Bootstrapping

Stochastic Models

- Data is stochastic
- We model the data-generating process using some probability distribution
- Inferences from the data (ie. mean, relations between variables) are thus functions of stochastic functions
-
    - The quantities are statistics
    - Are functions of functions, so called **functionals** or **statistical functionals**
    - They have a distribution themselves
- To get confidence/standard error intervals for functionals, need to know the underlying distribution
-
    - The "sampling distribution of the estimators"
    - Usually intractable, have to rely on special cases/asymptotics

Parametric Bootstrapping

We have a set of data x and a type of model which we wish to fit to it, parameterized by a set of unknown true parameters θ_0. We have a statistic T which estimates the true functional of the data, t_0.

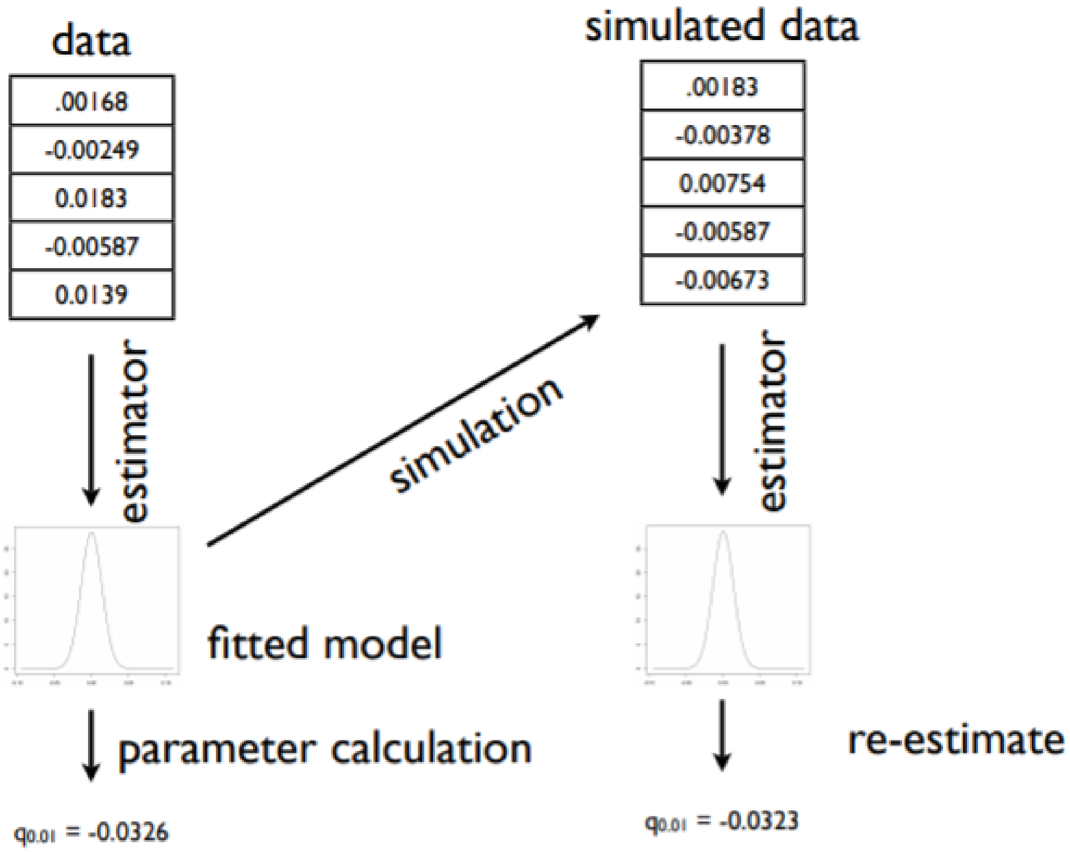


Figure 1: Schematic for model-based bootstrapping: simulated values are generated from the fitted model, then treated like the original data, yielding a new estimate of the functional of interest, here called $q_{0.01}$ (as a hint towards this week's homework).

1. Fit a model to the true data x by estimating parameters $\hat{\theta}$.
2. Estimate functional $\hat{t} = T(x)$.
3. Simulate data $\tilde{X}_i$ using model
4. Estimate functionals $\tilde{t}_i = T(\tilde{X}_i)$ on simulated data.
5. Estimate parameters $\tilde{\theta}_i$ from simulated data

Variance and Bias

Assuming that our model is correct and that our estimate $\hat{\theta}$ is close to θ_0

$Var[\hat{t}_0] \approx Var[t_0]$

In terms of bias

$E[\hat{t}] - t_0 \approx E[\tilde{t}] - \hat{t}$

The left side is our model's bias which we wish to know and the right side is our bootstrapping bias, which we have. This is valid as long as the distribution of $\hat{t} - t_0$ is close to $\tilde{t} - \hat{t}$.

Confidence Intervals

A confidence interval is a region C which contains the truth t_0 with high probability. So we say $Pr(t_0 \in C) = 1 - \alpha$ for a confidence level $1 - \alpha$. Once again, since the sampling distribution is unknown, our confidence interval is an estimate of the true confidence interval.

