Bootstrapping

Goal: take repeated samples to allow for the effects of sampling variation.

Problem: Repeated sampling from a population may be impractical, expensive or not possible (sample items destroyed during sampling)

Solution: If we cannot resample from the population, (the true sampling distribution is unavailable) then we resample from the best approximation of the population we have - which is the sample itself (producing a bootstrap distribution)

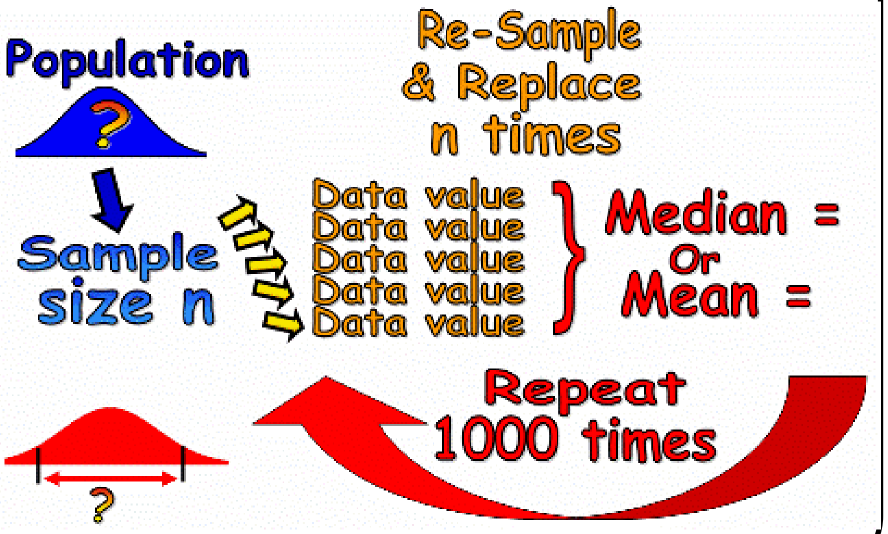
> Take repeated re-samples from the original sample. (repeat 'n' times)
> Use these re-samples to calculate an estimate for the population statistic (mean or median) This is called **Bootstrapping**

For Bootstrapping a **sample size of at least 5** values is required

To calculate the bootstrapping confidence interval many (1000) re-samples are needed.

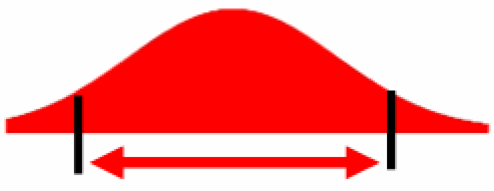
Technology is essential (iNZight software)

(class practice of bootstrapping – re-sampling by hand and collect class results on poster paper)



