Bootstrap Methods

Paweł Polak

April, 2016

STAT W4413: Nonparametric Statistics - Lecture 17

Bootstraping

- bootstrap for the *iid* observations $y_1, \ldots, y_n$.
- bootstrap estimates of the standard errors
- bootstrap estimates of the bias of the estimator
- the construction of confidence intervals
- bootstrap for the parameters of the linear regression model.
- bootstrap for the parameters of any regression model with heteroscedasticity.
- nonparametric hypothesis testing using bootstrap, as an alternative to permutation and Monte Carlo tests.
- parametric vs. nonparametric bootstrap
- smoothed bootstrap
- bootstrap for dependent data
- when bootstrap fails

The plug-in principle

Let $Y_1, \ldots, Y_n$ be independent random variables with common distribution function F.

The empirical distribution puts probability mass $1/n$ at each Y_i, or equivalently the empirical distribution function is given by

$$\widehat{F}(y) = n^{-1} \sum_{i=1}^{n} \mathbb{I}_{\{Y_i \leq y\}}.$$

The "plug-in estimate" of $\theta = g(F)$ is $\widehat{\theta} = g(\widehat{F})$, where g is a functional of F, e.g., the method of moments estimates the kth moment of F, for which

$$g(F) = \int y^k \mathrm{d}F(y),$$

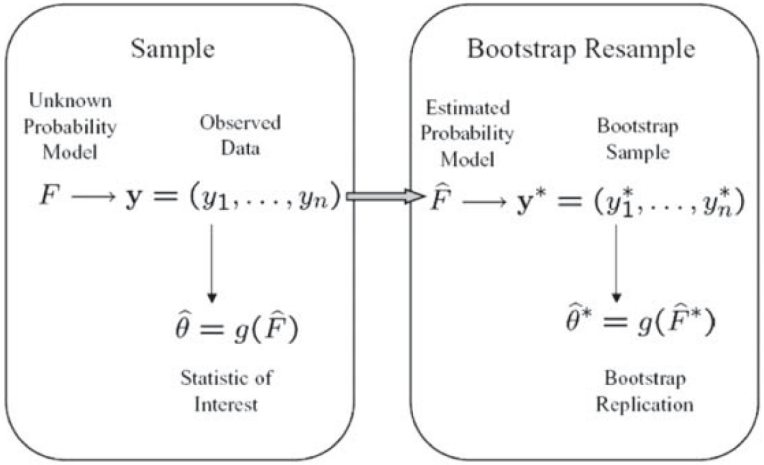
by the kth sample moment

$$\int y^k d\widehat{F}(y) = n^{-1} \sum_{i=1}^{n} y_i^k.$$

What people do before they invented bootstrap?

- To get things like standard errors or confidence intervals, we need to know the distribution of our estimates $\widehat{\theta} = g(\widehat{F})$ around the true values of our functionals $\theta = g(F)$.
- These sampling distributions follow, from the distribution of the data, since our estimates are functions of the data.
- The two classical responses of statisticians were to focus on tractable special cases, and to appeal to asymptotics.
- The bootstrap approach propsed by Efron (1979) combines estimation with simulation.

From the observed sample to the bootstrap sample



- Given the empirical distribution function $\widehat{F}$, a bootstrap sample $y^* = (y_1^*, \ldots, y_n^*)$ is obtained by sampling with replacement from $\widehat{F}$ so that the y_i^* are independent and have a common distribution function $\widehat{F}$.
- This bootstrap sample is used to form a bootstrap replicate of $\widehat{\theta}$ via $\widehat{\theta}^* = g(y^*)$.[1]
- The sampling distribution of $\widehat{\theta}$ can be estimated by simulations involving a large number of bootstrap replicates generated from $\widehat{F}$.

[1] For notational simplicity, denote $g(\widehat{F})$ by $g(y)$, and $g(\widehat{F}^*)$ by $g(y^*)$.

