SDA Group Submission Assignment 6

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

You want to investigate the relationship between an email containing the word 'Urgent' in the subject line and the email being spam. You analyze the last $n = 45$ emails you have received from unknown senders and classify them as follows:

```
##            Spam Not_Spam
## Urgent        9        4
## Not_Urgent   11       21
```

a. Create the above contingency table as a matrix in R (double-check rows and columns), and visualize the data using a so-called mosaic plot. Change the colors of the plot so that the two mosaic tiles corresponding to *Spam* are in a warm color of your choice, and the two corresponding to *Not Spam* are in a cold color of your choice.

   Hint: if you like (and depending on what you have fed to the function as first argument), you can 'flip' the resulting mosaic plot using `mosaicplot(..., sort = 1:2, dir = c("h", "v"), ...)`.

b. How can you (not `R`) estimate the conditional probability $\mathbb{P}\left(\texttt{Spam}\middle|\texttt{Urgent}\right)$ from the given contingency table? Show your calculation and briefly explain.

c. Compute the following sample odds ratio using the corresponding sample (conditional) probabilities, which you can compute from the given contingency table:

$$\frac{\mathbb{P}\left(\texttt{Spam}\middle|\texttt{Urgent}\right) / \mathbb{P}\left(\texttt{Not\_Spam}\middle|\texttt{Urgent}\right)}{\mathbb{P}\left(\texttt{Spam}\middle|\texttt{Not\_Urgent}\right) / \mathbb{P}\left(\texttt{Not\_Spam}\middle|\texttt{Not\_Urgent}\right)}$$

   You can (but do not have to) use `R` to aid in calculations here, but still *compute* the odds ratio; i.e. do not extract it from the test below.

d. Use Fisher's exact test in `R` to test the *one-sided* alternative that makes most sense to you given what you have found in c. (briefly motivate your choice). Is the null rejected (use a significance level of $\alpha = 5\%$)?

   Hint: recall the formulation of the alternative stated in the output of `fisher.test`. ϟ The odds ratio in the resulting `R` output might differ a bit from the one you calculated as it does not use sample odds ratio, but the conditional MLE one. Do not worry about that.

Deliverables

- For all subtasks, to aid in correction: all relevant code
- The graphical output required by a.
- Your calculation and explanation for b.
- Your result for c.
- Your motivation, test output and test decision for d.

Useful `R` functions

`matrix`, `mosaicplot`, `fisher.test`, ...

Exercise 2

Data about passengers of the Titanic are available in `R`. From them, extract the contingency table with variable `Class` on the rows and `Survived` on the columns as follows

```
titanic_CS <- margin.table(Titanic, margin = c(1, 4))
titanic_CS
```

```
##       Survived
## Class   No Yes
##   1st  122 203
##   2nd  167 118
##   3rd  528 178
##   Crew 673 212
```

a. Visualize the data using a mosaic plot (if you wish to 'flip' it, see hint of Exercise 1 a.). Comment on the plot: is there an apparent relationship between the fate of the passengers and their economic status?

b. Perform a chi-square test for these data and comment on the results (use a significance level of $\alpha = 5\%$): is the null rejected?

c. From the test output of b., extract the expected counts under the null. For which class do you find the difference between the observed (original data) and expected counts most striking and why?

d. Perform a one-sided Fisher's exact test to test whether the odds of surviving in 3^{rd} class are larger than in the crew ($\alpha = 5\%$). To do so, first construct the appropriate 2×2 matrix from the original table.

Hint: ϟ make sure you select the appropriate alternative! To do so, pay attention to the categories corresponding to the (1, 1) cell in the 2×2 table.

Deliverables

- For all subtasks, to aid in correction: all relevant code
- The graphical output and comment required by a.
- The test result and decision for b.
- The required output and your answer for c.
- The required data extraction, test result and decision for d.

Useful `R` functions

`mosaicplot`, `chisq.test`, `fisher.test`, ...