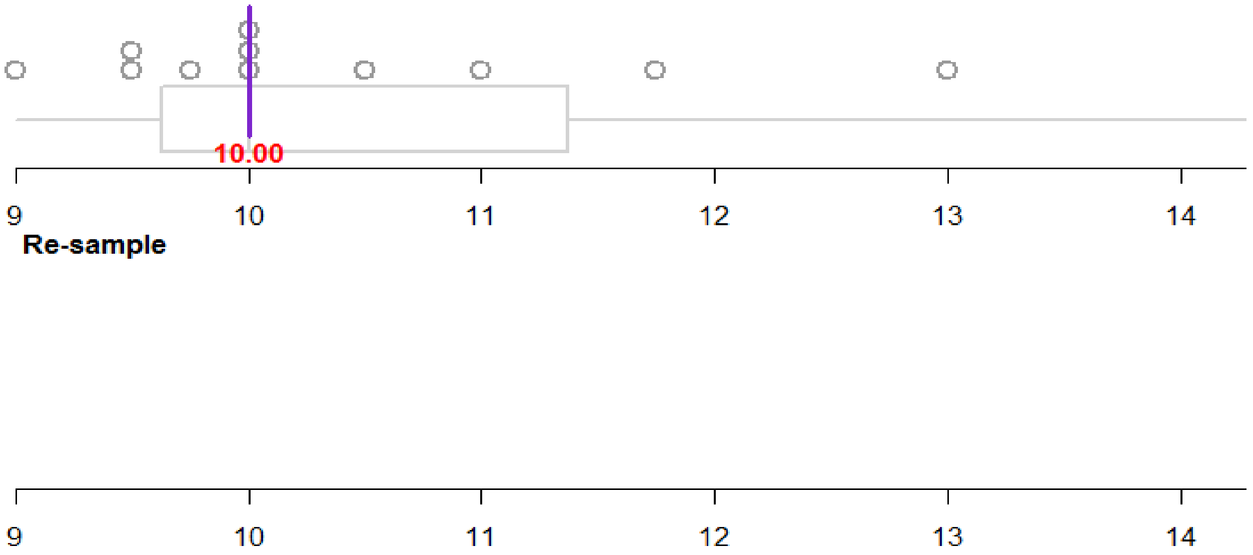
Bootstrap Details

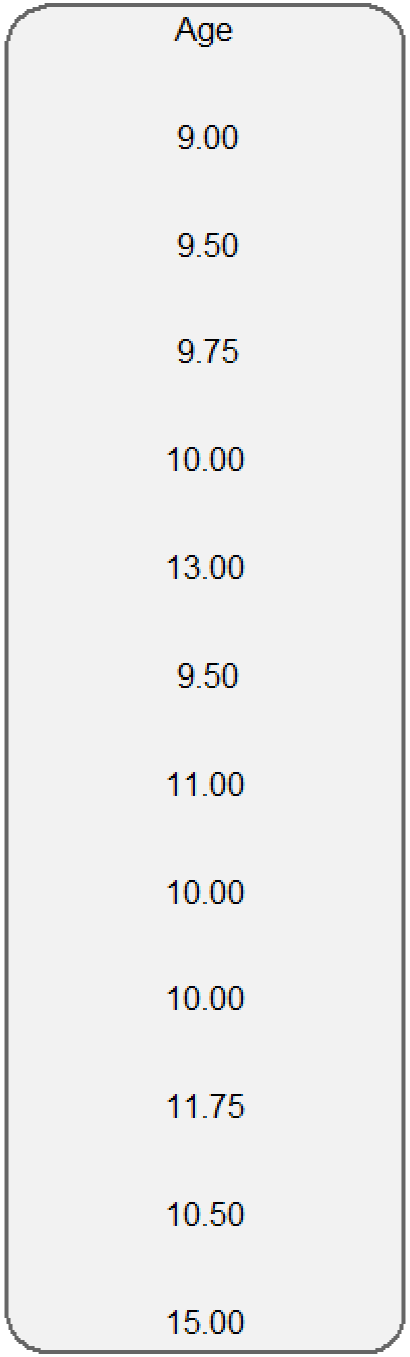
The re-sampling produces a distribution of 1000 means (or medians) which form a distribution



A 95% confidence interval provides an estimate for the population mean
How?
By counting in 2.5% (ie 25) of the 1000 values from each end