Bootstrap estimates of standard errors

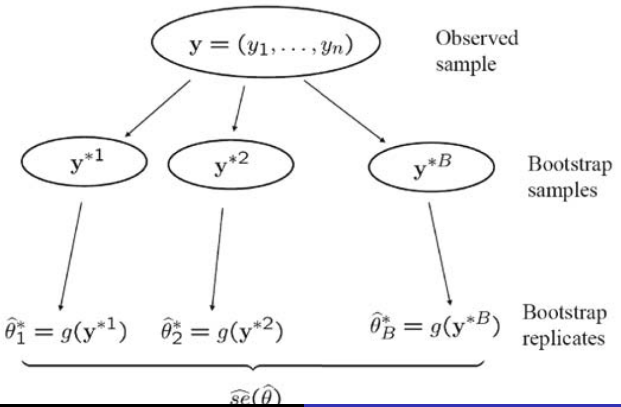
The standard error $se(\widehat{\theta})$ of the parameter estimates $\widehat{\theta}$ tells us

by how much would our estimate of the functional $\theta = g(F)$ vary, typically, from one replication of the experiment to another.

It measures this "typical variation" using the standard deviation of the sampling distribution. It can be estimated as follows:

1. Draw B independent with replacement bootstrap samples $y_1^*, \ldots, y_B^*$, each consisting of n independent observations from $\widehat{F}$.
2. Evaluate $\widehat{\theta}_b^* = g(y_b^*)$, for $b = 1, \ldots, B$.
3. Compute the average $\bar{\theta}^* = B^{-1} \sum_{b=1}^{B} \widehat{\theta}_b^*$ of the bootstrap replicates $\widehat{\theta}_1^*, \ldots, \widehat{\theta}_B^*$, and estimate the standard error

$$\widehat{se}(\widehat{\theta}) = \left\{ \frac{1}{B-1} \sum_{b=1}^{B} \left(\widehat{\theta}_b^* - \bar{\theta}^* \right)^2 \right\}^{1/2}.$$



Bootstrap the bias of the estimator

Remark

The procedure for obtaining the standard errors of the parameter estimates can be extended to other functionals of the sampling distribution of $\widehat{\theta}$. *For example, the bias of* $\widehat{\theta}$ *is* $b\left(\widehat{\theta}\right) = \mathbb{E}\left(\widehat{\theta}\right) - \theta$, *and the bootstrap estimate of the bias is* $\widehat{b}\left(\widehat{\theta}\right) = \bar{\theta}^* - \widehat{\theta}$.

Bootstrapping confidence intervals

A function of the data for which the sampling distribution does not depend on the unknown F is called a *pivot*.[2]

An approximate pivot is $(\widehat{\theta} - \theta)/se(\widehat{\theta})$, which is asymptotically standard normal (as $n \to \infty$) but may deviate substantially from normality for the finite sample size actually used.

Typically $se(\widehat{\theta})$ involves unknown parameters and needs to be estimated. Let $\widehat{se}$ be a consistent estimate of the standard error so that $(\widehat{\theta} - \theta)/\widehat{se}(\widehat{\theta})$ is also an approximate pivot.

[2] One of the simplest pivotal quantities is the z-score; given a normal distribution with mean μ and variance σ^2, and an observation x, the z-score: $z = \frac{x-\mu}{\sigma}$, has distribution $N(0, 1)$ - a normal distribution with mean 0 and variance 1. Similarly, since the n-sample mean has sampling distribution $N(\mu, \sigma^2/n)$, the z-score of the mean $z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ also has distribution $N(0, 1)$. Note that while these functions depend on the parameters - and thus one can only compute them if the parameters are known - the distribution is independent of the parameters.

Bootstrapping confidence intervals

From B bootstrap samples $y_1^*, \ldots, y_B^*$, we can compute the quantiles of

$$Z_b^* = (\widehat{\theta}_b^* - \widehat{\theta})/\widehat{se}_b^*, \quad b = 1, ..., B,$$

where $\widehat{se}_b^*$ is the estimated standard error of $\widehat{\theta}_b^*$ based on the bootstrap sample y_b^*. Let $\widehat{t}_\alpha$ and $\widehat{t}_{1-\alpha}$ be the αth and $(1-\alpha)$th quantiles of $\{Z_b^*, 1 \leq b \leq B\}$.

Then the *bootstrap-t* interval, with confidence level $1 - 2\alpha$, is

$$(\widehat{\theta} - \widehat{t}_{1-\alpha}\widehat{se}, \widehat{\theta} - \widehat{t}_\alpha \widehat{se}).$$

Bootstrapping regression models

In linear regression models, in which the observations are $(\mathbf{x}_i, y_i), 1 \leq i \leq n$, an alternative to inference based on approximate normal distribution theory for the least squares estimates is to make use of bootstrap resampling.

