SDA Group Submission Assignment 5

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 ϟ indicates something generally worthy of attention. Have fun!

Exercise 1

A large class takes a statistics exam and their (*fictitious*) grades are collected in file `grades.RDS` on Canvas. In particular, the `R` object saved in that file is a list of two elements (which can be extracted 'by name' using the operator `$`): vector `on_time`, containing the grades of the 250 student who started the exam on time, and vector `late`, containing the grades of the 25 who started the exam late due to traffic on their route to uni.

a. 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? Is a one-sample t-test an appropriate location test for this data and why/why not?

b. Test the null $H_0 : m \leq 7$ against the alternative $H_1 : m > 7$ for the median m of the `on_time` sample using a sign test. Is the null rejected (use a significance level of $\alpha = 5\%$)?

Recall that the test statistic has null distribution $B(n, 0.5)$ (that is, binomial with parameters n – where n is the size of the sample – and $p = 1/2$) and that, for large n, the binomial distribution can be approximated with a normal distribution with mean np and standard deviation $\sqrt{n \cdot p \cdot (1-p)}$. Compute the p-value of the preceding (right-sided) test using this normal approximation. How does it compare to the 'true' p-value and does it lead to the same test decision?

Hint: recall to check if any values in the sample are equal to $m_0 = 7$ before performing the sign test and, if necessary, adapt the testing procedure accordingly.

c. Repeat the analysis you performed in b. for the `late` sample. How does the normal approximation perform in this case and why?

Deliverables

- For all subtasks, to aid in correction: all relevant code
- The test results and comments required by a.
- The test results, computations and comments required by b.
- The test results, computations and comments required by c.

Useful `R` functions

`readRDS`, `$`, `shapiro.wilk`, `any`, `sum`, `binom.test`, `pnorm`, ...

Exercise 2

Canadian-American astronomer and mathematician Newcomb conducted a series of experiments to determine the speed of light. The file `newcomb.RDS` on Canvas gives $n = 66$ measurements to be so interpreted: these values times 10^{-3} plus 24.8 are the times, in millionths of a second, that light took to travel a known distance. *Note*: you do not need to manipulate the given values and can use them as they are for the exercise.

a. Split the data into the first 20 and the last 46 observations, and consider these as two distinct samples. Explore the distribution of the data in the two samples graphically: for each, plot an histogram (scaled to density), a boxplot and the empirical CDF of each sample side-by-side, and comment on what you see; e.g. do the samples look like they could originate from a normal distribution, is the data distribution symmetric, are there any outliers...

b. Assuming the first 20 observations originate from distribution F and the last 46 from distribution G, you want to test $H_0 : F = G$ vs. $H_1 : F \neq G$. Given what you have observed in a., which test (Wilcoxon two-sample test or Kolmogorov-Smirnov two-sample test) do you deem more suitable for this task and why?

c. Perform the test you chose in b. and comment on the results (use a significance level of $\alpha = 5\%$).

Deliverables

- For all subtasks, to aid in correction: all relevant code
- The graphical output and comment required by a.
- The decision and corresponding explanation for b.
- The test results and comment for c.

Useful `R` functions

`readRDS`, `hist`, `boxplot`, `ecdf`, `plot`, `ks.test`, ...