1.6 Smooth bootstrap

Usually the true underlying distribution is continuous. In other words, the original sample is drawn from a **continuous** distribution. On the other hand, bootstrap samples are drawn from the **discrete** empirical distribution of the observed sample. In a number of applications performance of the bootstrap can be improved by **smoothing** the empirical distribution.

The basic approach of the smooth bootstrap can be described as follows:

- Original sample $\mathcal{Y}_n = \{Y_1, \ldots, Y_n\} \Rightarrow$ Estimator $\hat{\theta}$ of the parameter θ of interest.
- An i.i.d sample $Y_1^*, \ldots, Y_n^*$ is drawn at random (with replacement) from $\mathcal{Y}_n$.
- A parameter $h > 0$ is selected, and the final bootstrap sample is generated by adding smoothing (noise)

  $$\widetilde{Y}_i = Y_i^* + h\varepsilon_i,$$

  where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. $\mathcal{N}(0,1)$ random variables.
  $\rightarrow$ Bootstrap estimator $\hat{\theta}^*$

Note: For any real number y we have

$$\begin{aligned} P(\widetilde{Y} \leq y|\mathcal{Y}_n) &= \frac{1}{n}\sum_{i=1}^{n} P(Y_i^* + h\varepsilon < y|\mathcal{Y}_n) \\ &= \frac{1}{n}\sum_{i=1}^{n} \Phi(\frac{y - Y_i^*}{h}), \end{aligned}$$

where Φ and ϕ denote the distribution function and density of the standard normal distribution. This is a continuous distribution with density

$$\hat{f}(y) = \frac{1}{nh}\sum_{i=1}^{n} \phi(\frac{y - Y_i^*}{h})$$

From the theory of kernel density estimation it is known that asymptotically ($n \to \infty$, $h \to 0$, $nh \to \infty$) $\hat{f}$ converges to the true density f of the underlying distribution. The smooth bootstrap is thus consistent.

Example: Median estimation

Let μ_{med} and $\hat{\mu}_{med}$ and denote true and estimated median (normal sample of size n). We will use $\widehat{\mu}^*_{med}$ to denote bootstrap estimates from the usual non-parametric bootstrap, while $\widetilde{\mu}^*_{med}$ refers to median estimates from the smooth bootstrap.

Bootstrap aims to approximate the sampling distribution of $\hat{\mu}_{med} - \mu_{med}$ by the conditional distributions of $\widetilde{\mu}^*_{med} - \hat{\mu}_{med}$ or $\widehat{\mu}^*_{med} - \hat{\mu}_{med}$ (i.e., conditional on the bootstrapped sample). I think they mean the original data

In a simulation study (`nboot=200, nsims=1000`) we now compare the quality of these approximations for the usual and the smoothed bootstrap. We concentrate on the quality of approximating $\text{var}(\hat{\mu}_{med})$ and measure the error by

$$\sqrt{\mathbb{E}\left(\text{var}(\widehat{\mu}^*_{med}|\mathcal{Y}_n)\right) - \text{var}(\hat{\mu}_{med})\big)^2}$$

We compare the variablity of sample median and bootstrapped sample median by (the sqrt of) the squared difference between the variance of the sample estimator and the varianceof the bootstrapped median estimator averaged over the bootstrapped samples. The variablility for both estimators should be as similar as possible. Similarly we consider.

$$\sqrt{\mathbb{E}\left(\text{var}(\widetilde{\mu}^*_{med}|\mathcal{Y}_n)\right) - \text{var}(\hat{\mu}_{med})\big)^2}$$

Results: for sample of size n=11 we obtain $\sqrt{\mathbb{E}\left(\text{var}(\widehat{\mu}^*_{med}|\mathcal{Y}_n)\right) - \text{var}(\hat{\mu}_{med})\big)^2}$ =**1.4942** for the bootstrapped estimator, while for the smoothed bootstrap estimator $\sqrt{\mathbb{E}\left(var(\widetilde{\mu}^*_{med}|\mathcal{Y}_n)\right) - var(\hat{\mu}_{med})\big)^2}$we obtain depending on the degree of smoothing

`h`	`0.10`	`0.25`	`0.50`	`0.75`
n=11	`1.4012`	`1.2593`	`1.1395`	`1.3089`

Similarly for sample size n=81 we obtain

`h`	`0.10`	`0.25`	`0.50`	`0.75`
n=81	`0.5276`	`0.3938`	`0.4527`	`0.9366`

where $\sqrt{\mathbb{E}\left(\text{var}(\widehat{\mu}^*_{med}|\mathcal{Y}_n)\right) - \text{var}(\hat{\mu}_{med})\big)^2}$ =**0.6415**

Smooth bootstrap for maximum estimation

- Original sample $\mathcal{Y}_n = \{Y_1, \ldots, Y_n\} \Rightarrow$ Estimator $\hat{\theta} = Y_{(n)}$ of the maximum θ of Y.
- An i.i.d sample $Y_1^*, \ldots, Y_n^*$ is drawn at random (with replacement) from $\mathcal{Y}_n$.
- A parameter $h > 0$ is selected, and the final bootstrap sample is generated by adding smoothing (noise)

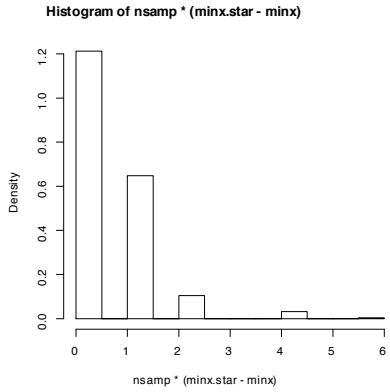
$$\widetilde{Y}_i = Y_i^* - h\varepsilon_i,$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. $\mathcal{U}(0,1)$ random variables (or alternatively $\varepsilon_i = |\zeta_i|$, $\zeta \sim \mathcal{N}(0,1)$).
$\rightarrow$ Bootstrap estimator $\hat{\theta}^*$

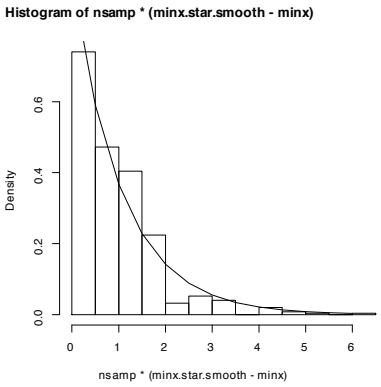
The distribution of $\max(\widetilde{Y}_j)$ is non-degenerate, exponentially shaped, unlike the distribution of $\widetilde{\theta} = Y_{(n)}^*$ which is degenerate. The smooth bootstrap is **consistent**.

Example: Sample $X_1, \ldots, X_n$, where $X_i = -|\zeta_i|$, $\zeta \sim \mathcal{N}(0,1)$. The true maximum of X is therefore $\theta = 0$.

We use a sample size `n=1000` and smooth using ε_j iid $\mathcal{U}(0,1)$ with `h=0.001`. We estimate the distributions of $n\ (\theta - \hat{\theta}^*)$ and of $n\ (\theta - \widetilde{\theta})$



unsmoothed



smoothed