Summative Assessment for EMATM0061

Statistical Computing and Empirical Methods, Teaching Block 1, 2023

Introduction

This document contains the specification for the summative assessment for the unit EMATM0061, TB1 2023. Please read carefully the following instructions before you start answering the questions.

Deadline. Your report is due on Wednesday 10th January 2024 at 13:00.

Rules: This is *an independent task*. For the summative assessment you should not share your answers with your colleagues.The experience of solving the problems in this project will prepare you for real problems in your career as a data scientist. If someone asks you for the answer, resist! Instead, you can demonstrate how you would solve a similar problem.

Support: Whilst this is an independent task, there is a lot of support available if you need it. Talk to your classmates and book office hours. If you are unclear about what is required for any part of the assessment then discuss this issue with the our teaching team in the computer lab or email `rihuan.ke@bristol.ac.uk`, including the unit code EMATM0061 in the subject of your email.

Plagiarism: Be very careful to avoid plagiarism. For more details, you should consult the "Plagiarism" section within the central Blackboard page for the Data Science MSc.

Extenuating circumstances: For more details on the procedure for extenuating circumstances consult the "Extenuating circumstances" section within the central Blackboard page for the Data Science MSc.

Clarity: Clarity is highly important. Be careful to make sure you clearly explain each step in your answer. You should also include comments within your code when necessary. Your answer should clearly demarcate which part of the question you are answering. Whenever possible, include pieces of well-written codes in your report to promote clarity.

Programming language: *For Part I(A)* of this coursework you should use `Tidyverse` methods within the `R` programming language. For *Part I(B)* and *Part II*, you can use either R, Python or Julia. Regardless of your choice of language, it is essential that your answers are clear and well-written.

Submission points: To submit your solutions, please visit the "Assessment, submission and feedback" tab on the course webpage at Blackboard. Make sure your submission follows the submission structure described below.

Submission structure: Your submission should include the following 4 parts (the submission points will be available at Blackboards):

- Part IA Submission: A report in a single PDF file containing your answers for Part I (A). The report should be submitted to the submission point called Part IA Submission.

- Part IB Submission: A report in a single PDF file containing your answers for Part I (B). The report should be submitted to the submission point called Part IB Submission.
- Part II Submission: A report in a single PDF file containing your answers for Part II. The report should be submitted to the submission point called Part II Submission.
- Source code and data: A single folder that contains three subfolders entitled `username_PartIA`, `username_PartIB` and `username_PartII` where `username` is replaced by your unique UoB username. Each subfolder should include any source code (such as .Rmd, .R, .py files) and data (if applicable) that you used for the corresponding sections (Part I (A), Part I (B) and Part II).

For each of Part IA, Part IB and Part II Submission, if your solutions are in a file format (HTML, images, etc) other than a PDF file, then please convert it to a single PDF file before the submission. It is important that your approach to solving the questions is visible within these reports (PDF files) and we recommend including pieces of clear and well-written code along with explanatory pros within the report itself.

Time allocation: Both Part I and Part II contain 50 marks, but we recommend that you allocate more time for the tasks in Part II, for example 60% on Part II and 40% on Part I.

Part I

Part I (A) (20 marks)

General instruction: In this part of your assessment, you will perform a data wrangling task using *R programming*. Note that clarity is highly important. Be careful to make sure you clearly explain each step in your answer. You should also include comments within your code when necessary. In addition, make the structure of your answer clear through the use of headings. You should also make sure your code is clean by making careful use of Tidyverse methods in R.

1 . First download the files entailed "`global_financial_development.csv`" and "`GFD_indicators.csv`" available within the Assessment section within Blackboard.

The file "`global_financial_development.csv`" contains data about the financial development of different countries (Aruba, Afghanistan, $\cdots$), measured by different indicators (such as stock market return, numbers of bank branches per 100000 adults), from 1960 to 2021. The indicators are represented by the codes on the `indicator_code` column. The file "`GFD_indicators.csv`" contains a list of the indicator names as well as their associated indicator codes.

Load the file "`global_financial_development.csv`" into an R data frame called `df`, and load the file "`GFD_indicators.csv`" into an R data frame called `df_code`.

Display a subset of the data frame `df` consisting of the first 5 rows and the 3 columns `country`, `indicator_code`, `year_2019`.

Display a subset of the data frame `df_cdoe` consisting of the first 5 rows and all columns.

2 . Merge the two data frames `df` and `df_code` into a single data frame called `df_merged` based on the column `indicator_code`. The new data frame should contain all columns of `df` and all columns of `df_code`. Make sure that for each row of `df_merged`, the `indicator_name` is matched to the `indicator_code` according to the correspondence between the indicator names and codes described in `df_code`.

Then remove the `indicator_code` column from the data frame `df_merged`.

Display a subset of `df_merged` consisting of the first 6 rows and the columns `country`, `indicator_name`, `year_2019`.

3 . Use the data frame `df_merged` to create a new data frame called `df_stock` that contains all rows of `df_merged` where the `indicator_name` is `StockMarketReturn`.

