DATA TO INSIGHT: AN INTRODUCTION TO DATA ANALYSIS

THE UNIVERSITY OF AUCKLAND

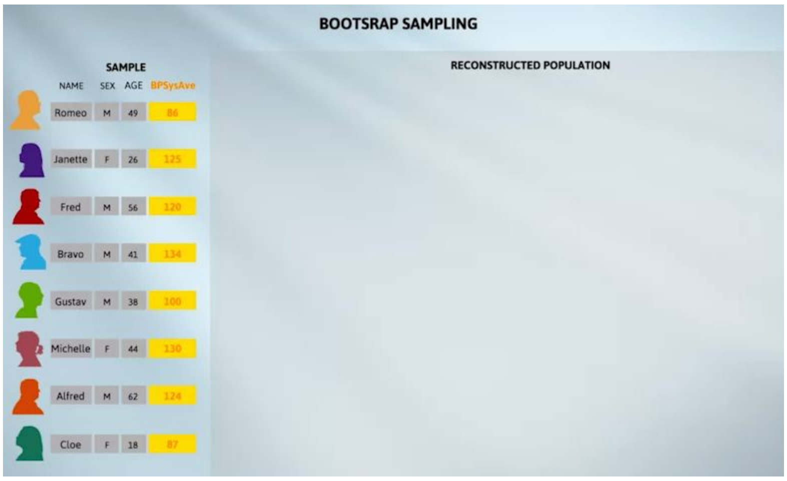
WEEK 6

CONFIDENCE INTERVALS FROM BOOTSTRAP RE-SAMPLING by Chris Wild

Hi. Our last video left us with the seemingly insoluble problem of gauging the extent of sampling error when all we have is a single sample. We held out the hope of a solution using something called bootstrap re-sampling.

So what is this bootstrap re-sampling? The name was taken from a 19th century expression, "to pull yourself up by your bootstraps". This has come to mean getting out of a difficult situation by your own efforts.

Bootstrap re-sampling in statistics began with a 1979 paper by the famous Stanford University statistician, Brad Efron. A bootstrap resample is generated by sampling from the sample (our data) with replacement. What does this mean? We'll illustrate using this scenario.



BOOTSRAP SAMPLING

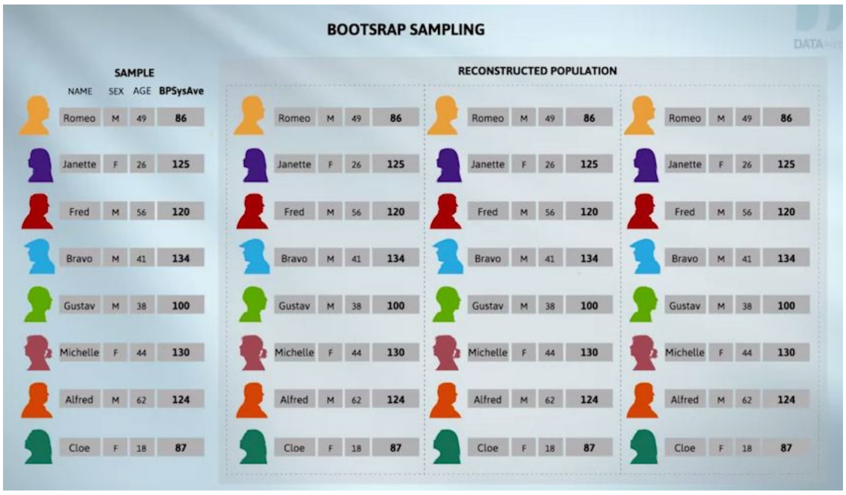
SAMPLE

NAME	SEX	AGE	BPSysAve
Romeo	M	49	86
Janette	F	26	125
Fred	M	56	120
Bravo	M	41	134
Gustav	M	38	100
Michelle	F	44	130
Alfred	M	62	124
Cloe	F	18	87

RECONSTRUCTED POPULATION

Our data corresponds to this sample from some population of interest. We've emphasised the variable, BPSysAve because our first example is going to be about systolic blood pressures. For illustrative purposes, we're using a much smaller sample than we'd ever apply re-sampling to in practise.