For the linear regression model

$$y_i = \boldsymbol{\beta}^T \mathbf{x}_i + \varepsilon_i \quad (i = 1, \ldots, n),$$

there are two bootstrap approaches:

(I) Bootstrapping Residuals,

(II) Bootstrapping Pairs.

Bootstrapping regression models: Bootstrapping Residuals

(I) Bootstrapping Residuals: the regressors $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$ are assumed to be fixed (i.e., they do not vary across samples), and the sampling variability of the least squares estimate is due to the random disturbances ε_i, which are assumed to be independent and have the same distribution F with mean 0.

- The distribution F is estimated by the empirical distribution $\widehat{F}$ of the centered residuals $\widehat{\varepsilon}_i = e_i - \bar{e}$, where $e_i = y_i - \widehat{y}_i$ and $\bar{e} = n^{-1} \sum_{i=1}^{n} e_i$.
- Note that $\bar{e} = 0$ if the regression model has an intercept term. Bootstrapping residuals involves:
  - (i) defining $y_i^* = \widehat{\boldsymbol{\beta}}^T \mathbf{x}_i + \varepsilon_i^*$, $(i = 1, \ldots, n)$ from a bootstrap sample $(\varepsilon_1^*, \ldots, \varepsilon_n^*)$ drawn with replacement from $\widehat{F}$, and
  - (ii) computing the OLS estimate $\widehat{\boldsymbol{\beta}}$ based on $(x_1, y_1^*), \ldots, (x_n, y_n^*)$. It uses the empirical distribution of B bootstrap replicates $\widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_B$ to estimate the sampling distribution of $\widehat{\boldsymbol{\beta}}$.

Bootstrapping regression models

(II) Bootstrapping Pairs: assumes that $\mathbf{x}_i$ are *iid*.

- Since the ε_i are assumed to be *iid*, the pairs $(\mathbf{x}_i, y_i)$ are also i.i.d.
- Their common distribution Ψ can be estimated by the empirical distribution $\widehat{\Psi}$.
- Bootstrapping pairs involves drawing B bootstrap samples $\left\{(\mathbf{x}^*_{i,b}, y_{i,b}) : 1 \leq i \leq n\right\}$ from $\widehat{\Psi}$ and
- computing the OLS estimate $\widehat{\beta}^*_b$ from the bth bootstrap sample, $1 \leq b \leq B$.

Re-sampling Residuals with Heteroskedasticity

In resampling the residuals we assumed that the distribution of fluctuations around the regression curve is the same for all values of the input x_i. In practice, it does not necessarily have to be the case - specifically if we look at heteroskedasticity, and estimating the conditional variance function. If we have a conditional variance function $\sigma^2(x)$, or a conditional standard deviation function $\sigma(x)$, as well as the estimated regression function $r(x)$, we can combine them to re-sample heteroskedastic residuals.

1. Construct the standardized residuals, by dividing the actual residuals by the conditional standard deviation:
$$\eta_i = \varepsilon_i / \widehat{\sigma}(x_i)$$
The η_i should now be all the same size (in distribution!), no matter at what x_i is.
2. Re-sample the η_i with replacement, to get $\widetilde{\eta}_1, \widetilde{\eta}_2, \ldots, \widetilde{\eta}_n$.
3. Set $\widetilde{x}_i = x_i$.
4. Set $\widetilde{y}_i = \widehat{r}(\widetilde{x}_i) + \widehat{\sigma}(\widetilde{x}_i)\widetilde{\eta}_i$.
5. Analyze the surrogate data $(\widetilde{x}_1; \widetilde{y}_1), \ldots, (\widetilde{x}_n; \widetilde{y}_n)$ like it was real data.

Of course, this still assumes that the only difference in distribution for the noise at different values of x is its scale.

Bootstrap testing procedures

The bootstrap tests are analogous to the permutation tests described in the previous lectures.

In the permutation tests, the test statistic is calculated for all possible samples of the data drawn without replacement from the combined data.

As we discussed the permutation tests are often approximated by Monte Carlo methods, in which case it is quite similar to the bootstrap test except, in the case of the bootstrap, the sampling is done with replacement.

Usually, the permutation tests and the bootstrap test give very similar solutions.

