Short research paper

- Teams of 3 or 4 people.
- By September 12th: Confirm team members and topic.
- Due on October 10th (earlier is better): Simulation plan (to confirm the scope is fine).
- Due on October 31th: Research paper as a pdf with a maximum of 8 pages, all inclusive.
- Produced in RMarkdown, Quatro or RSweave – you must provide your R code as well as the pdf of the paper.
- A Google Sheet may be found online to help you form teams.
- This work is worth 20% of the final grade – submit your work by email (golshid.aflaki@hec.ca).

Research in statistics consists in developing new methods for analyzing or collecting data. The new method is described, and a paper will typically provide:

- Mathematical properties of the method,
- An illustration of its application to a real dataset,
- Monte Carlo simulation studies that evaluate the performance of the method for data sets of finite size (as opposed to mathematical results for samples that tend towards infinity).

Those steps are also often found in MSc theses in data science that have an academic research component.

When analyzing a dataset, it is impossible to know if we have found “the right answer.” It is therefore difficult to really know if our analysis has successfully extracted the information contained in the data. In contrast, in a Monte Carlo simulation, we know the actual data generation process (the true model). It is therefore possible to compare the result of an analysis with the actual answer. Furthermore, since we are generating multiple datasets from the same structure, we can see how close (or not) we are to the "right answer."

Step 1: Form a team, plan your simulation

Once your team is formed, you will need to pick a research question. All topics require the instructor’s approval to ensure a proper scope. Send me an email describing what you plan to do. Many of the topics may look like:

$$\text{Comparing the} \left\{ \begin{array}{c} \text{robustness} \\ \text{bias} \\ \text{mean squared error} \\ \text{time complexity} \\ \text{performance} \\ \text{stability} \\ \ldots \end{array} \right\} \text{de} \left\{ \begin{array}{c} \text{estimates} \\ \text{algorithms} \\ \text{strategies} \\ \text{models} \\ \text{methods} \\ \ldots \end{array} \right\} \text{when} \left\{ \begin{array}{c} \text{predictors are correlated} \\ \text{data contain outliers} \\ \text{some data are missing} \\ \text{changing their parameters} \\ \text{variables are error-prone} \\ \ldots \end{array} \right\}$$

You may follow the structure above or suggest a different topic. In previous years, topics ranged from evaluating the robustness of algorithms with extreme data, to simulating the greenhouse gas production of a city with different incentives in place for electric vehicles. The key with simulations is to take advantage of your ability to generate a large amount of data – repetitions of the similar datasets to see how a method varies from set to set. The approval process allows to ensure that your question is achievable, and that no two projects are on the exact same problem. You must then write a one-page simulation plan that contains:

- A clear description of your research question(s),
- How you will generate your data, and how you will measure the results,
- How it answers your research question / how you will change parameters of your simulation to do so.

The simulation plan ensures that all teams have different topics and that the scope of your project is appropriate. Prepare your plan as soon as possible and submit it. This way, you can get to the heart of the work without delay.

Step 2: Running the simulation

You should have the following structure:

1. Description of the methods compared (precise enough to be reproducible)
2. Research question (what we want to know about the methods)
3. Monte Carlo Simulation:
    - Describe precisely what you simulate and why; it must be reproducible.
    - Display and interpret the results obtained.
    - Design insightful graphs and charts to answer your question.
4. Conclusion (why was this useful and what have we learned about the method).

A good paper should therefore identify and clearly explain a relevant problem, then provide an answer to that question through a numerical Monte Carlo study.

Your final paper must be a pdf of at most 8 pages, all inclusive.

We will talk about Monte Carlo simulations in week 2 of the course. Start thinking about your topic early. Read the simulation section of some research papers to see how simulations are often run and presented. Some examples are given as references on ZoneCours.

Example of potential titles:

Here are a few examples of topics that could be considered. In 8 pages, it may not be possible to answer those questions fully, but it is possible to give some insight from simulating a few scenarios. This is the purpose of the assignment – designing a simulation to answer a research question.

- Can the stepwise method recover the true model that was simulated?
- How does the model from a stepwise method compare in performance with respect to the true model?
- How bad is it to ignore censored data?
- Is choosing the best AUC a good idea when you know the costs of errors in selection?
- Are random forests systematically better than regression trees?
- How bad can informative missing values be?
- How do measurement errors affect the AUC of a classification problem
- Do measurement errors affect the RMSE of a prediction problem?
- What happens when the target variable of a prediction model has measurement error?
- What models are most affected by the addition of noise variables?
- How does (a specific model) react to outliers?
- How stable is hierarchical clustering when you change the distance?
- Is multicollinearity bad for predictions?