We want to take one bootstrap resample from these eight people. We want it to be the same size as our data sample. And if we take several of these resamples, we want them to be different because we want to imitate sampling variation.



BOOTSRAP SAMPLING

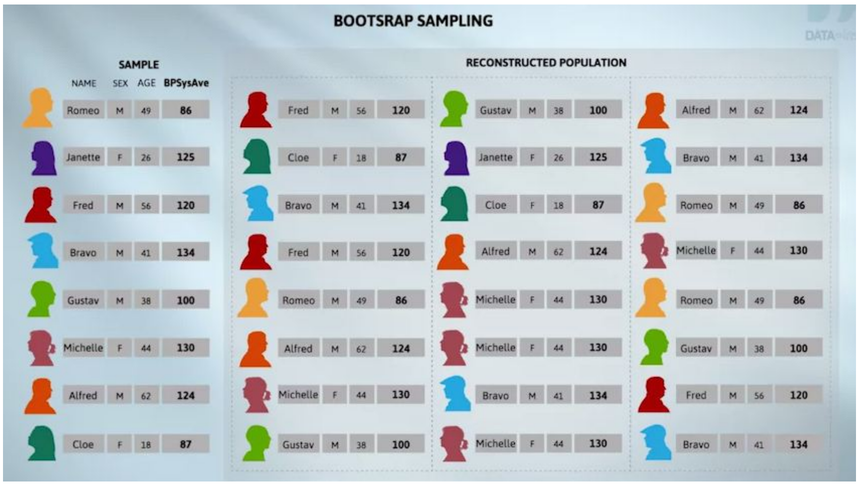
SAMPLE

NAME	SEX	AGE	BPSysAve
Romeo	M	49	86
Janette	F	26	125
Fred	M	56	120
Bravo	M	41	134
Gustav	M	38	100
Michelle	F	44	130
Alfred	M	62	124
Cloe	F	18	87

RECONSTRUCTED POPULATION

Romeo	M	49	86	Romeo	M	49	86	Romeo	M	49	86
Janette	F	26	125	Janette	F	26	125	Janette	F	26	125
Fred	M	56	120	Fred	M	56	120	Fred	M	56	120
Bravo	M	41	134	Bravo	M	41	134	Bravo	M	41	134
Gustav	M	38	100	Gustav	M	38	100	Gustav	M	38	100
Michelle	F	44	130	Michelle	F	44	130	Michelle	F	44	130
Alfred	M	62	124	Alfred	M	62	124	Alfred	M	62	124
Cloe	F	18	87	Cloe	F	18	87	Cloe	F	18	87

If they were all identical, there would be no variation. Now, if I was to sample eight people from these eight in the normal way, I'd just get everybody. I'd just get the whole sample back. Every time I did it, the result would be the same-- the whole sample. So that's no good.