Three sources of approximation error

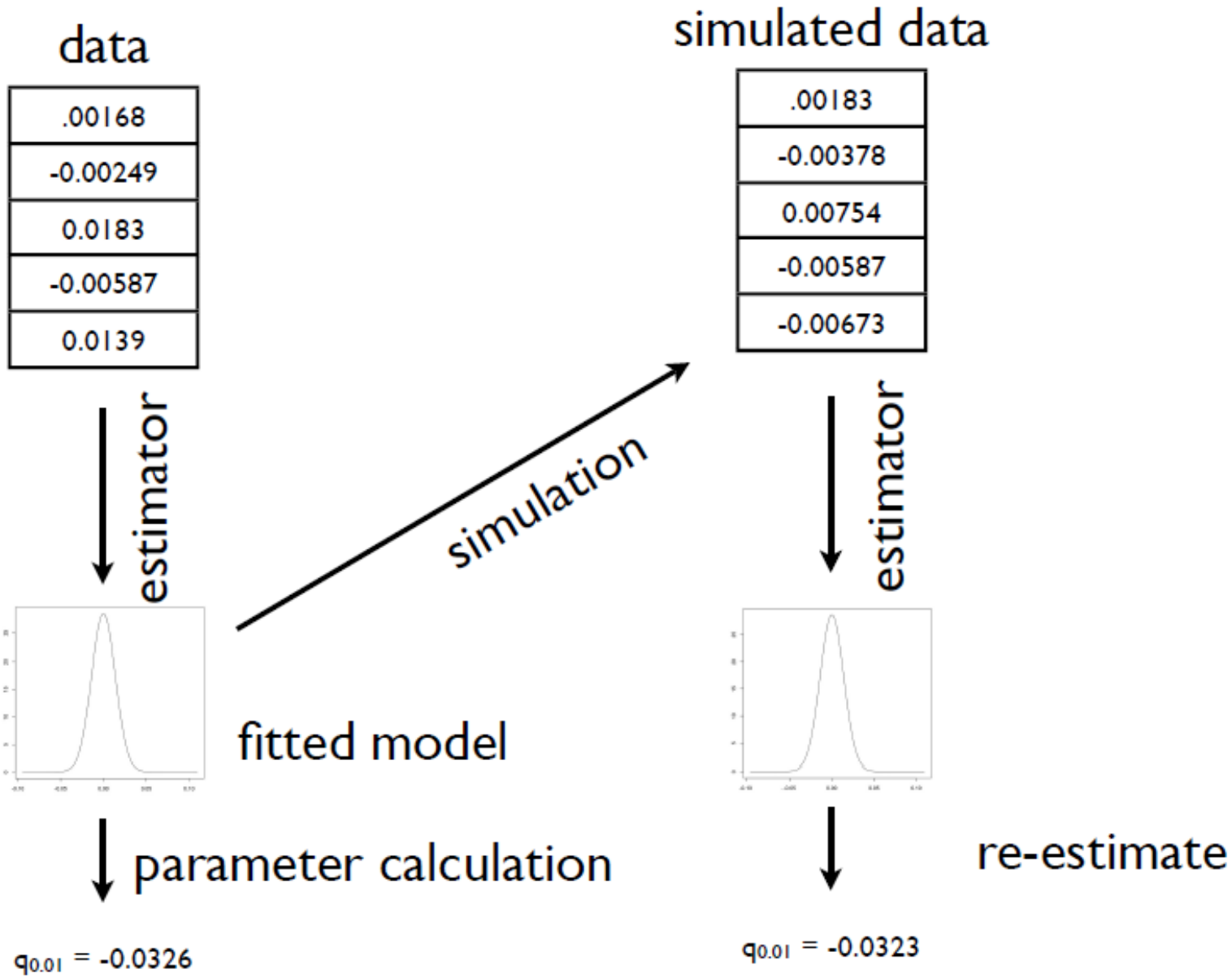
The bootstrap approximates the sampling distribution, with three sources of approximation error:

1. simulation error: using finitely many replications to stand for the full sampling distribution. Clever simulation design can shrink this, but brute force - just using enough replicates - can also make it arbitrarily small.

2. statistical error: the sampling distribution of the bootstrap re-estimates under our estimated model is not exactly the same as the sampling distribution of estimates under the true data-generating process. The sampling distribution changes with the parameters, and our initial estimate is not completely accurate. But it often turns out that distribution of estimates around the truth is more nearly invariant than the distribution of estimates themselves, so subtracting the initial estimate from the bootstrapped values helps reduce the statistical error; there are many subtler tricks to the same end.

