Bootstrap Methods

Typically we use (asymptotic) theory to derive the sampling distribution of a statistic. From the sampling distribution, we can obtain the variance, construct confidence intervals, perform hypothesis tests, and more.

Challenge:

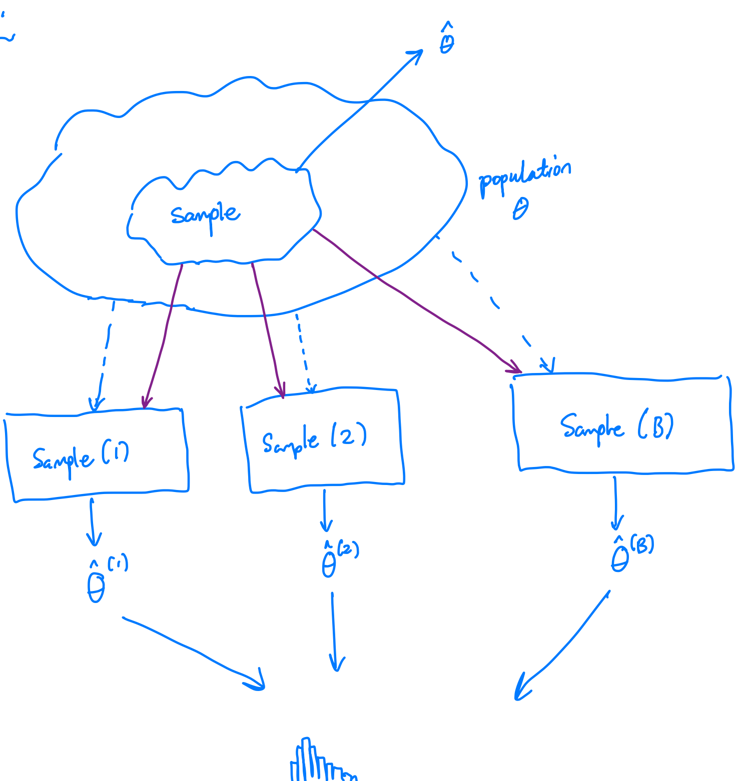
what if the sampling distribution is impossible to obtain or asymptotic theory doesn't hold?

Basic idea of bootstrapping:

Use the data to approximate the sampling distribution of the statistic.

How? Estimate the sampling distribution by creating a large # of data sets that we might have seen. and compute the statistic on each of the data sets.

E.g.



estimate of sampling dsn of $\hat{\theta}$.

Goals of bootstrap:

estimate bias, se, CI's when

1. there is doubt about if distributional assumptions are met
2. there is doubt about whether asymptotic results hold.
3. The theory to derive the dsn of a statistic is too hard.

In reality, we only have a sample and need to make sample$^{(1)}$, ..., sample$^{(B)}$.

"Bootstrap World" where the data analyst knows everything.

idea: treat the sample $Y_1, \dots, Y_n$ as the population

$\approx$ the population consists of an infinite number of Y_1 values & infinite # of Y_2 values,
each occurring $\frac{1}{n}$ proportion of the time.

$\Rightarrow$ we can resample

e.g. We are instrested in the variance of the estimator.

- In "bootstrap world" we can calculate the exact variance b/c we have access to "population"

- In practice, estimate variance by repeatedly sampling from pseudo-population.

↳ we have a represative sample of the population.

Real World

True pop. $Y_1, Y_2, \dots$ w/ dsn F_θ

True pop. parameter θ.

↓ sample

$Y_1, \dots, Y_n \Rightarrow \hat{\theta}(Y_1, \dots, Y_n)$ is estimator.

$MSE = E_F[(\hat{\theta} - \theta)^2]$

If we don't have access to F (we don't),
can't take expectation.

If we have access to population θ (we don't),
we could estimate MSE w/:

$$\widehat{MSE} = \frac{1}{rep} \sum_{i=1}^{rep} (\hat{\theta}_i - \theta)^2$$

Bootstrap World

$Y_1, \dots, Y_n$ is population

$\hat{\theta}$ is true value of parameter

Since we have access to the population,

$$\widehat{MSE}_{Boot} = \frac{1}{Bootrep} \sum_{i=1}^{Bootrep} (\hat{\theta}_i^* - \hat{\theta})^2$$

We hope $\widehat{MSE}_{Boot} \approx MSE$.