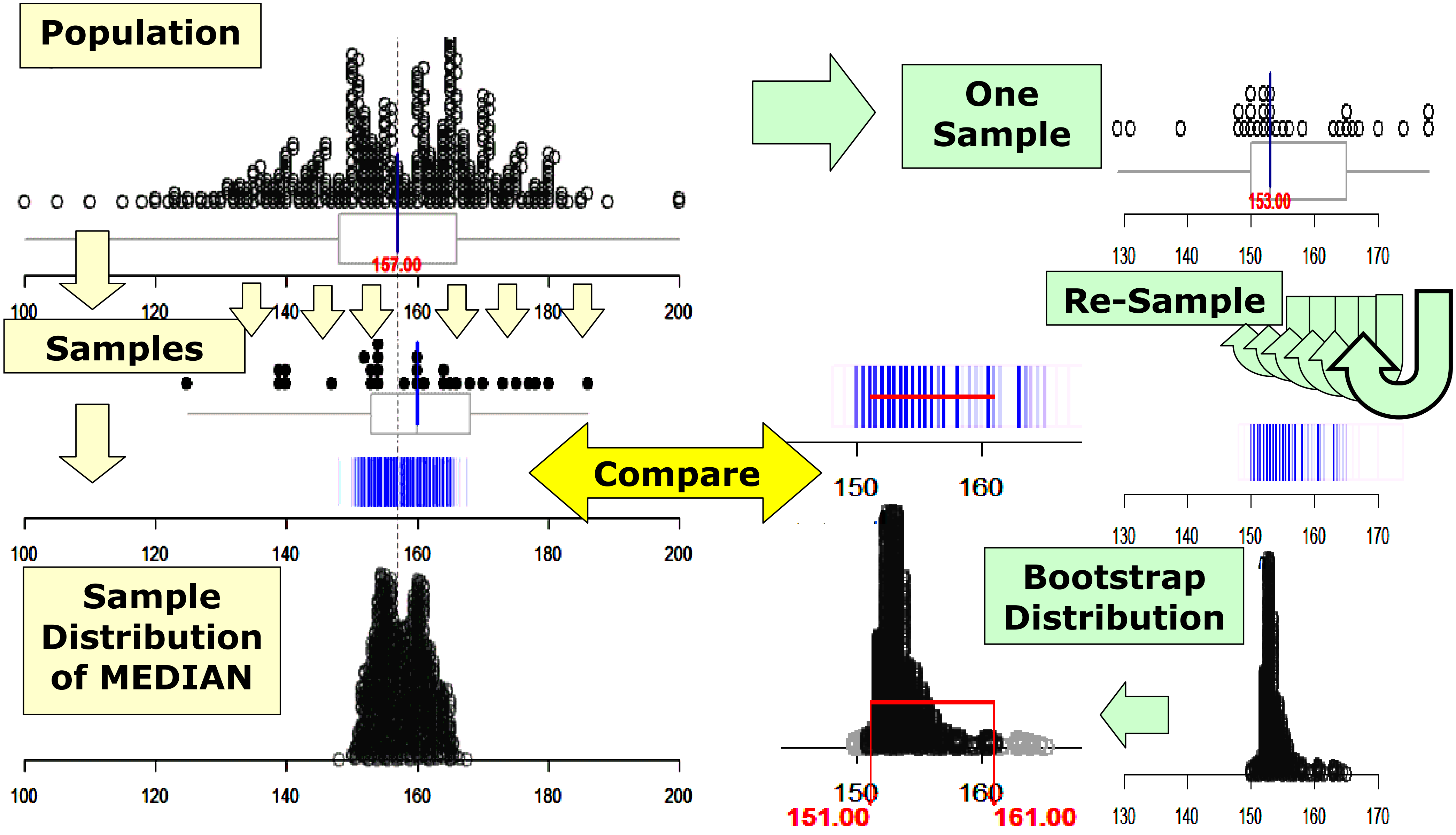
Assuming our sample was representative of the population then the bootstrapped confidence interval can be used as a estimate of the true population mean (or median