3. specification error: the data source does not exactly follow our model at all. Simulating the model then never quite matches the actual sampling distribution.

Parametric bootstrap



Nonparametric bootstrap

Note that our intitial collection of data gives us a lot of information about the relative probabilities of different values.

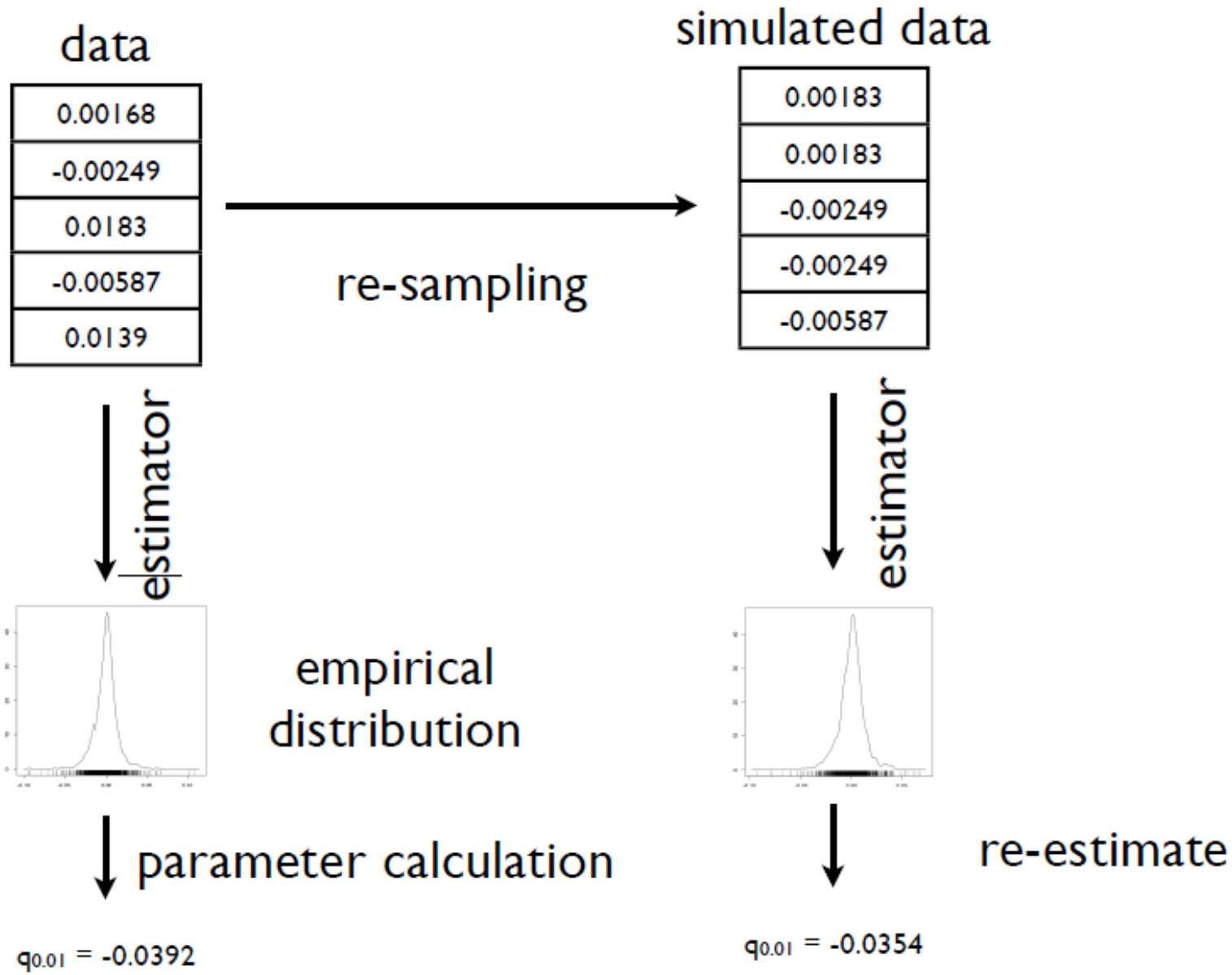
Hence, to address specification error we can replacing simulation from the model with re-sampling from the data.

