# MATH1712 Probability and Statistics II
# Practical 2

`http://www1.maths.leeds.ac.uk/~arief/MATH1712/`
Arief Gusnanto, MATH1712@leeds.ac.uk
2023/24, semester 2

In this practical we consider a data set about UK fishing vessels from the module external webpage

`https://www1.maths.leeds.ac.uk/~arief/MATH1712`

under the heading *Data*. The original source of the dataset is from the UK government website at

`https://www.gov.uk/government/statistical-data-sets/vessel-lists-over-10-metres`

The UK government web page provides several versions of the data set and for the practical we will use the data from January 2020. The practical counts 10% towards the final mark of the module.

The deadline for handing in your report is **Wednesday, 20th March 2024, 12noon**. No submission is possible after 5pm on the day. **There are some important instructions on how to write and submit your report in Page 3 of this document**, including submission to both *Gradescope* and *Turnitin* via Minerva.

You can get help with the use of R in the timetabled practical sessions. These sessions take place in the week beginning 11th March. You can also refer to the 'Introduction to R' on the module web page for more explanations about the required R commands.

**Task 1.** Import the data into R and give an overview over the data, using summary statistics as appropriate. If you access the data directly from the UK government webpage, it may be easiest to first load the data into Microsoft Excel and then to save the relevant section of the spreadsheet as a csv file. Make sure to read the explanations in the Excel file.

As hinted at in the original spreadsheet, rows are duplicated, with the duplicates differing only in the column 'Licence Category'. Since we are not interested in licence categories, remove the column 'Licence Category' from the data, and then use the command `unique()` to remove the excess rows. (You can type '`help(unique)`' to learn how this command works.) Your report should at least address the following issues:

- Give some general information about the data set.
- Convince the reader that you have imported the data correctly.
- State how many rows were removed as duplicates.
- What is the most common vessel name?

**Task 2.** For the remaining tasks we will consider only vessels where the home port is either Ardglass (Northern Ireland) or Newlyn (Cornwall). From the full data set (after we remove the duplicates in Task 1), extract two subsets, corresponding to all vessels which have Ardglass or Newlyn as their home ports, respectively. Your report should at least address the following issues:

- Explain how you split out the rows corresponding to a given home port.
- How many vessels have Ardglass as their home port?
- How many vessels have Newlyn as their home port?
- Comparing the two subsets to the full data set, would you consider vessels from the two ports 'typical'?
  *Note:* The two subsets here refer to vessels whose home port are Ardglass and Newlyn. The comparison is in terms of vessels' overall length and engine power, between these two subsets and the whole sample (all vessels). Use suitable plot to make the comparison. There is no need to perform statistical test in this task.

**Task 3.** We want to compare overall lengths and engine powers of vessels based in the two ports. As a first step, plot histograms of overall length and engine power for both ports (*i.e.* four histograms in total). Plot your histograms so that it is easy to compare the two ports.

**Task 4.** Still considering Ardglass and Newlyn, use a Welch $t$-test to test the following hypotheses at 5%-level:

$H_0$ : overall vessel lengths at both ports have the same mean
$H_1$ : overall vessel lengths at both ports have different means

and

$H_0$ : vessel engine powers at both ports have the same mean
$H_1$ : vessel engine powers at both ports have different means.

For both tests, first compute the test statistic yourself (using simple R commands), and then re-do the test using the R function `t.test()` (see `help(t.test)` for how to use this function). Make sure that both methods give the same result. Comment on the applicability of the chosen test, and discuss the results of the test.
*Note:* By default, the command `t.test()` will perform Welch $t$-test in `R`.

