# F20SA/F21SA 2023-24: Assessed Project

In this project we analyse the distribution of file sizes sent through an internet network. We will model those sizes by using a Pareto statistical model, which we will fit to the data using both maximum likelihood estimation and the method of moments. Note that where appropriate, you are expected to show full details of mathematical derivations, either writing in Word/Latex, or by writing out clearly and legibly by hand and then scanning and embedding the image in your report.

## Project description

You have been asked to analyse file sizes sent through an internet network. You have access to the values $\underline{x} = (x_1, x_2, \dots, x_{1500})$ of sizes in kB of 1500 randomly selected files (of size at least 2000kB) sent through a network in the recent past. The data is provided in the `filesize.csv` file.

You are advised to assume a Pareto model for the file sizes, with probability density function given by

$$f(x, \alpha, x_m) = \begin{cases} \dfrac{\alpha x_m^\alpha}{x^{\alpha+1}} & \text{for } x \geq x_{m,} \\ 0 & \text{otherwise,} \end{cases}$$

where $x_m = 2000$ is the minimum possible value and $\alpha > 1$ is an unknown parameter.

1. Summarise the data by plotting a histogram and calculating numerical summaries, such as the sample mean, standard deviation, median, and other quantiles. Briefly comment on your results.

[3 marks]

2. Derive the formula for the maximum likelihood estimator for $\alpha$ (denoted henceforth by $\hat{\alpha}$).

[4 marks]

3. Derive the Fisher information for $\alpha$ and use it to approximate the distribution of $\hat{\alpha}$.

[3 marks]

4. Given that the expected value of a Pareto random variable is $E(X) = \frac{\alpha x_m}{\alpha - 1}$, derive the method of moments estimator for $\alpha$ (denoted henceforth by $\tilde{\alpha}$).

[3 marks]

5. Using the results of parts 2 and 3, and the given data, calculate the numerical value of $\hat{\alpha}(\underline{x})$ and report an approximate 90% equal-tailed confidence interval $I = [\alpha_L(\underline{x}), \alpha_U(\underline{x})]$ for $\alpha$.

[4 marks]

6. Using the result of part 4, and the given data, calculate the numerical value of $\tilde{\alpha}(\underline{x})$, compare this value with the value of $\hat{\alpha}(\underline{x})$, and comment on the values of the two estimators.

[3 marks]

7. For any $i = 1 \dots 1500$, let $X_i' \sim \mathrm{Pareto}\left(\hat{\alpha}(\underline{x}), x_m\right)$ denote the predicted file sizes coming through a network next month. Let $Y' = \frac{1}{1500}\sum_{i=1}^{1500} X_i'$ be the predicted mean file size (assume that predicted individual sizes are independent of each other). Use simulation in R to estimate the distribution of $Y'$. Report this distribution by plotting a histogram and calculating numerical summaries. (Hint: Use the `Pareto` R package).

[5 marks]

Your findings should be presented in the form of a report, which should:

- have a clear and logical structure;
- include an introduction and clearly stated conclusions that can be understood by any numerate scientist (not necessarily a statistician);
- include details of your mathematical calculations;
- include clearly labelled and correctly referenced tables and diagrams, as appropriate;
- include the R code you used in an appendix (you do not need to explain individual R commands but some comments should be included to indicate the purpose of each section of code);
- include citation and referencing for any material (books, papers, websites etc) used.
- **maximum page limit of four (4) pages (11-point font, A4 size, 4 pages = 2 sheets of paper, additional pages are allowed for the R code). Excluding R code, only the first 4 pages of your submission will be marked. No feedback will be given, or marks awarded, for any work (apart from the R code) appearing on subsequent pages**.

A total of **5 Marks** is available for these aspects of your report. This will be marked according to the rubric given in the Appendix.

[Total: 30 Marks]

## Notes

- This assignment counts for 30% of the course assessment.
- You may discuss this project with your colleagues, but your report must be your own work.
- **Plagiarism** is a serious academic offence and carries a range of penalties, some very serious. Copying a friend's report or code or copying text into your report from another source (such as a book or website) without citing and referencing that source, is plagiarism. **Collusion** is also a serious academic offence. You must not share a copy of your report (as a hard copy or in electronic form) or your computer code with anyone else. Penalties for plagiarism or collusion can include voiding of your mark for the course.
- **Your report should be submitted through Canvas by 15:30 GMT (for Edinburgh students) or by 23:59 local time (for Dubai students) on Friday the 24th of November 2023. A link to the submission page is available through the 'Assignments' section of the course Canvas page. Please use the submission link appropriate for the campus where you are studying (Edinburgh or Dubai). Please submit one single pdf file.**
- For late submissions, 30% will be deducted for work submitted up to 5 working days late. Submissions that are 6 days late or more will receive 0 marks.

## Appendix

The five marks available for the exposition of your report will be awarded according to the scale below:

| | |
|---|---|
| 0-1 Marks will be awarded for | • Lack of clear and logical structure<br>• Conclusions missing or not suitable for a non-statistician<br>• Statistical calculations and methodology not clearly set out for the reader<br>• Tables and figures unclear, badly labelled or not correctly referred to<br>• R code not included, or no comments included in it<br>• Sources used not clearly referenced |
| 2-3 Marks will be awarded for | • Clear and logical structure<br>• Conclusions generally suitable for a non-statistician<br>• Statistical calculations and methodology generally set out clearly for the reader<br>• Tables and figures often clear and correctly referred to<br>• R code included with some comments<br>• Sources used clearly referenced |
| 4-5 Marks will be awarded for | • Clear and logical structure<br>• Conclusions suitable for a non-statistician<br>• Statistical calculations and methodology set out clearly for the reader<br>• Tables and figures clear, correctly referred to and easy to interpret<br>• R code included with comments<br>• Sources used clearly and correctly referenced |