In a sense the empirical distribution is the least prejudiced estimate possible of the underlying distribution - anything else imposes biases or pre-conceptions, possibly accurate but also potentially misleading. Lots of quantities can be estimated directly from the empirical distribution, without the mediation of a parametric model.

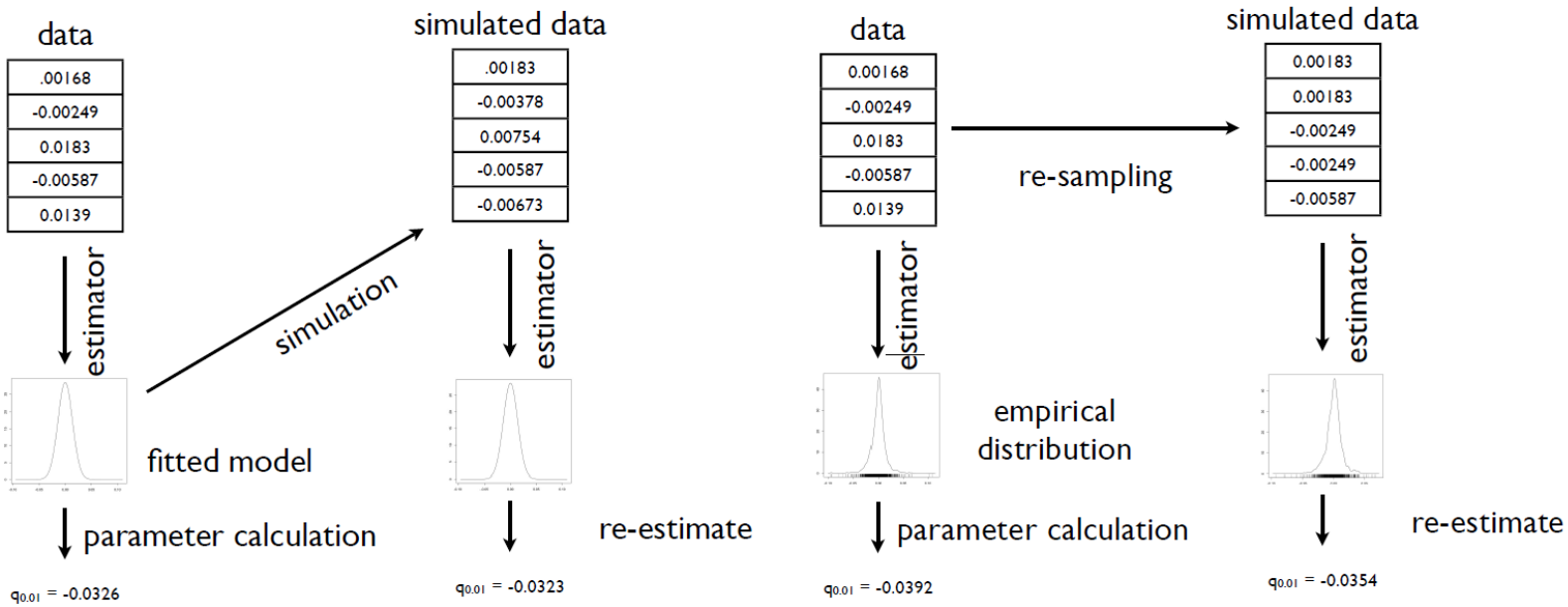
This is the idea behind the non-parametric bootstrap. It treats the original data set as a complete population and draws a new, simulated sample from it, picking each observation with equal probability (allowing repeated values) and then re-running the estimation. In fact, this is usually what people mean when they talk about the bootstrap without any modifier.

Everything we did with parametric bootstrapping can also be done with nonparametric bootstrapping - the only thing that's changing is the distribution the surrogate data is coming from.

Nonparametric bootstrap



Parametric vs. nonparametric bootstrap



When we have a properly specified model, simulating from the model gives more accurate results (at the same n) than does re-sampling the empirical distribution - parametric estimates of the distribution converge faster than the empirical distribution does. If on the other hand the parametric model is misspecified, then it is rapidly converging to the wrong distribution. This is of course just another bias-variance trade-off. If you are suspicious of your parametric modeling assumptions, choose resampling (when you can figure out how to do it, or at least until you have convinced yourself that a parametric model is very good approximation to reality).