BOOTSRAP SAMPLING

SAMPLE

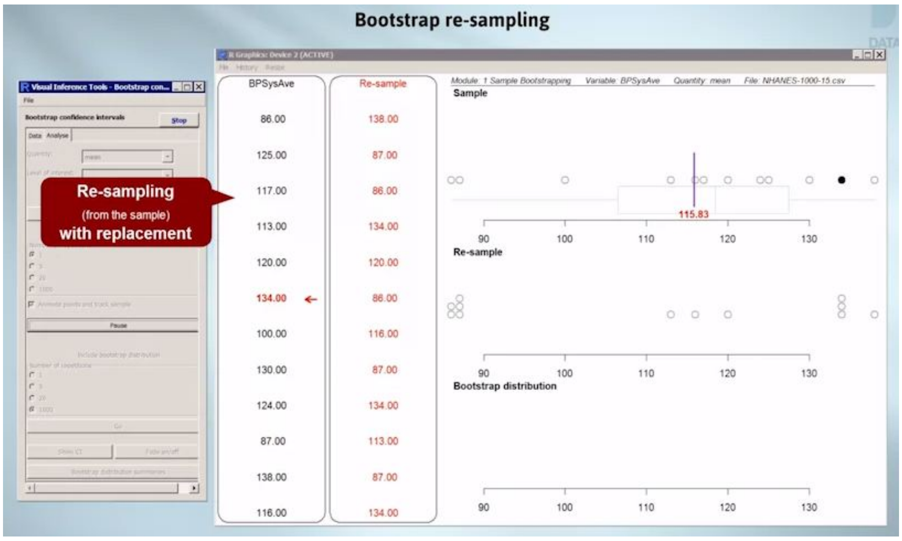
NAME	SEX	AGE	BPSysAve
Romeo	M	49	86
Janette	F	26	125
Fred	M	56	120
Bravo	M	41	134
Gustav	M	38	100
Michelle	F	44	130
Alfred	M	62	124
Cloe	F	18	87

RECONSTRUCTED POPULATION

Fred	M	56	120	Gustav	M	38	100	Alfred	M	62	124
Cloe	F	18	87	Janette	F	26	125	Bravo	M	41	134
Bravo	M	41	134	Cloe	F	18	87	Romeo	M	49	86
Fred	M	56	120	Alfred	M	62	124	Michelle	F	44	130
Romeo	M	49	86	Michelle	F	44	130	Romeo	M	49	86
Alfred	M	62	124	Michelle	F	44	130	Gustav	M	38	100
Michelle	F	44	130	Bravo	M	41	134	Fred	M	56	120
Gustav	M	38	100	Michelle	F	44	130	Bravo	M	41	134

Here, I've randomly drawn the first person to be in my resample. It's Fred. When I go back to get the next person, I randomly draw from the entire list again. That's what the "with replacement" is all about. Fred is still eligible to be chosen again, but this time, I get Chloe. And so it goes on. Oh, there's Fred again. When we finish, we see that Fred has been chosen twice, and Janet doesn't appear. I'll draw a new resample. Michelle has been chosen three times, and several people were not chosen at all. One more time. So we see these resamples are different. The process looks quite a lot like sampling and is going to give us variation.

Now, we'll start doing this with real data on systolic blood pressures. We'll be estimating the mean. This time, we'll use a slightly larger sample, a sample of size 12 from NHANES. We'll watch what happens when we do bootstrap re-sampling using animations from VIT's bootstrap module.



Bootstrap re-sampling

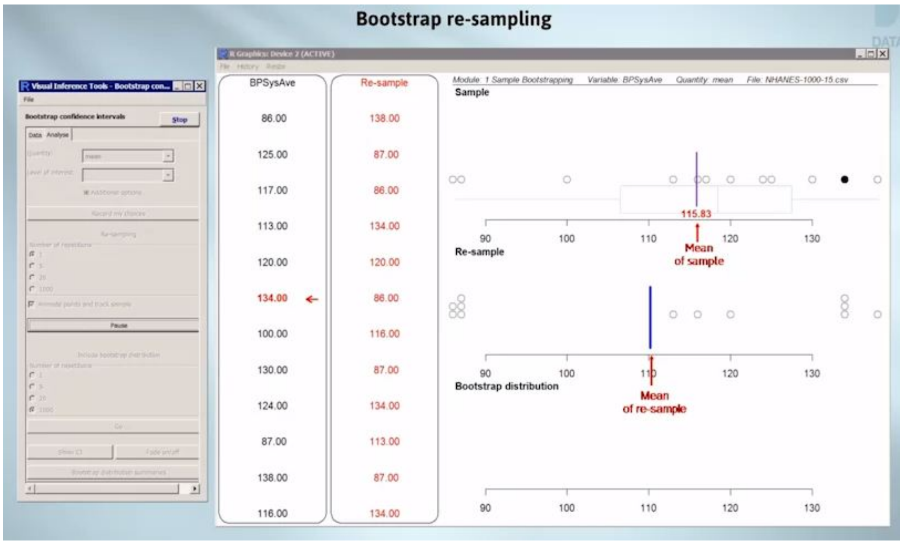
Module: 1 Sample Bootstrapping   Variable: BPSysAve   Quantity: mean   File: NHANES-1000-15.csv

