MT5763: Software for Data Analysis
Assignment 2

Rachel Sippy and Lindesay Scott-Hayward

14 September 2022

Housekeeping

- **PLEASE** read the coursework instructions carefully *before* you start (probably more than once!).
- This **individual** coursework 2 comprises **15%** of your overall module mark.
- This is an **individual** project. The submitted coursework should reflect the work of you as an **individual**. Suspected cases of copying will be taken very seriously, so **please** adhere to the University's guidelines on good academic practice. If you have any uncertainties or questions about this, please contact me.
- I recommend you attempt every part of the assignment; even if you do not complete everything, marks are likely to be awarded for incomplete tasks / code. Remember, I cannot allocate marks to a blank sheet of paper, so help me to help you.
- **All** of the tasks / analysis should be completed in R as instructed. There should be **no** manual manipulation of files / datasets.

Submission

- You are required to upload to Moodle a **single** file - a compiled / knitted R Markdown report, either as a PDF or HTML. Name the file `MT5763_<ID>`, where `<ID>` is your University student ID e.g. `MT5763_12345.pdf` or `MT5763_12345.html`.
- The "deliverables" section will instruct you about what to include in your report. **Please** read these sections carefully.
- Deadline is **Tuesday, 25th October 2022, 23:59 (UK time)**. **PLEASE** do not leave it to the last minute to upload your work.
- The School has a lateness policy. The standard policy is an initial penalty of 15% of the maximum available mark, then a further 5% per 8-hour period, or part thereof.

Marking guidance

- Monte Carlo simulation (R)
  - Problem A - 33%
  - Problem B - 66%

You will be assessed on the following criteria:

- Successfully answering the tasks and adhering to the specifications set out in each problem.
- Providing the *deliverables* asked for.
- Clear documentation of what you have done, together with the relevant analysis and interpretation of the results.
- Code that is readable, logical, reproducible, tidy and appropriately commented.
- Appropriate use of version control.

Assignment

Version control

- Create a **private** GitHub repository called `MT5763_2_<ID>`, where `<ID>` is your University student ID. It is important that you set the repo to private to avoid any temptation of peeking at your colleague's repos.
- You **only** need to version control your `*.Rmd` file, nothing else.
- Make sure you **commit** changes often and include a succinct commit message that clearly describes what you changed and why. You will **not** be penalised for committing changes to fix mistakes in your code - this is one of the reasons why version control is used!
- *Before* submitting your coursework invite us as a collaborator to your repos so that we are able to access them. Instructions can be found here. Our usernames are `rsippy` and `lindesaysh`.
- Include a link to your GitHub repositories in your report.

Task: Monte Carlo simulation (R)

- Write efficient R code (use parallel computation techniques where applicable) to solve the following problems using Monte Carlo simulation.

Problem A

- Consider the following independent random variables:
  - $X \sim \mathcal{N}(\mu = 4, \sigma^2 = 10)$
  - $Y \sim \mathcal{U}(a = 2, b = 8)$
- Compute the probability that $X > Y$, i.e. $\Pr(X > Y)$.
- Use bootstrapping to derive the sampling distribution for your estimate of $\Pr(X > Y)$.
- Show how the *sample variance* of this sampling distribution changes as a function of the *number* of Monte Carlo simulations.

Problem B

- Consider the following football tournament format: a team keeps playing until they accrue 7 wins or 3 losses (whichever comes first - no draws allowed). Assume a fixed win rate $p \in [0, 1]$ across all rounds (they are paired at random).
- Plot how the *total* number of matches played (i.e. wins + losses) varies as a function of p.
- Comment on the *observed* win rate relative to the assumed win rate p (i.e. if a team obtains 2 wins - 3 losses, the maximum likelihood point estimate for their win rate is 40%). Specifically, focus on the effect driven by the *format* of this tournament.

Deliverables

- Include the code used to perform the simulations / calculations.
- Show, interpret and discuss the results obtained.