Smoothed bootstrap

An important variant is the smoothed bootstrap, where we re-sample the data points and then perturb each by a small amount of noise, generally Gaussian. This corresponds exactly to sampling from a kernel density estimate

Bootstrap for dependent data

- If the data point we are looking are vectors (or more complicated structures) with dependence between components, but each data point is independently generated from the same distribution, then dependence isn't really an issue. We re-sample vectors, or generate vectors from our model, and proceed as usual.

- If there is dependence across data points, things are more tricky. If our model incorporates this dependence, then we can just simulate whole data sets from it. An appropriate re-sampling method from the data is trickier - just re-sampling individual data points destroys the dependence, so it won't do. One can use so called block bootstrap which samples whole blocks of data to keep the dependence.

Why Does the Bootstrap Work?

Let

$$F_n(t) = \mathbb{P}\left(\sqrt{n}\left(\widehat{\theta}_n - \theta\right) \leq t\right)$$

If we knew F_n we could easily construct a confidence interval. Let

$$C_n = \left[\widehat{\theta}_n - \frac{t_{1-\alpha/2}}{\sqrt{n}}, \quad \widehat{\theta}_n - \frac{t_{\alpha/2}}{\sqrt{n}}\right],$$

where $t_\alpha = F_n^{-1}(\alpha)$. Then

$$\begin{aligned}
\mathbb{P}(\theta \in C_n) &= \mathbb{P}\left(\widehat{\theta}_n - \frac{t_{1-\alpha/2}}{\sqrt{n}} \leq \theta \leq \widehat{\theta}_n - \frac{t_{\alpha/2}}{\sqrt{n}}\right) \\
&= \mathbb{P}\left(t_{1-\alpha/2} \leq \sqrt{n}\left(\widehat{\theta}_n - \theta\right) \leq t_{\alpha/2}\right) \\
&= F_n\left(t_{1-\alpha/2}\right) - F_n\left(t_{\alpha/2}\right) \\
&= F_n\left(F_n^{-1}\left(1-\alpha/2\right)\right) - F_n\left(F_n^{-1}\left(\alpha/2\right)\right) = 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha
\end{aligned}$$

The problem is that we do not know F_n

Why Does the Bootstrap Work?

The bootstrap estimates F_n with

$$\widehat{F}_n(t) = \mathbb{P}\left(\sqrt{n}\left(\widehat{\theta}^* - \widehat{\theta}_n\right) \leq t \mid X_1, \ldots, X_n\right).$$

If $\widehat{F}_n \approx F_n$, then the bootstrap will work.
Usually F_n will be close to some limiting distribution L. Similarly, $\widehat{F}_n$ will be close to some limiting distribution $\widehat{L}$. Moreover, L and $\widehat{L}$ will be close which implies that F_n and $\widehat{F}_n$ are close. In practice, we usually approximate $\widehat{F}_n$ by its Monte Carlo version

$$\bar{F}(t) = \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}\left(\sqrt{n}\left(\widehat{\theta}_b^* - \widehat{\theta}_b\right) \leq t\right)$$

But $\bar{F}$ is close to $\widehat{F}_n$ as long as we take B large.

Why Does the Bootstrap Work? Example

Suppose we have a random sample $X_1, \ldots, X_n \sim P$, where X_i has a mean μ and variance σ^2. Suppose we want to construct a confidence interval for μ.
Let $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and define

$$F_n(t) = \mathbb{P}(\sqrt{n}(\widehat{\mu}_n - \mu) \leq t).$$

We want to show that

$$\widehat{F}_n(t) = \mathbb{P}(\sqrt{n}(\widehat{\mu}_n^* - \widehat{\mu}_n) \leq t \mid X_1, \ldots < X_n)$$

is close to $F_n(t)$.

Theorem

(Bootstrap Theorem) Suppose that $\mu_3 = \mathbb{E}\left|X_i\right|^3 < \infty$. *Then,*

$$\sup_t \left|\widehat{F}_n(t) - F_n(t)\right| = O_p(\frac{1}{\sqrt{n}})$$

Why Does the Bootstrap Work? Proof of Bootstrap Theorem

To prove the bootstrap theorem let us recall Berry-Esseen Theorem

Theorem

Let $X_1, \ldots, X_n$ be iid with mean μ and variance σ^2. Let $\mu_3 = \mathbb{E}[|X_i - \mu|^3] < \infty$. Let $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$ be the sample mean and let Ψ be the cdf of a $N(0,1)$ random variable. Let $Z_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$. Then

$$\sup_z |\mathbb{P}(Z_n \leq z) - \Phi(z)| \leq \frac{33}{4} \frac{\mu_3}{\sigma^3 \sqrt{n}}$$

Why Does the Bootstrap Work? Proof of Bootstrap Theorem

Proof.