BPSysAve	Re-sample
86.00	138.00
125.00	87.00
117.00	86.00
113.00	134.00
120.00	120.00
134.00 ←	86.00
100.00	116.00
130.00	87.00
124.00	134.00
87.00	113.00
138.00	87.00
116.00	134.00

Re-sampling (from the sample) with replacement

Still from animation

Here is the blood pressure data for our 12 people and a dotplot of the data. We're interested in the mean. The position of the mean for the sample is marked at 115.83. Now, we'll start sampling from these people with replacement.



Bootstrap re-sampling

Module: 1 Sample Bootstrapping   Variable: BPSysAve   Quantity: mean   File: NHANES-1000-15.csv

BPSysAve	Re-sample
86.00	138.00
125.00	87.00
117.00	86.00
113.00	134.00
120.00	120.00
134.00 ←	86.00
100.00	116.00
130.00	87.00
124.00	134.00
87.00	113.00
138.00	87.00
116.00	134.00

Mean of sample

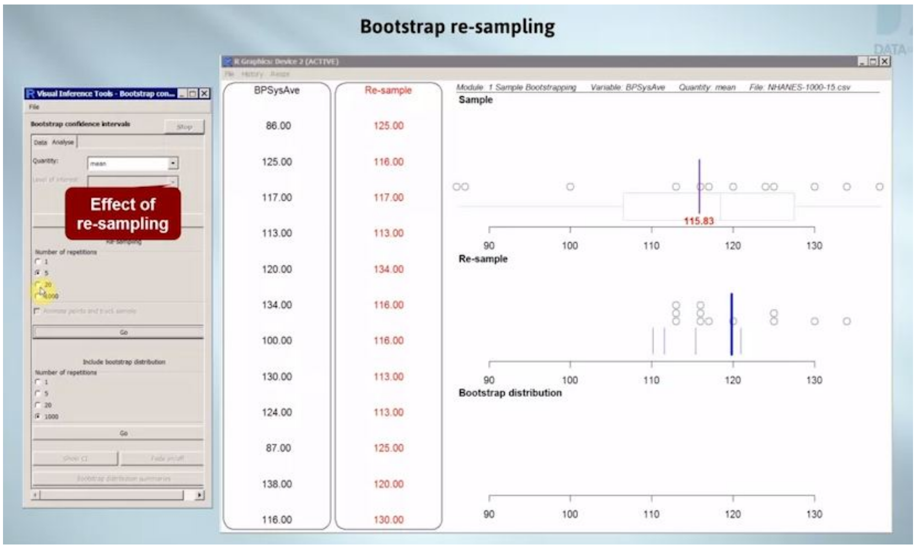
Mean of re-sample

Still from animation

So as we've said, on each draw we draw from the whole list, regardless of whether someone has been used before or not. Each time we choose someone, their value moves across to the resample column. At the same time, a copy of their point is moved down into the middle panel, which forms the dot plot of the resample. That's it finished. We've filled up the resample column, finished the dot plot, and marked the position of the mean for the resample.

The resample's dotplot is quite different from the original, and the mean is in a different place. How did the resample get made up? We'll track what happened to each of the original points. The first point was used twice, the next two not at all, the fourth one once, and so on. So this is how the resamples get to be different.

We'll do it a few more times, fast. The mean of each resample leaves a blue footprint. And now, we see a familiar pattern of variation emerging.



