Department of Statistics
STATS 782 Statistical Computing
Assignment 3(2022.1)

Total: 50 marks **Due: 23:59 NZT, Thursday 19 May 2022**

1. Please read these instructions carefully. Further instructions might be posted on the class webpage.

2. Upload your soft copy (assignment source) to Canvas: the file should end in `.Rmd`, or possibly `.R` or `.Rnw`. The marker may run or knit your R code, so include your name and ID in all files. The file names should contain your UPI. RMarkdown is strongly recommended.

3. Also upload your `.pdf` to Canvas too. **Note the time difference between countries**.

4. Coversheet: please make sure you do **one** of the following else your assignment will not be marked:

   (a) Sign the Cover Sheet and combine with your assignment document (pdf or Word) into a single file before submission, OR

   (b) Type or write for the following at the beginning of your assignment: Your name (as it appears in Canvas), your UPI, and the following statement: "I have read the declaration on the cover sheet and confirm my agreement with it."

5. Include everything in your report: R code (tidied up), outputs (including error/warning messages), and your explanations (if any). Please comment on almost all of your output, especially parts that need human interpretation, else marks will be deducted. That is, you need to convince the marker that you understand what the data or solution is saying.

6. Print some intermediate results to show how your code works step by step, if not obvious. Comment your code if appropriate, e.g., for functions, blocks of code, and key variables.

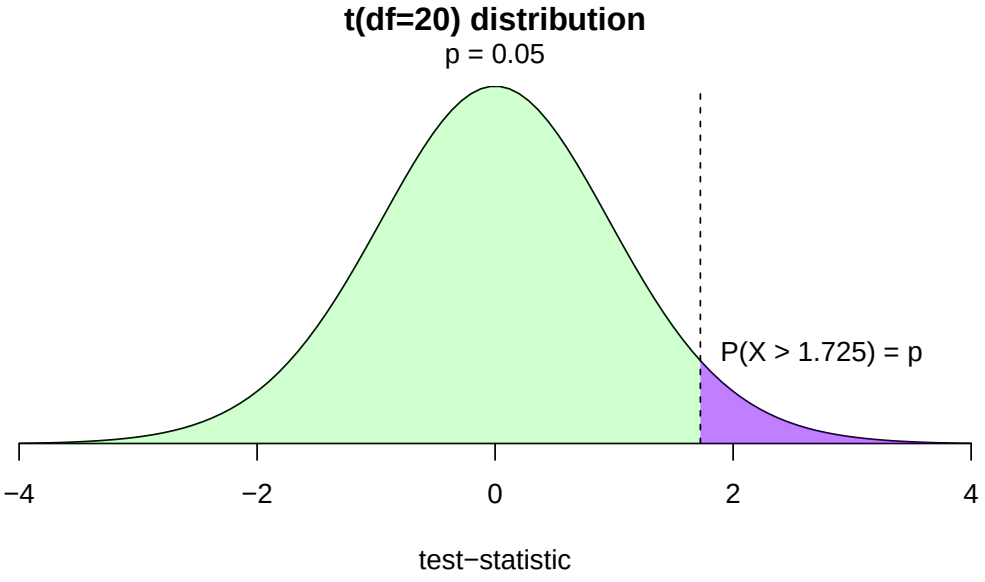
7. Type `help.start()` when you open R. You need to use the online help to find details and functions that may not be covered directly in the coursebook. This requires maturity; we cannot cover everything in class or the coursebook.

8. Your mark for this assignment will depend on getting the right answer, the elegance/efficiency of your approach, and the tidiness and documentation of your code/report. Avoid copy/paste - use programmatic ways to repeat things if needed. **Marks (up to 7) will be deducted for messy code, etc**.

9. Use base R graphics as discussed in the course for this assignment, do not use `ggplot` or other 3rd-party tools, because this assignment is about understanding and building graphics from ground up.

