SDA Group Submission Assignment 4

CD

For your group submission, upload one .pdf in Canvas ☺

If you follow all installation instructions given in Canvas and still have big issues with R Markdown installation, you can exceptionally submit a .pdf generated in a different way (e.g., by copying code 'manually' into a text editor), which should still include all relevant explanations, code snippets and graphics.

Always remember that there are several ways to code something, and that the useful `R` functions provided for each task are just some suggestions. Finally, the symbol ⚡ indicates something generally worthy of attention. Have fun!

Exercise 1

Load the data set `Davis`, available in the package `carData`,[1] remove row 12 (likely contains an input error), and inspect it using

```
library(carData)
Davis <- Davis[-12, ]
head(Davis)
```

As our variable of interest for this exercise, we will focus on *measured* (not reported) height in cm, which corresponds to column `height` in the data frame. We will split the subjects in two groups: those with `sex == M` and those with `sex == F`. Our purpose is to estimate the distribution (particularly, the standard deviation) of the *difference* between the sample mean in group `M` (denoted by $\bar{X}_M$) and the sample mean in group `F` (denoted by $\bar{X}_F$).

a. Explore the distribution of the height data in the two groups graphically: for each group, plot an histogram (scaled to density) and a normal QQ-plot side-by-side, and comment on what you see. Moreover, investigate the normality of each of the two samples by using the Shapiro-Wilk test with significance level $\alpha = 5\%$. Comment of the result of each test: is the null hypothesis rejected or not?

   Hint: note that heights have been rounded to the nearest cm, which should be taken into account when assessing the QQ-plots.

b. Set a seed (for reproducibility) and compute a bootstrap estimator for the standard deviation of the *difference* between sample means $\left(\bar{X}_M - \bar{X}_F\right)$ using both:

   - the *parametric* bootstrap (given what you have established in a.);
   - the *empirical* bootstrap.

   In both cases, remember to report the standard deviation of the empirical distribution of bootstrap values for the difference).

   Hint: note that this is a *two-sample* problem, like the one of Example 5.4 in the Syllabus. ⚡ Note that the two original samples do not have the same sample size.

c. For two independent populations distributed as $\mathcal{N}(\mu_M, \sigma_M^2)$ and $\mathcal{N}(\mu_F, \sigma_F^2)$, the sampling distribution of the difference in sample means (with sample sizes n_F and n_M, respectively) is known theoretically to

[1] Do **not** include the package installation command in your .Rmd: this would install the package every time you knit 😁 If needed, more details on package installation and loading can be found in the footnote of Submission Assignment 2 – Exercise 1.

be:

$$\left(\bar{X}_M - \bar{X}_F\right) \sim \mathcal{N}\left(\mu_M - \mu_F, \frac{\sigma_M^2}{n_M} + \frac{\sigma_F^2}{n_F}\right).$$

Assuming that the data originate from two independent normal distributions with expectation $\mu_M = 177.8$ resp. $\mu_F = 164.6$ and standard deviation $\sigma_M = 6.5$ resp. $\sigma_F = 5.5$, what is the theoretical standard deviation of the sampling distribution of the difference in sample means? Which of the estimates you found in b. is closer to this value?

d. Now, disregard what you have learned in a. and compute again a parametric bootstrap estimator, but this time based on a uniform distribution with suitably estimated parameters (you may take, for each of the two groups, the corresponding sample minimum and sample maximum as estimates for the boundaries of the supports). Report the standard deviation of the empirical distribution of bootstrap values for the difference of sample means resulting from this misspecified parametric bootstrap. How does it compare with the theoretical one you found in c.?

Deliverables

- For all subtasks, to aid in correction: all relevant code
- The graphical output, test results and comments required by a.
- The code and numerical results required by b.
- The answers to the questions in c.
- The code, numerical result and comparison for d.