Bootstrap re-sampling

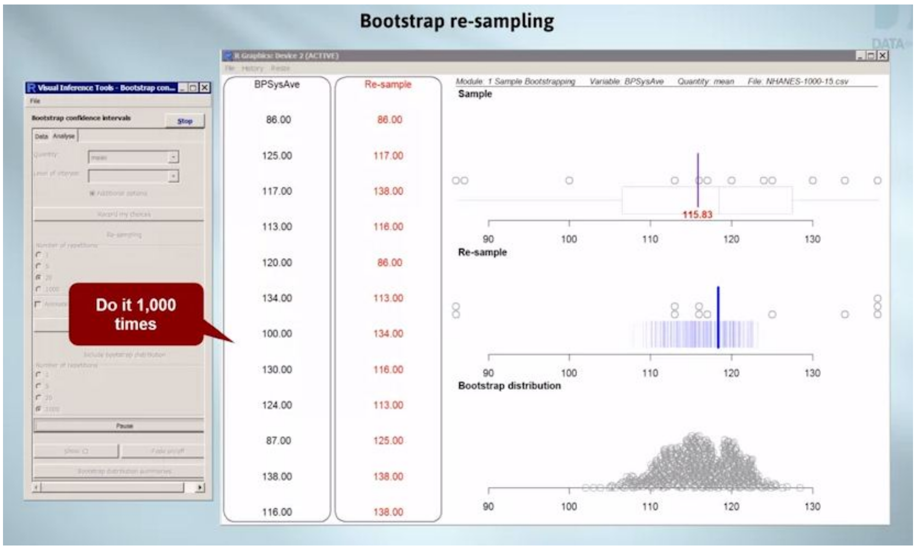
Effect of re-sampling

Module: 1 Sample Bootstrapping Variable: BPSysAve Quantity: mean File: NHANES-1000-15.csv

BPSysAve	Re-sample
86.00	125.00
125.00	116.00
117.00	117.00
113.00	113.00
120.00	134.00
134.00	116.00
100.00	116.00
130.00	113.00
124.00	113.00
87.00	125.00
138.00	120.00
116.00	130.00

Still from animation

Do some fast and see the pattern building up -- a reminder of how the bottom display builds up.



Bootstrap re-sampling

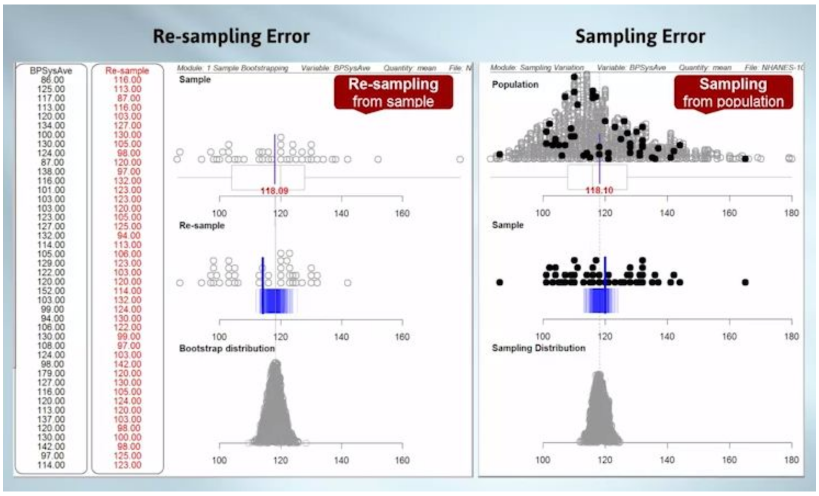
Do it 1,000 times

Module: 1 Sample Bootstrapping Variable: BPSysAve Quantity: mean File: NHANES-1000-15.csv

BPSysAve	Re-sample
86.00	86.00
125.00	117.00
117.00	138.00
113.00	116.00
120.00	86.00
134.00	113.00
100.00	134.00
130.00	116.00
124.00	113.00
87.00	125.00
138.00	138.00
116.00	138.00

Still from animation

Now, we do it 1,000 times, and watch the pattern of variation build up. The pattern generated by re-sampling from the sample does seem similar to what we saw with sampling from the population.