Reorder the rows of the data frame `df_stock` such that the values in the columns `year_2019` are in descending order.

Display a subset of `df_stock` consisting of the first 5 rows and the 4 columns `country`, `year_2019`, `year_2020`, `year_2021`.

4 . Use the data frame to generate a data frame called `df_summary` that contains the following information:

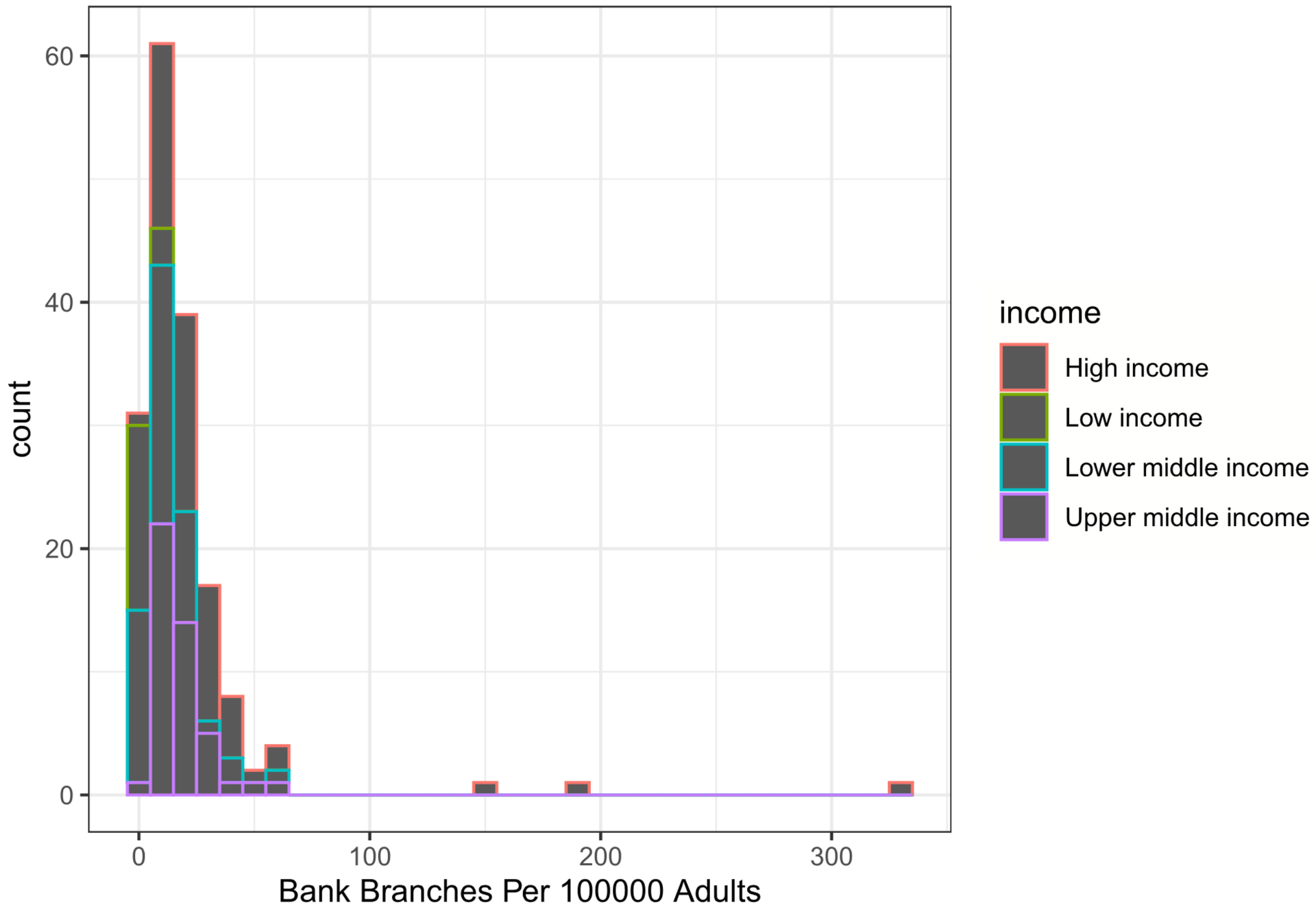
Given a country A and an indicator B, we are interested in a statistic that is defined by the average of the 5 largest values of the indicator B for the country A over all different years. For example, if the values of indicator `"CompaniesPer1000000People"` for country "United Kingdom" for all different years are $a_{1960}, a_{1961}, \cdots, a_{2021}$, we first find the 5 largest non-missing values (i.e., excluding all missing values) of $a_{1960}, \cdots, a_{2021}$ and compute their average. If there are less than 5 non-missing values, compute the average of these non-missing values. If all values are missing, then the computed result should be represented by "NaN" (not a number).