Useful `R` functions

`subset`, `par(mfrow = c(1, 2))`, `hist`, `qqnorm`, `qqline`, `shapiro.test`, `length`, `mean`, `replicate`, `sqrt`, `min`, `max`, …

Exercise 2

American physicists Michelson and Morley conducted a series of experiments to determine the speed of light. Some of this measurements (in km/sec, with 299000 subtracted) are available in `R` and can be loaded and inspected with

```
data(morley)
head(morley)
l_speed <- morley$Speed
```

In this exercise, we assume that `l_speed` is a sample of realizations from i.i.d. $X_1, \dots, X_n$.

For the *composite* null hypothesis *"H_0: $X_1, \dots, X_n \overset{i.i.d.}{\sim} F \in \mathcal{F}_0$, where $\mathcal{F}_0$ is the location-scale family of normal distributions"* the standard Kolmogorov-Smirnov test cannot be used. To test this null hypothesis, an *adjusted* Kolmogorov-Smirnov statistic $\tilde{D}_n$ can be used, whose null distribution (and corresponding p-values) is unknown. This adjusted Kolmogorov-Smirnov statistic is given by

$$\tilde{D}_n = \sup_x |\hat{F}_n(x) - \Phi((x - \bar{X})/S)|.$$

Its distribution under the null hypothesis, and the corresponding p-values, can be estimated by means of the bootstrap method.

a. Explain why, under the null, $\tilde{D}_n$ is independent of the location and scale parameters of the underlying data distribution F. *Hint*: you may use the following representation:

$$\tilde{D}_n = \sup_x \left|\hat{F}_n(x) - \Phi\left(\frac{x - \bar{X}}{S}\right)\right| = \max_i \max\left(\left|\frac{i-1}{n} - \Phi\left(\frac{X_{(i)} - \bar{X}}{S}\right)\right|, \left|\frac{i}{n} - \Phi\left(\frac{X_{(i)} - \bar{X}}{S}\right)\right|\right)$$

b. For normality testing, run `ks.test` on the data using sample mean and sample standard deviation for the `par` argument. Extract the test statistic $\tilde{D}_n$ (using `$`) and comment on the p-value and corresponding test decision (use a significance level $\alpha = 5\%$). Is this p-value reliable? Why/why not?

c. Set a seed (for reproducibility) and implement a parametric bootstrap where you compute, for the i-th bootstrap sample, the value of the bootstrapped test statistic $\tilde{D}^*_{n,i}$, $i = 1, \ldots, B$ (use $B = 1000$). Finally, use these bootstrapped values of the test statistic to find a suitable p-value for the test.

   Notes:

   - Use the (default) argument `alternative = two.sided`;[2]
   - You should make sure to set *appropriate* arguments `dist` (distribution) and `par` (parameters) when using `ks.test` in the bootstrapping procedure. Perhaps the bootstrap version of the t-test statistic $t = \frac{\bar{X}-\mu_0}{S/\sqrt{n}}$, i.e. $t = \frac{\bar{X}^*-\bar{X}}{S^*/\sqrt{n}}$, might give you an idea of what an appropriate bootstrap version of the Kolmogorov-Smirnov test statistic could look like. Here $\bar{X}^*$ resp. S^* are the sample mean resp. standard deviation of a bootstrapped sample;
   - Ignore warning messages of about ties for this exercise;[3]
   - Perhaps the procedure in Example 5.5 in the Syllabus helps you to get an idea about how the bootstrap samples and values should be generated.

d. Are the two p-values found in b. and in c. (and the corresponding test decisions) the same or do they differ, and if yes how?

Deliverables

- For all subtasks, to aid in correction: all relevant code
- Your explanation for a.
- The code, result and test decision required by b.
- The code for the bootstrap procedure and the bootstrap p-value for c.
- Your answer for d.

Useful `R` functions

`ks.test`, `mean`, `sd`, `length`, `curve`, `dexp`, ...

[2] Yet, the test is right-tailed, i.e. the null hypothesis is rejected for sufficiently large values of the test statistic.

[3] You can suppress them in your R Markdown output by setting `warning = FALSE` in the chunk header.