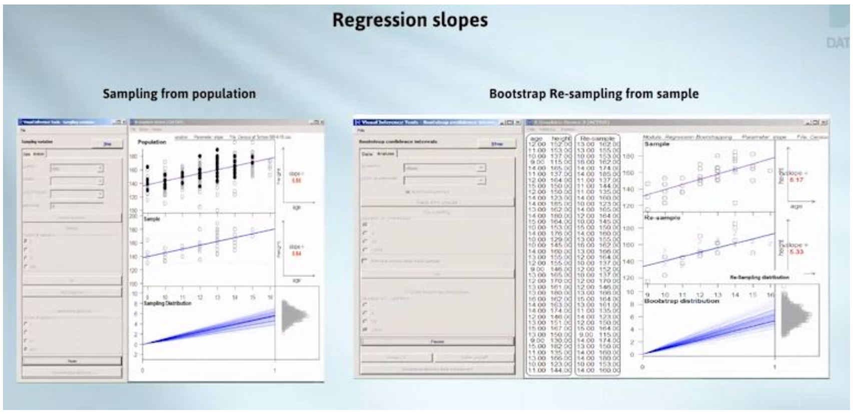


BPSysAve	Re-sample
86.00	116.00
125.00	113.00
117.00	87.00
113.00	116.00
120.00	103.00
134.00	127.00
100.00	130.00
130.00	105.00
124.00	98.00
87.00	120.00
138.00	97.00
116.00	132.00
101.00	123.00
103.00	123.00
103.00	120.00
123.00	105.00
127.00	125.00
132.00	94.00
114.00	113.00
105.00	106.00
129.00	123.00
122.00	103.00
120.00	120.00
152.00	114.00
103.00	132.00
99.00	124.00
94.00	130.00
106.00	122.00
130.00	99.00
108.00	97.00
124.00	103.00
98.00	142.00
179.00	120.00
127.00	130.00
116.00	105.00
120.00	124.00
113.00	120.00
137.00	103.00
120.00	98.00
130.00	100.00
142.00	98.00
97.00	125.00
114.00	123.00

