TEXAS McCombs

MASTER OF SCIENCE IN BUSINESS ANALYTICS

Resampling

Outline

Quantifying uncertainty

A seemingly hacky solution: The bootstrap

Quantifying uncertainty

In data science, we equate trustworthiness with stability:

Key question: If our data had been different merely due to chance, would our answer have been different, too? Or would the answer have been stable, even with different data?

Confidence in your estimates $\iff$ Stability of those estimates under the influence of chance

Example: Quantifying uncertainty

For example:

- If doctors had taken a different sample of 503 cancer patients and gotten a drastically different estimate of the new treatment's effect, then the original estimate isn't very trustworthy.

- If, on the other hand, pretty much any sample of 503 patients would have led to the same estimates, then their answer for this particular subset of 503 is probably accurate.

Some notation

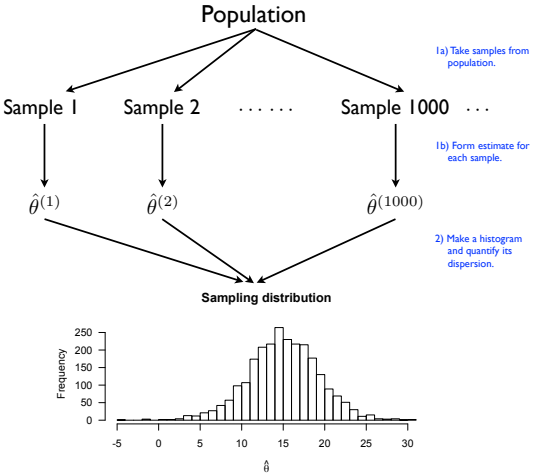
Suppose we are trying to estimate some population-level feature of interest, θ. This might be something very complicated!

So we take a sample from the population: $X_1, X_2, \ldots, X_N$. We use the data to form an estimate $\hat{\theta}_N$ of the parameter.**Key insight: $\hat{\theta}_N$ is a random variable.**

($\hat{\theta}_N$ can be the slope of a least squares regression)

$\rightarrow$ Now imagine repeating this process thousands of times! Since $\hat{\theta}_N$ is a random variable, it has a probability distribution: the sampling distribution.

Visualizing this procedure



Revisiting the standard error

Standard error: the standard deviation of an estimator's sampling distribution:

$$\begin{aligned} \text{se}(\hat{\theta}_N) &= \sqrt{\text{var}(\hat{\theta}_N)} \\ &= \sqrt{E[(\hat{\theta}_N - \bar{\theta}_N)^2]} \\ &= \text{Typical deviation of } \hat{\theta}_N \text{ from its average} \end{aligned}$$

"If I were to take repeated samples from the population and use this estimator for every sample, how much does the answer vary, on average?"

Standard error

But there's a problem here...

Knowing the standard error requires knowing what happens across many separate samples. But we've only got our one sample!

So how can we ever calculate the standard error?

Standard error

> *Two roads diverged in a yellow wood And sorry I could not travel both And be one traveler, long I stood And looked down one as far as I could To where it bent in the undergrowth...*

– Robert Frost, The Road Not Taken, 1916

Quantifying our uncertainty would seem to require knowing all the roads not taken—an impossible task.

The bootstrap

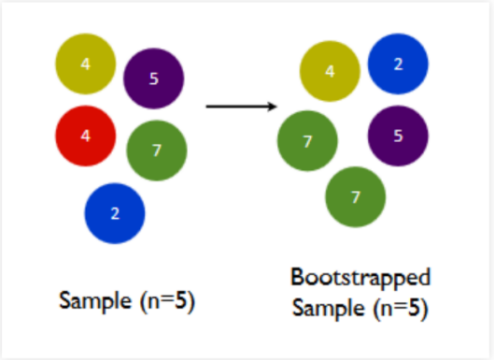
Problem: we can't take repeated samples of size N from the population, to see how our estimate changes across samples.

Seemingly hacky solution: Take repeated samples of size N, with replacement, from the sample itself, and see how our estimate changes across samples. This is something we can easily simulate on a computer (with `R`).

Basically, we pretend that our sample is the whole population and we charge ahead! This is called bootstrap resampling, or just bootstrapping.

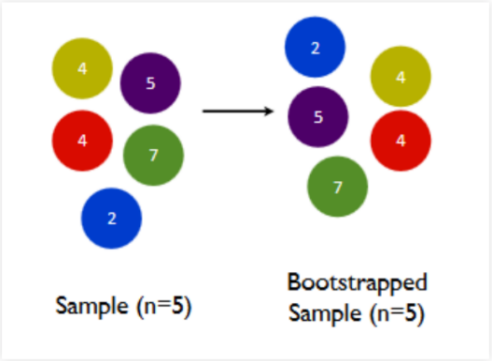
Sampling with replacement is key!

boostrapped sample 1



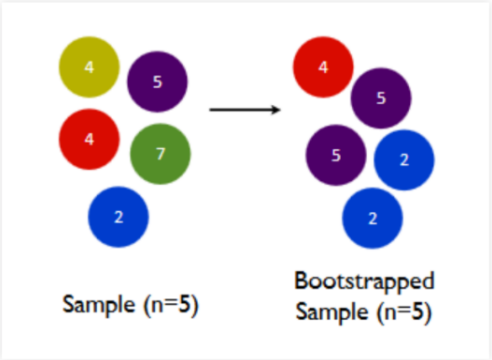
Sampling with replacement is key!

boostrapped sample 2

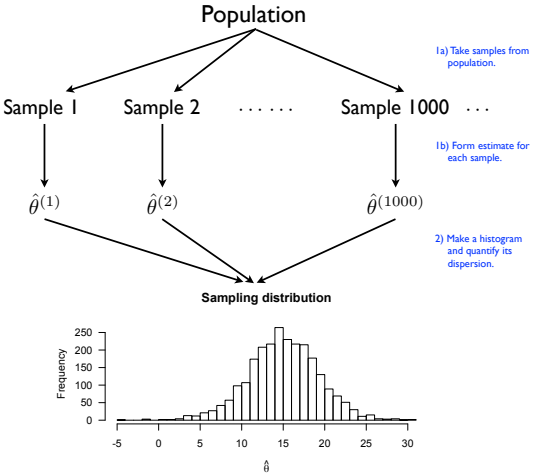


Sampling with replacement is key!

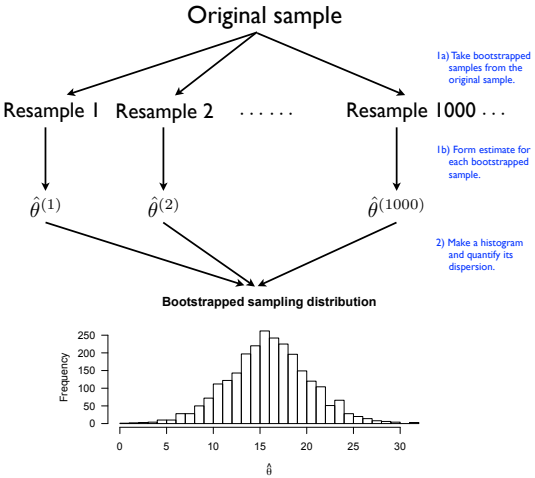
boostrapped sample 3



The true sampling distribution



The bootstrapped sampling distribution



The bootstrapped sampling distribution

- Each bootstrapped sample has its own pattern of duplicates and omissions from the original sample.

- These duplicates and omissions create variability in $\hat{\theta}$ from one bootstrapped sample to the next.

- This variability mimics the true sampling variability you'd expect to see across real repeated samples from the population.