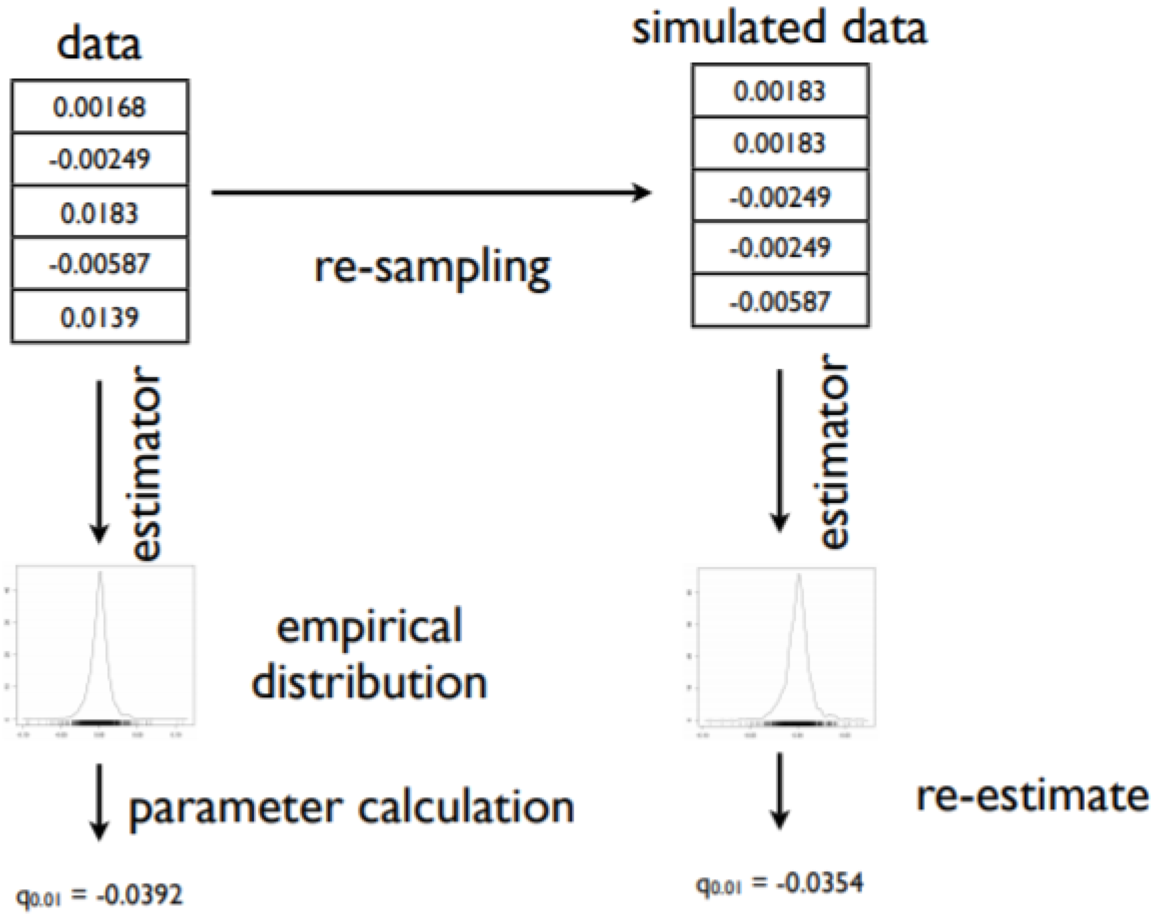
Since our functional t_0 is a real number, let C be the interval $[C_l, C_u]$. Letting this be a symmetric case, we wish equal confidence $\alpha/2$ on both sides. Looking at the left side,

$$\begin{aligned}\alpha/2 &= Pr(C_l \geq t_0) \\ &= Pr(C_l - \hat{t} \geq t_0 - \hat{t}) \\ &= Pr(\hat{t} - C_l \leq \hat{t} - t_0)\end{aligned}$$

Notice now that we have $\hat{t} - t_0$ in this inequality, which is close to our bootstrap result $\tilde{t} - \hat{t}$. This lets us approximate the true confidence interval. If we knew the underlying distribution, we could use something akin to a Z-score to find out our $\alpha/2$ quantile. Yet we don't know it and so we find such a quantile using our bootstrapped data $\tilde{t}$.

Non-parametric Bootstrapping

We have **specification error** in the case that the data generating process doesn't follow our model. The only thing we know is that our data is from the true distribution. By treating our data x as the population and sampling not from a model but from x, we can repeat everything from the parametric bootstrap.



Bootstrapping Assumptions

Bootstrapping relies on the idea that the sampling distribution is close to the true distribution. In the parametric case, it is required that small parameter changes don't lead to large changes in the resulting functionals and that the model is close to correct. In the non-parametric case, adding or removing a few data points shouldn't lead to large changes in the functional. This means that outliers occurring above their true frequency are dangerous as all observed data points are sampled uniformly

Reference

https://www.stat.cmu.edu/~cshalizi/402/lectures/08-bootstrap/lecture-08.pdf