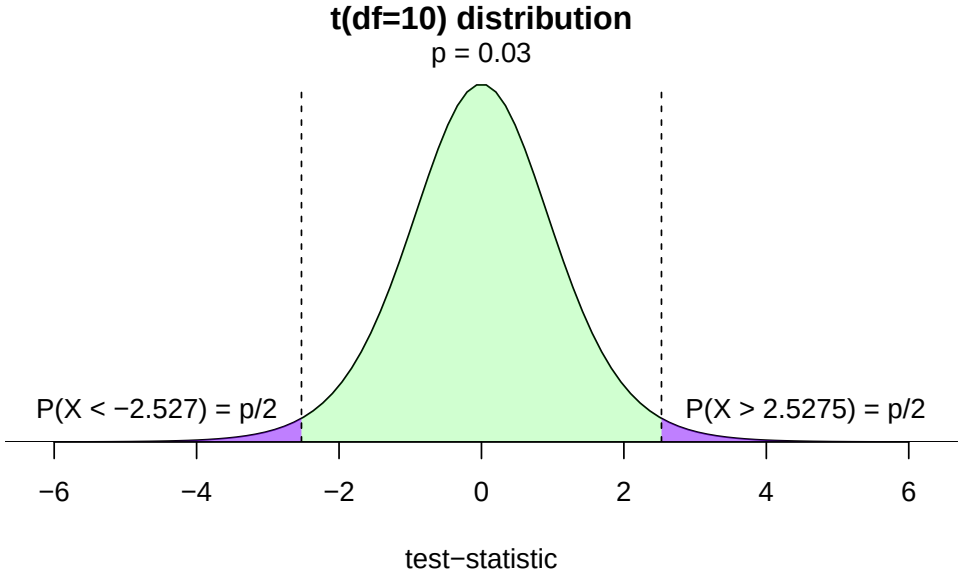
10. This PDF file may contain colour that is important to see.

1. [**19 marks**] You want to explain the one-sample t-test to your friend so you decide to create a plot of the t-distribution with 20 degreees of freedom for illustration purposes. To keep things simple you decide to show the one-sided test for $p = 0.05$ and come up with the following graphic:



(a) Re-create the plot using R as closely as possible. [7 marks]

(b) Write a function `f` with the following parameters: `df` numeric, number of degrees of freedom, `p` numeric, the p-value and `onesided` logical, `TRUE` for two-sided test and `FALSE` for one-sided test. The function should automatically compute the necessary quantiles based on the p value and draw the plot. Therefore calling `f(20, 0.5, TRUE)` should draw the same plot as the previous question (except for the range of x which may vary). For a two-sided test you colour $p/2$ from each (left and right) tail of the distribution.
Test your code with `f(20, 0.5, TRUE)` and `f(10, 0.03, FALSE)` where the latter should look someting like this: [7 marks]



(c) The t-test statistic for the hypothesis that the sample $X = \{x_1, ..., x_n\}$ follows a distribution with mean μ is computed as

$$t = \frac{\bar{X} - \mu}{\frac{\hat{\sigma}}{\sqrt{n}}}$$

where $\bar{X}$ is the sample mean and $\hat{\sigma}$ the sample standard deviation. This statistic follows the t distribution with $n - 1$ degrees of freedom.

Sample 30 values from the normal distribution with mean 3 and standard deviation 2. Compute the t-test statistic for that sample and the five hypotheses: $\mu = \{2, 2.5, 3, 3.5, 4\}$. Use the function `f` from above to draw the corresponding distribution for $p = 0.05$ (two-sided) and superimpose the five resulting test statistics as a thick vertical lines each labeled with the corresponding value of μ. Which of the values fall into the region indicating the the hypothesis is accepted? [5 marks]

2. **[31 marks]** Stats NZ publishes[1] detailed datasets on monthly imports and exports to/from Aotearoa New Zealand which includes the countries and categories of goods. In this question we will use the monthly import statistics for years 2000 through 2021. The orignal data is reasonably large (over 18 million records, 3.2Gb) so we will restrict ourselves to the value of imported goods aggregated by country and month. The resulting dataset can be found in the `imports-by-country.csv` file with the following columns: `"yearmonth"` specifying the year and month in the form `YYYYMM` where `YYYY` is the year and `MM` is the month, `"country"` name of the county the goods are imported from and `"value"` the value of the goods (in NZD) imported that month. Answer the following questions based on this dataset.

   (a) Compute the total value of imports by country over the entire period. List the top three countries from which New Zealand imports (by total value of imports). [4 marks]

   (b) Draw a pie chart of the total value of imports by country using all countries. Discuss two issues with such visualisation (one sentence each). [3 marks]

   (c) Draw a bar chart showing the average annual import value in billions of NZD for the top 15 countries. Pick a suitable orientation and margins such that the names of all countries are fully visible. [5 marks]

   (d) We want to look at the development of the imports over time. To make things more manageable, we want to focus on the top 11 countries and aggregate all other countries into one category `"other"`. Draw a line plot with x-axis being time and y axis the monthly import value in billions. Each of the top 11 countries and `"other"` should by represented by one line (hence 12 lines total). Add a corresponding legend for countries. Discuss the trade evolution of the top three countries over time. [7 marks]