**There are some important instructions on how to write and submit your report in the next page. Please turn over.**

**Writing your report:**

a) Clearly mark your report with your name and your student ID.
b) The format of the report should follow the tasks (i.e. address each task, one at a time). So, your report should be either numbered 1, 2, etc according to the task number, or you have some sections with headings 'Task 1', 'Task 2', etc. The most important thing is that the layout and format of the report should be in such a way that it is clear that you are addressing each task individually. These tasks are of course inter related and you may refer to other tasks when addressing one task. For example, suppose you wish to refer to your previous result in Task 1 while explaining the results of Task 2, then you can say "As described in the answer to Task 1, ...", or something along this line.
c) Your report must be typeset (not handwritten) and must not exceed four pages, including all figures and R code. Since space is limited, focus your discussions on the essentials and think about what is most important to include in your report. The academic integrity form does *not* count towards the page limit.
d) Use LaTeX, Microsoft Words, or other text editing softwares. Do not use markdown languages (for example, R markdown). The output should be in (or should be converted to) pdf format.
e) Write complete sentences, including correct punctuation.
f) Use plots to illustrate your results where appropriate, and describe/discuss each plot you include in the report. Make sure you use good axis labels, captions, *etc.* for your plots and make sure that your plots are meaningful and easy to interpret. Do not use screenshots for the plot. R and Rstudio allow you to save the figure into file (either in jpg, pdf, ps, or tiff), which you can import to your report. If you are new to LaTeX, there are some internet resources that describe how to include figure in the tex file.
g) Include and explain all the main relevant R code you use to derive your results, but do not include unused or unnecessary R code. Include the code in the main text rather than into an appendix. If you use LaTeX, put the R codes in verbatim mode. If you use Microsoft Word, put the R codes in `Courier New` font type and a font size of 10 (just for the R codes and its output – not the whole report). Again, do not use screenshots.

**Submitting your report:**

a) The deadline to submit the report is Wednesday 20th March 2024 at 12noon. No submission is possible after 5pm on the day.
b) Don't forget to attach the academic integrity form and submit the report as a single pdf file.
c) You can merge two pdf files into a single file using online resources. Google 'merge PDF online'.
d) You will need to submit the pdf file to **both** *Gradescope* and *Turnitin* via Minerva. The relevant link is 'Submit My Work'.
e) The marking will be done in Gradescope and Turnitin will check for plagiarism. At the moment both features cannot be handled by Gradescope alone, hence submission to two repositories.
f) In Gradescope, **assign all pages to Question 1 or 'Q1'** for the purpose of marking.
g) The deadline applies to submission to *both* Gradescope and Turnitin. I.e., your submission is considered on-time, if you submit the report to both of Gradescope and Turnitin on-time. If one of them is late, then it will considered as late submission. See also Point (i) below.
h) As a consequence of the above point, if you submit your report only to one repository, it will be considered as non-submission.
i) Given the time to submit the report, it is not recommended to submit your report within one hour of the deadline. Please allow yourselves some times. This will make sure if there is a problem in the submission, you are able to deal with that.
j) Submission by email is absolutely not permitted and will not be marked (consequently, zero mark).
k) Note that submission to Turnitin can not be over-written, while submission to Gradescope can be over-written (as long as before the deadline).

**Frequently asked questions**

a) *There is a problem with my internet and the deadline has now passed. What should I do?* Please send the report even if it has passed the deadline (12noon on Wed 20th March 2024). It is better to submit the report late (and get the penalty) than no submission at all or submission via email, which will get zero mark. If you consider this a mitigating

circumstance, then submit your case online. If your case is approved, we will exempt your report.

b) *In the past week, I have not been feeling well. Can I get an extension to the deadline?*
Extension is not possible as explained by your lecturer in Week 1. Submit your mitigating circumstance online and, if approved, then we will exempt your report.

c) *The deadline has passed, can I just send the report to you by email?*
Submission by email will get zero mark. If you are late (but still before 5pm on 20th March 2024), then just submit it to both Gradescope and Turnitin. A late submission is still better than submitting via email, which will get zero mark. If you have mitigating circumstance that has affected you, then submit your case online and, if approved, we will exempt your report.

d) *Hi, I am not sure whether I wrote the report in the right format. So, can I send it to you to check?*
We are not able to check your report before you submit it. As long as you have written the report by addressing each task individually as described above, then it should be OK.

e) *I forgot to attach the Academic Integrity Form. What should I do?*
If you forgot to attach Academic Integrity Form (AIF), send it to MATH1712@leeds.ac.uk (*not* to your lecturer's email address), with the text/statement that the AIF is for Practical 1.

f) *I have had a problem with my R (or laptop). Can I get an extension to submit my report?*
As your lecturer explained in Week 1, extension to deadline is not possible. If it is a software problem, re-installation of R may help. Alternatively, you can use university computers (library, computer cluster, School of Maths), or you can still use Rstudio in the cloud using your web browser (assuming internet connection). Go to `http://rstudio.cloud` and login with your google account (create one if you haven't got one). If you have mitigating circumstance, then submit your case online and, if approved, we will exempt your report.