Series 5

1. We will derive the probability that a given observation is part of a bootstrap sample. Suppose that we obtain a bootstrap sample from a set of n observations.

a) Argue that the probability that the jth observation is *not* in the bootstrap sample is $(1 - 1/n)^n$.

b) When $n = 100$, what is the probability that the jth observation is in the bootstrap sample?

c) Create a plot that displays, for each integer value of n from 1 to 10,000, the probability that the jth observation is in the bootstrap sample. Comment on what you observe.
Hint: Use log-scale for n, e.g. `plot(..., log = "x")` for the x-axis.

d) Given a sample of size n (n large), what proportion of the original observations do you expect to be in a bootstrap sample of size n?

(Adapted from: G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning - with Applications in R*, 2017, p. 197, Ex. 2)

2. For any random variable Z, let $q_Z(\alpha)$ denote the α-quantile of Z. Consider the Bootstrap T confidence interval $\left[\hat{\theta} - q_{\frac{\hat{\theta}^* - \hat{\theta}}{\widehat{sd}(\hat{\theta}^*)}}(1-\alpha/2) \cdot \widehat{sd}(\hat{\theta}),\ \ \hat{\theta} - q_{\frac{\hat{\theta}^* - \hat{\theta}}{\widehat{sd}(\hat{\theta}^*)}}(\alpha/2) \cdot \widehat{sd}(\hat{\theta})\right]$ and show that it asymptotically has the correct coverage when assuming bootstrap consistency.
Hint: Revise the proof for the reversed quantile bootstrap confidence interval.

3. We now construct different bootstrap confidence intervals and check their empirical coverage for different sample sizes in a simulation study. A R-skeleton including hints is available on the course website.

a) We want to estimate the trimmed mean θ of the Gamma distribution where the 10% largest and 10% smallest observations are trimmed. Approximate θ based on a very large sample.

b) Now construct a sample of size 40 from the given Gamma distribution and estimate θ using the sample trimmed mean $\hat{\theta}$.

c) Construct the four different 95%-bootstrap confidence intervals (CI) for the trimmed mean θ based on the sample from task b). The four CIs are:

- "quantile": $\left[q_{\hat{\theta}^*}(\alpha/2),\ \ q_{\hat{\theta}^*}(1-\alpha/2)\right]$,
- "normal approximation": $\hat{\theta} \pm q_Z(1-\alpha/2) \cdot \widehat{sd}(\hat{\theta})$ where $Z \sim \mathcal{N}(0,1)$, $\widehat{sd}(\hat{\theta}) = \sqrt{\frac{1}{R-1}\sum_{i=1}^{R}\left(\hat{\theta}^{*i} - \overline{\hat{\theta}^*}\right)^2}$,
- "reversed quantile": $\left[\hat{\theta} - q_{\hat{\theta}^* - \hat{\theta}}(1-\alpha/2),\ \ \hat{\theta} - q_{\hat{\theta}^* - \hat{\theta}}(\alpha/2)\right]$, and
- "bootstrap T": see exercise 2, where $\widehat{sd}(\hat{\theta})$ is as above and for each bootstrapped estimate $\hat{\theta}^{*i}$, $\widehat{sd}(\hat{\theta}^{*i}) = \sqrt{\frac{1}{M-1}\sum_{j=1}^{M}\left(\hat{\theta}^{**j} - \overline{\hat{\theta}^{**}}\right)^2}$ where $\hat{\theta}^{**j}$'s are computed from a second layer of M bootstrap samples from $Z_1^{*i}, \ldots, Z_n^{*i}$.

Hint: See R-skeleton. If you want to use the R package `boot`, `type = "perc"` corresponds to "quantile" bootstrap CI, `type = "norm"` corresponds to "normal approximation" bootstrap CI, `type = "basic"` corresponds to "reversed quantile" bootstrap CI, and `type = "stud"` corresponds to "bootstrap T" bootstrap CI.

d) To investigate the performance of the different confidence intervals, we conduct a small simulation study. Simulate 200[1] new data sets (40 observations each) and construct the different bootstrap CIs based on 500 bootstrap replicates[2] for each data set. For each type of CI, compute the percentage of CIs that do not contain θ. Specifically, if the CI is denoted by (CI_l, CI_u), compute the percentage of times that $\theta < CI_l$ and the percentage of times that $\theta > CI_u$ (non-coverage rate of the upper and lower end of the CI). Ideally, both percentages should be 2.5%.

e) Repeat task d) for sample sizes $n = 10, 40, 160, 640$ and plot the upper and lower (one-sided) non-coverage as a function of n in two separate plots. See task d) for the definition of the non-coverage. What do you observe?

Preliminary discussion: Friday, March 29.

Deadline: Friday, April 05.

[1] Start with a smaller number of data sets and / or number of bootstrap replicates to try your code and to see if your code is running correctly. It depends on the computer time you can spend whether you try 50, 100, or 200 new data sets. It may need lots of time, because each time a complete bootstrap simulation has to be carried out. You can always downsize your simulations by simulating fewer data sets and / or varying the number of bootstrap replicates.

[2] We would clearly choose way more bootstrap replicates for a real data set but this is not feasible for this simulation or computationally very expensive. Note that we only chose 50 bootstrap replicates for the double bootstrap for the "bootstrap T" CIs, see function `tm_var` in the R-skeleton.