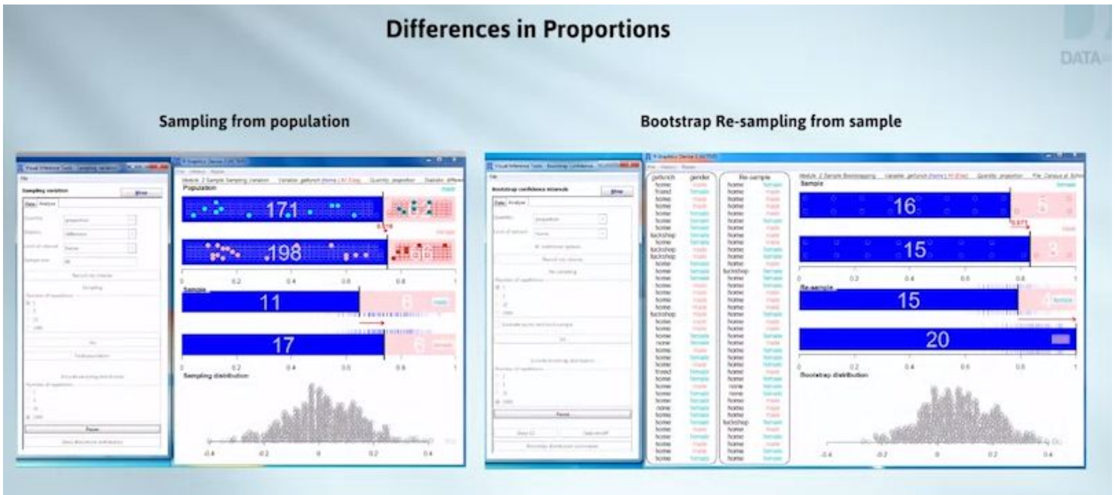
Still from animation

Here, we'll use a blood pressure data and show re-sampling error for one particular sample of 40 people from NHANES-1000 on the left-hand side. We'll compare that with sampling error for samples that are also of size 40, taken from everyone in NHANES-1000. That's on the right-hand side. Does it look similar to you? It looks pretty similar to me. This suggests that we should use bootstrap re-sampling error, which we can see, to estimate the size of sampling error, which, in practise, we can't see.

Here are several other comparisons of sampling and bootstrap re-sampling in other situations:



- applied to the slope of a trend line on a scatterplot…

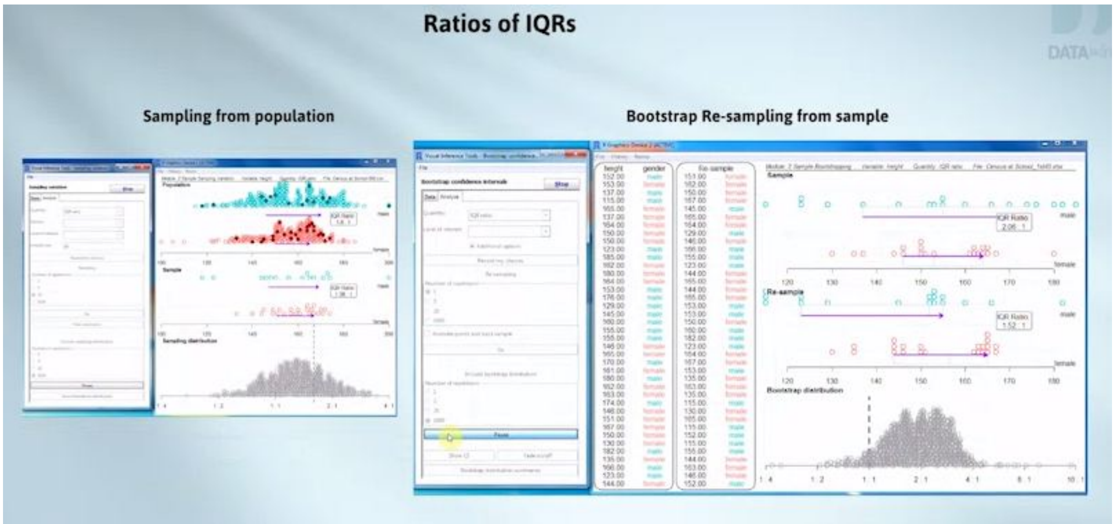


Differences in Proportions

Sampling from population

Bootstrap Re-sampling from sample

- applied to a difference in proportions or percentages,



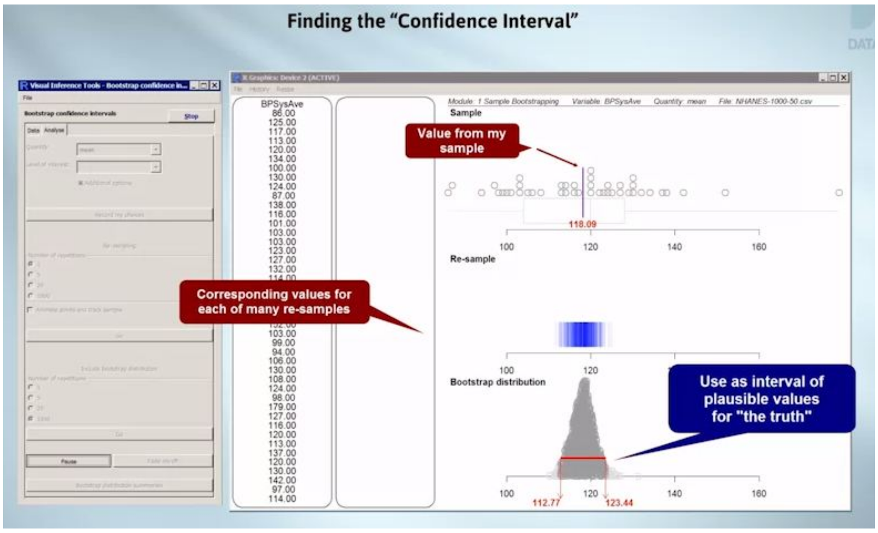
Ratios of IQRs

Sampling from population

Bootstrap Re-sampling from sample

- applied to the ratio of two interquartile ranges.

All I want you to take from these animations is re-sampling variation looks a lot like sampling variation. Our take-home message is to use bootstrap re-sampling error, which we can see, to estimate the extent of sampling error, which we can't see. In other words, we use bootstrap re-sampling to answer the question, "How wrong could I be?" But how do we use it? Here's the basic idea.



Finding the "Confidence Interval"

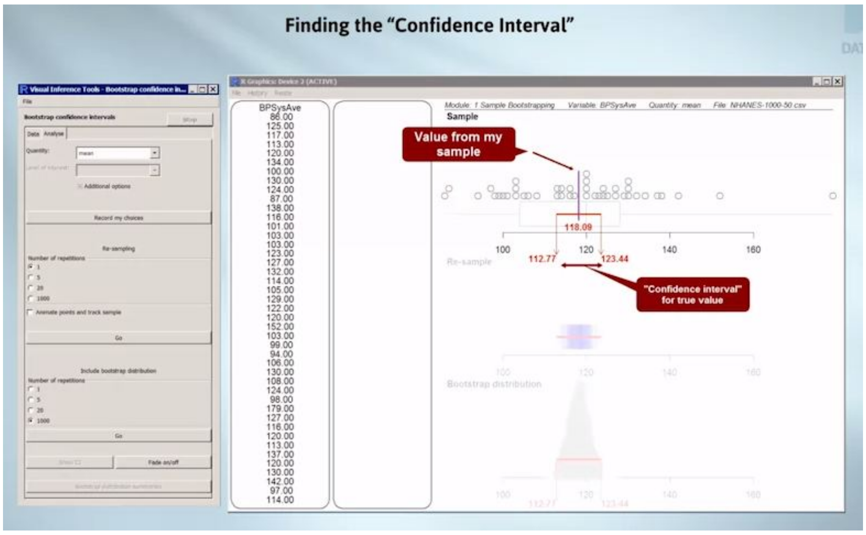
Value from my sample

Corresponding values for each of many re-samples

Use as interval of plausible values for "the truth"

Let's look at the blood pressures again. We're working with the mean, but the same basic idea applies to virtually any quantity we want to estimate (medians, percentages, slopes, and so on). In the top-right panel, we have the data with the position of our estimate marked (in this case, the mean). We then calculate the same quantity for each of a large number of resamples. This gives us a set of "How wrong can we be" values.

We trim off the few really big and really small ones and use a remaining interval as what statisticians call a confidence interval, an interval around our estimate that we're pretty sure will contain the truth. The recipe so far is entirely general.



Finding the "Confidence Interval"

Value from my sample

"Confidence interval" for true value

Now, we're moving the interval back up onto the dotplot at the top so that we can see the data, the mean of the data, and the confidence interval around the mean all in one view. Now we've pushed everything else into the visual background. This is like hiding the working in a long division sum and just showing the answer.

Our sample gave a mean blood pressure of 118.09 and we're fairly sure that the true population value lies somewhere between 112.77 and 123.44, or about 113 and 123. We'll tighten up this "fairly sure" language later.

So we have a general method, bootstrap re-sampling, for generating intervals around our estimates. We want to be able to say, we're pretty sure that our interval covers the true value. But the question now is, "Does our method work?" We'll answer that in the next video.