   (e) We want to look at the seasonal aspect of the imports for each country. Write R code to replicate Figure 1. The colours for years are generated using the `hcl()` function with chroma 60 and luminance 70. [8 marks]

   (f) Based on Figure 1, are any countries showing a seasonal effect (a cyclical pattern that is repeated each year) and if so, which? Does this plot enable us to reveal steadily increasing imports? If so, which countries and how can you tell from the plot? [4 marks]

[1] `https://www.stats.govt.nz/large-datasets/csv-files-for-download/overseas-merchandise-trade-datasets`

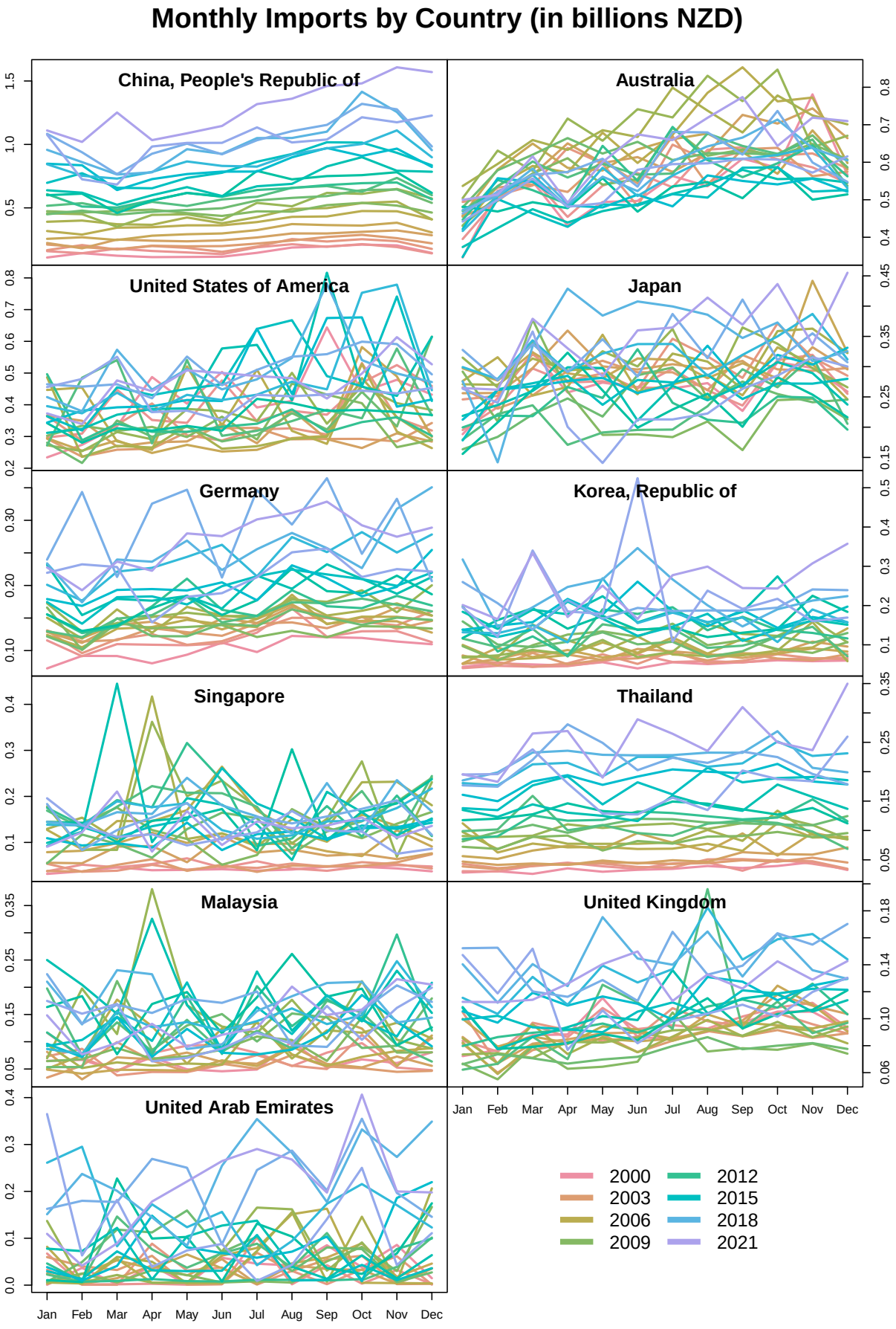


Figure 1: Monthly Imports by Country (in billions NZD)