Let $\Phi_\sigma(t)$ denote the cdf of a Normal with mean 0 and variance σ^2. Let $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \widehat{\mu}_n)^2$. Thus, $\widehat{\sigma}^2 = \text{Var}\left(\sqrt{n}(\widehat{\mu}_n^* - \widehat{\mu}_n) \mid X_1, \ldots, X_n\right)$. Now, by the triangle inequality,

$$\sup_t \left|\widehat{F}_n(t) - F_n(t)\right| \leq \sup_t |F_n(t) - \Phi_\sigma| + \sup_t |\Phi_\sigma(t) - \Phi_{\widehat{\sigma}}(t)| + \sup_t \left|\widehat{F}_n(t) - \Phi_{\widehat{\sigma}}(t)\right| = I + II + III$$

Let $Z \sim N(0,1)$. Then, $\sigma Z \sim N(0, \sigma^2)$ and from the Berry-Esseen theorem,

$$\begin{aligned} I &= \sup_t |F_n(t) - \Phi_\sigma| = \sup_t \left|\mathbb{P}\left(\sqrt{n}(\widehat{\mu}_n - \mu) \leq t\right) - \mathbb{P}(\sigma Z \leq t)\right| \\ &= \sup_t \left|\mathbb{P}\left(\frac{\sqrt{n}(\widehat{\mu}_n - \mu)}{\sigma} \leq \frac{t}{\sigma}\right) - \mathbb{P}\left(Z \leq \frac{t}{\sigma}\right)\right| \leq \frac{33}{4}\frac{\mu_3}{\sigma^3\sqrt{n}}. \end{aligned} \quad (1)$$

Using the same argument on III term, we have that

$$III = \sup_t \left|\widehat{F}_n(t) - \Phi_{\widehat{\sigma}}(t)\right| \leq \frac{33}{4}\frac{\widehat{\mu}_3}{\widehat{\sigma}^3\sqrt{n}}$$

where $\widehat{\mu}_3 = \frac{1}{n}\sum_{i=1}^n |X_i - \widehat{\mu}_n|^3$ is the empirical third moment. □

Why Does the Bootstrap Work? Proof of Bootstrap Theorem

Proof.

By the SLLN, $\widehat{\mu}_3$ converges almost surely to μ_3 and $\widehat{\sigma}$ converges alomst surely to σ.

So, almost surely for large n $\widehat{\mu}_3 \leq 2\mu_3$ and $III \leq \frac{33}{4}\frac{4\mu_3}{\sqrt{n}}$.

From the fact that $\widehat{\sigma} - \sigma = O_p(\frac{1}{\sqrt{n}})$ it may be shown that $II = \sup_t |\Phi_\sigma(t) - \Phi_{\widehat{\sigma}}(t)| = O_p(\sqrt{1/n})$. (This may be seen by Taylor expanding $\Phi_{\widehat{\sigma}}(t)$ around σ.)
This completes the proof. □

Why Does the Bootstrap Work? Proof of Bootstrap Theorem

We have shown that

$$\sup_t \left| \widehat{F}_n(t) - F_n(t) \right| = O_p(\frac{1}{\sqrt{n}})$$

From this, it may be shown that, for each $0 < \beta < 1$ $t_\beta - z_\beta = O_p(\frac{1}{\sqrt{n}})$, and from this one can prove that

$$\mathbb{P}(\mu \in C_n) = 1 - \alpha - O_p(\frac{1}{\sqrt{n}})$$

- So far we have focused on the mean.
- Similar results can be derived for more general parameters but this is beyond the scope of this course.

When bootstrap can fail

The principle behind bootstrapping is that sampling distributions under the true process should be close to sampling distributions under good estimates of the truth.

If small perturbations to the data-generating process produce huge swings in the sampling distribution, bootstrapping will not work well, and may fail spectacularly.

For parametric bootstrapping, this means that small changes to the underlying parameters must produced small changes to the functionals of interest.

Similarly, for non-parametric bootstrapping, it means that adding or removing a few data points must change the functionals only a little.