Age
9.00
9.50
9.75
10.00
13.00
9.50
11.00
10.00
10.00
11.75
10.50
15.00

Bootstrap Confidence Interval of Median



Population

One Sample

Re-Sample

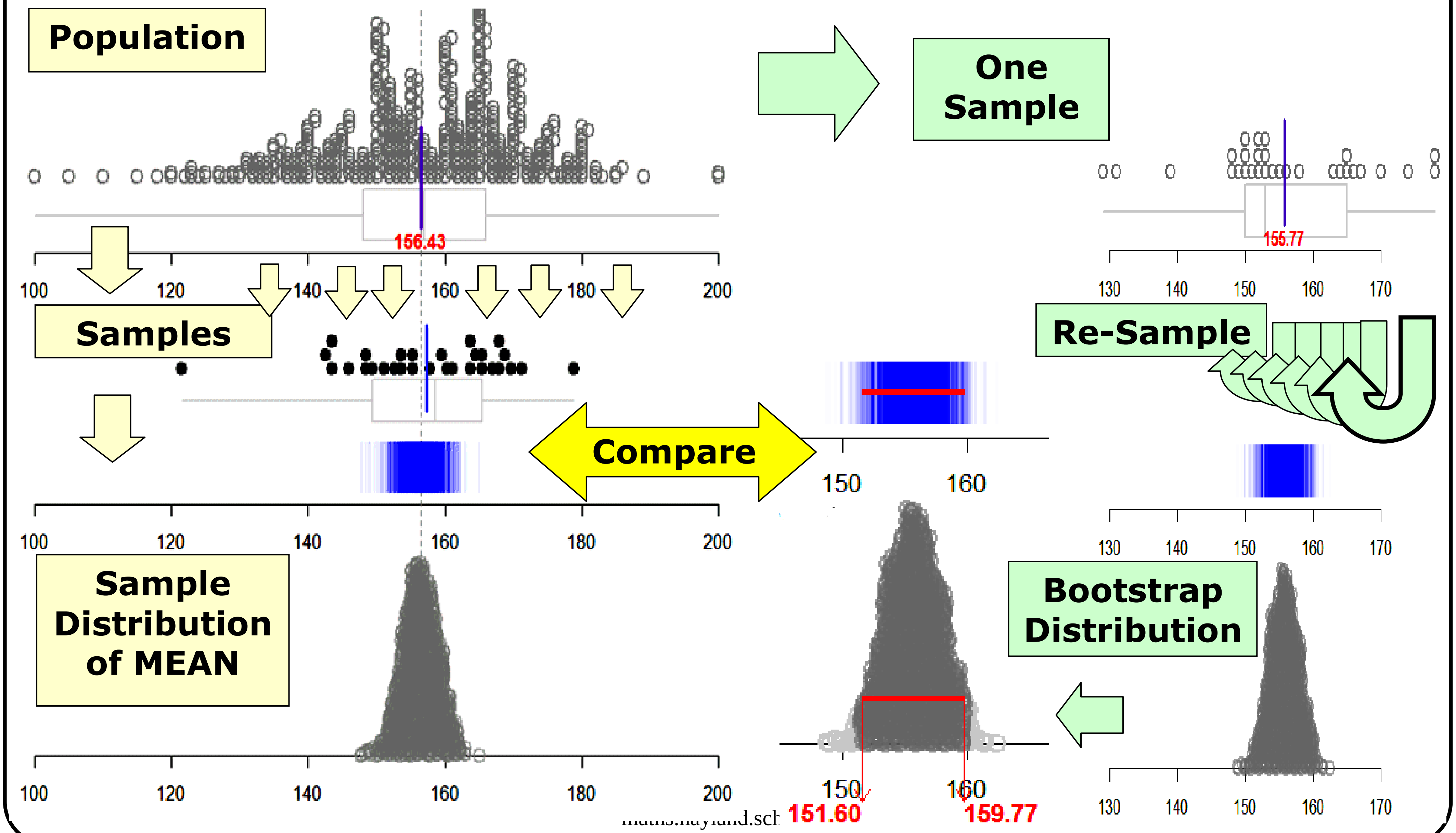
Samples

Compare

Sample Distribution of MEDIAN

Bootstrap Distribution

Bootstrap Confidence Interval of Mean



Population

One Sample

Samples

Re-Sample

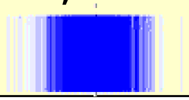
Compare

Sample Distribution of MEAN

Bootstrap Distribution

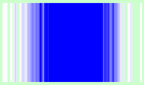
Bootstrap Confidence Interval - Overview

If we repeatedly sample from a population we get similar – but different samples.

The medians of the samples form a distribution, an interval within which we can be reasonably sure the population median lies 

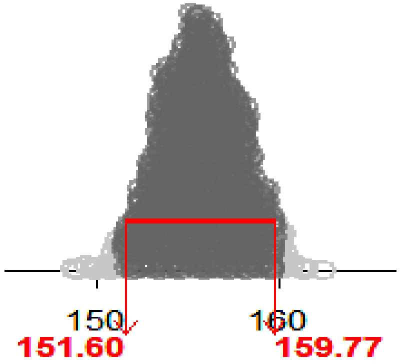
But what if we only have one sample?

If we resample (with replacement) from the sample we get a similar – but different ‘resample’ which has a median.

If we repeat the re-sampling process many times (bootstrapping) the resample medians form a distribution  which turns out to be the same as the sampling one!

By taking the middle 95% of the re-sampling distribution we make a confidence interval within which we can be reasonably sure the original population median will be within.

Assuming our sample was representative of the population then the bootstrapped confidence interval can be used as an estimate of the true population median (or mean)


150
160
151.60
159.77