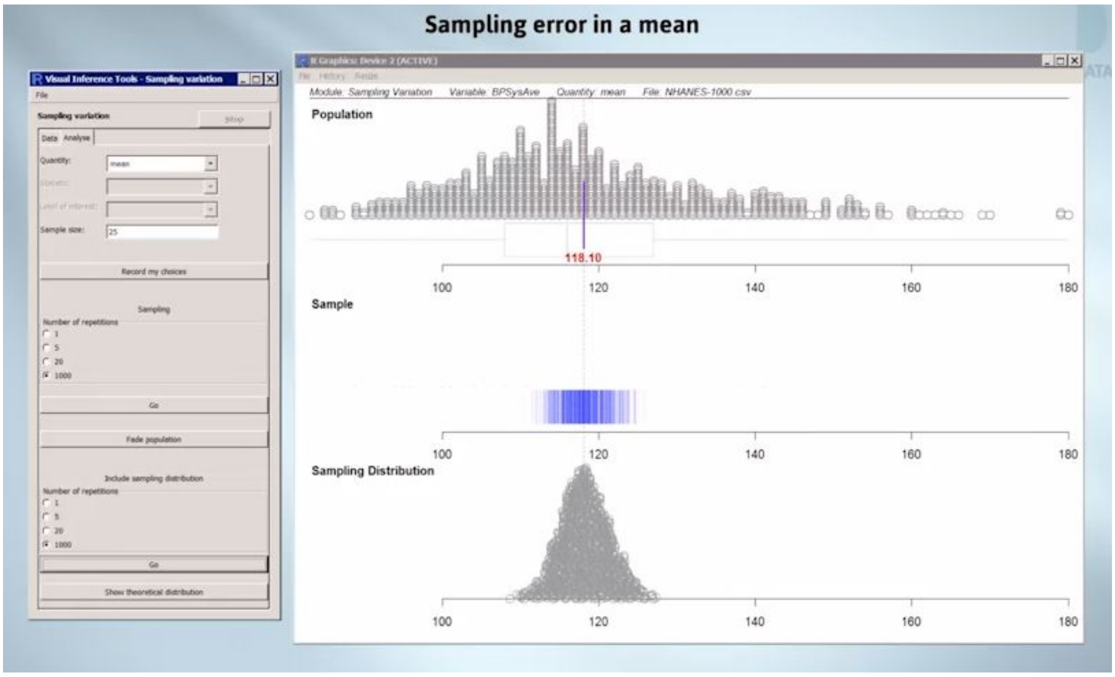
Before going on, I'm also going to capture all the answers we're getting in another way that is more conventional in statistics.



Module: Sampling Variation Variable: BPSysAve Quantity: mean File: NHANES-1000.csv

One example

I'll do a few slowly till you get the idea.



Module: Sampling Variation Variable: BPSysAve Quantity: mean File: NHANES-1000.csv

Quantity: mean

Sample size: 25

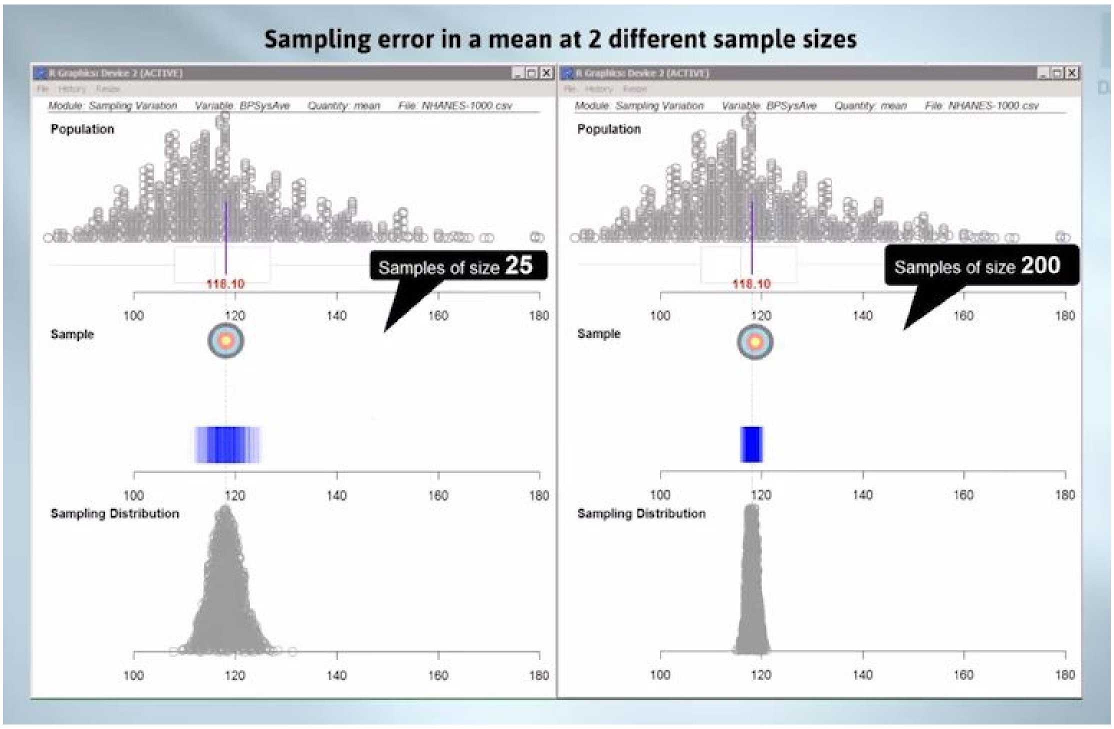
Number of repetitions: ○ 1 ○ 5 ○ 20 ● 1000

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

A thousand times

Now we'll do it 1,000 times. In the middle representation each blue bar is the answer given by one of our samples. The second representation is a dot plot. Each dot, or circle in the heap, is likewise the answer given by one of our samples.

Next, I'll try sampling using samples of 200-people. And I'll show it to you alongside the smallest samples of 25-people to make a point.


Sampling error in a mean at 2 different sample sizes
R Graphics: Device 2 (ACTIVE)
File History Resize
Module: Sampling Variation Variable: BPSysAve Quantity: mean File: NHANES-1000.csv
Population
118.10
Samples of size 25
100
120
140
160
180
Sample
Sampling Distribution
R Graphics: Device 3 (ACTIVE)
Samples of size 200

The good news is that the set of answers we've got with the bigger samples are generally much closer to the truth. We're making small errors. This illustrates a general principle: with random sampling errors the bigger the sample we take, the smaller the errors we make. So where possible, we can reduce error by increasing the size of our sample.

Unfortunately, collecting more data generally costs more money. Sampling errors are the errors we make when we use sample results to estimate population quantities. We've just seen what this looks like when we try to estimate the mean blood pressure in the whole population using the mean blood pressure from a sample. For any particular sample we take, we can never know how big our sampling error is. The best we can do is estimate how big it's likely to be. We'll see ways of doing that next week.