Series 6

1. 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.

2. 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 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 $\hat{\theta}^{*1}_n, \ldots, \hat{\theta}^{*1000}_n$.

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": $\hat{\theta}_n \pm q_Z(1-\alpha/2) \cdot \widehat{sd}(\hat{\theta}_n)$ where $Z \sim \mathcal{N}(0,1)$, $\widehat{sd}(\hat{\theta}_n) = \sqrt{\frac{1}{R-1}\sum_{i=1}^{R}\left(\hat{\theta}^{*i}_n - \overline{\hat{\theta}^*_n}\right)^2}$, 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 lecture.

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

3. Recall the Panini problem from the lecture where we considered the number of duplicate cards in 50 packs. Recall that there are 682 different cards and that one pack contains five cards. The null hypothesis is that the cards are packaged at random with replacement. We suspect that this is not the case, because we tend to get a lot of duplicates, especially of some card types. We want to test the null hypothesis by using a one-sided Monte Carlo test at significance level $\alpha = 0.05$. We want to know how many packs we should open in order to ensure that the power of the test is $\geqslant 80\%$ against the following alternative: the cards are packaged at random with replacement, but $k = 100$ of the card types are 5 times more common than the others.
To answer this question: Consider `npack = 25, 30, 35, 40` and plot the resulting null and alternative distributions of the number of duplicate cards. Compute the power of the test in each case.

What is the minimum number of packs we should open to obtain the desired power? This is called a sample size calculation.

Hint: Simulate the distribution of number of duplicate cards under the null for each of the `npack = 25, 30, 35, 40` and calculate the rejection region for the one-sided Monte Carlo test. Then simulate the distribution of number of duplicate cards under the alternative for each of the `npack = 25, 30, 35, 40` and calculate the power (the probability that the observed number of duplicate cards under the alternative is in the rejection region).

Preliminary discussion: Friday, April 05.

Deadline: Friday, April 12.