Series 6

1. In this excercise, we revisit exercise 2 of Series 5 applying a correction using the double bootstrap. A R-skeleton including hints is available on the course website. We first rerun some lines to recreate the set-up. First, we 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. Now construct a sample of size 40 from the given Gamma distribution and estimate θ using the sample trimmed mean $\hat{\theta}$.

a) We want to construct 90%-bootstrap confidence intervals (CI) for the trimmed mean θ based on the sample we created. The CIs under consideration are:

- "reversed quantile": $\left[\hat{\theta} - q_{\hat{\theta}^*-\hat{\theta}}(1-\alpha/2),\ \ \hat{\theta} - q_{\hat{\theta}^*-\hat{\theta}}(\alpha/2)\right]$ and
- "reversed quantile with corrected level': $\left[\hat{\theta} - q_{\hat{\theta}^*-\hat{\theta}}(1-\alpha'/2),\ \ \hat{\theta} - q_{\hat{\theta}^*-\hat{\theta}}(\alpha'/2)\right]$, where α' is an adjusted level such that we have 90% coverage of $\hat{\theta}$ in confidence intervals calculated on a second layer of bootstrap.

For the double bootstrap, we use $M = 50$ samples for the outer bootstrap and $B = 500$ for every second layer bootstrap. In the end, we calculate our bootstrap confidence interval based on $B = 500$ samples with nominal level α (for the reversed quantile) and α' (for the the interval corrected by the double bootstrap.).

b) 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 as described in a) for each data set.[2] 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 5%.

c) Repeat task b) 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 b) for the definition of the non-coverage. What do you observe?

2. Show that the reversed quantile bootstrap confidence interval $\left[\hat{\theta}_n - q_{\hat{\theta}_n^*-\hat{\theta}_n}(1-\alpha/2),\ \ \hat{\theta}_n - q_{\hat{\theta}_n^*-\hat{\theta}_n}(\alpha/2)\right]$ is identical to the quantile bootstrap confidence interval $\left[q_{\hat{\theta}_n^*}(\alpha/2),\ \ q_{\hat{\theta}_n^*}(1-\alpha/2)\right]$ if the distribution of $\hat{\theta}_n^* - \hat{\theta}_n$ is symmetric around zero.

[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 more bootstrap replicates for a real data set but this is not feasible for this simulation or computationally very expensive.

3. In this exercise, we apply parametric bootstrap by hand and compare it with the output of the R package `boot`. The quantity of interest θ is the 75% percentile of the variable `boogg` which measures the waiting time till the head of the Böögg explodes on Sechseläuten[3] in Zürich between the year 2000 and 2018.

```
> # The values are rounded to minutes (from 2000 to 2018).
> boogg <- c(17, 26, 12, 6, 12, 18, 10, 12, 26, 13, 13, 11, 12, 35, 7, 21, 44, 10, 21)
```

a) Plot the data using the function `stripchart`. What are the maximum likelihood estimates of the shape and rate parameter if you fit a Gamma distribution to the data?
Hint:
- `stripchart(..., method = "stack")`
- Look at the help file of `fitdistr` of the package `MASS`, i.e.
`require(MASS); fit.gamma <- fistdistr(...)`.

b) Plot a histogram of the variable `boogg` and add the density curve of the Gamma distribution with the estimated shape and rate from a).
Hint: `hist(..., breaks = ..., prob = ); lines(x = ..., y = dgamma(...))`

c) Generate 1000 bootstrap samples using parametric bootstrap by hand (i.e., without using the package `boot`) and compute the corresponding $\hat{\theta}_n^{*1}, \dots, \hat{\theta}_n^{*1000}$.

d) Construct the following bootstrap confidence intervals for θ by hand based on the generated bootstrap samples:
- "quantile": $\left[q_{\hat{\theta}_n^*}(\alpha/2),\ \ q_{\hat{\theta}_n^*}(1-\alpha/2)\right]$,
- "normal approximation": $2\hat{\theta}_n - \overline{\hat{\theta}_n^*} \pm q_Z(1-\alpha/2) \cdot \widehat{sd}(\hat{\theta}_n)$ where $Z \sim \mathcal{N}(0,1)$,
$\widehat{sd}(\hat{\theta}_n) = \sqrt{\frac{1}{R-1}\sum_{i=1}^{R}\left(\hat{\theta}_n^{*i} - \overline{\hat{\theta}_n^*}\right)^2}$, and
- "reversed quantile": $\left[\hat{\theta}_n - q_{\hat{\theta}_n^* - \hat{\theta}_n}(1-\alpha/2),\ \ \hat{\theta}_n - q_{\hat{\theta}_n^* - \hat{\theta}_n}(\alpha/2)\right]$.

e) Conduct the same types of confidence intervals using the package `boot` and compare them to the confidence intervals computed by hand.
Hint: Use `boot(..., sim = "parametric", ran.gen = ..., mle = ...)`. See the help file of the function `boot` or the example shown in the exercise class.

f) Compare the parametric confidence intervals to the confidence intervals using non-parametric bootstrap (calculated by hand, i.e., without the package `boot`).

Preliminary discussion: Friday, April 08.

Deadline: Friday, April 29.

[3]The Sechseläuten is one of the most important cultural events in the city of Zurich. After being cancelled twice due to the pandemic, it will finally take place again this year. It is on the Monday afternoon following the easter break, and ETH even closes then. So for those who have never seen it, you should consider going.