Your data frame `df_summary` should have N rows (where N is the number of different countries in the `country` columns of the data frame `df_merged`) and 6 columns. The first column is called `country` and the other 5 columns are `"BankAccountsPer1000Adults"`, `"BankBranchesPer100000Adults"`, `Top5BankAsset"`, `"CompaniesPer1000000People"`, `"StockMarketReturn"` corresponding to 5 different indicators. Each row of `df_summary` should contain the name of a country and the values of above defined statistics of the 5 corresponding indicators for that country.

Display the first 6 rows of the data frame `df_summary`.

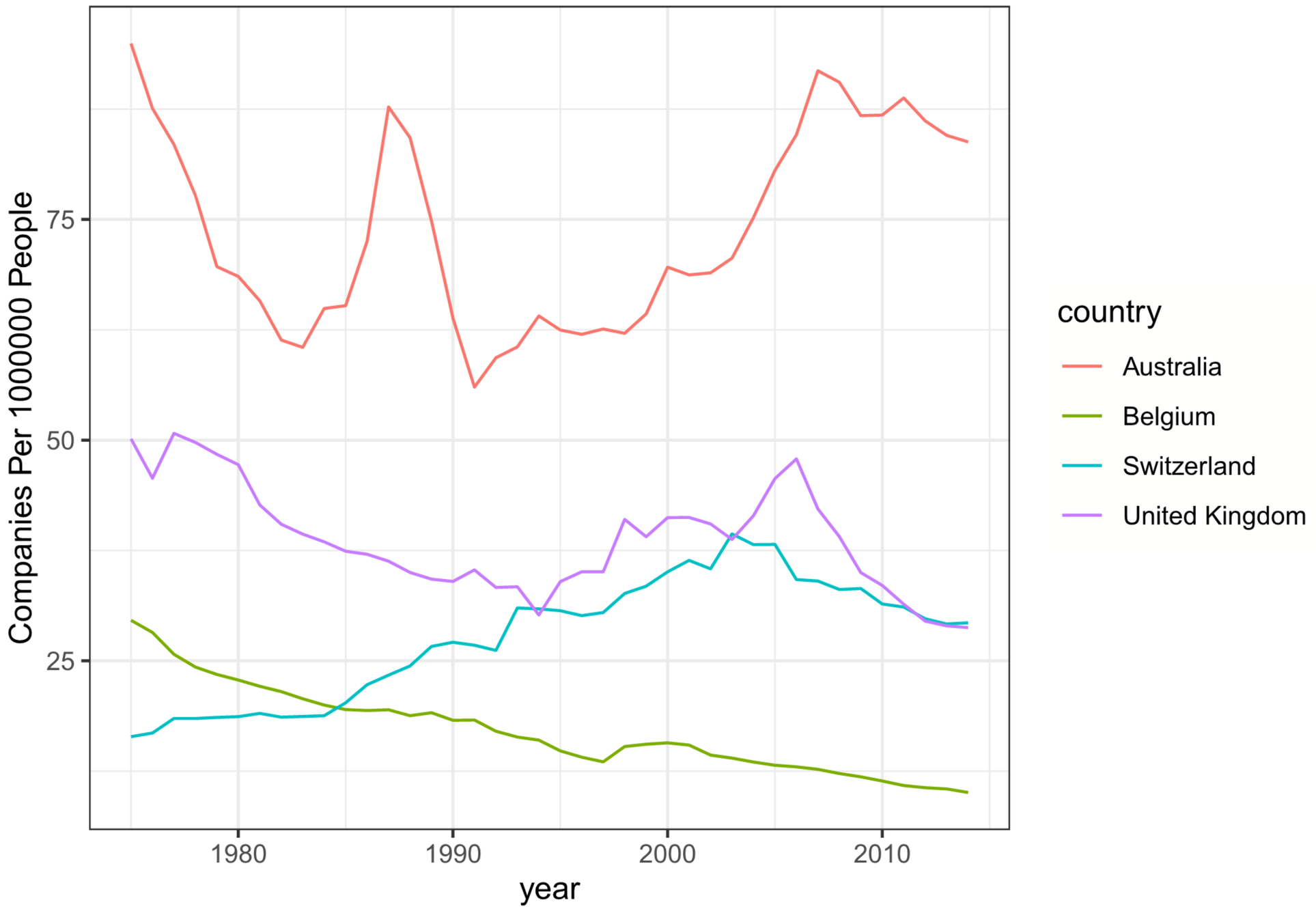
5 . Your data frame `df_merged` contains information about 4 different income groups(from High to Low) of countries. Based on the data frame `df_merged`, create a plot for the histogram of the values of `BankBranchesPer100000Adults` for each of the 4 different income groups in 2019. The bin width of your histogram plot should be set to 10. Ignore all missing values and the `BankBranchesPer100000Adults` values that are bigger than 500.

Your plot is expected to look as follows.



6 . Based on the data frame `df_merged`, create a plot to show the indicator `CompaniesPer1000000People` as a function of `year` (from 1975 to 2014) for the following countries: Australia, Belgium, Switzerland, and United Kingdom.

Your plot is expected to look as follows.



Part I (B) (30 marks)

Let X be a continuous random variable for the time that a customer spends in a supermarket in Bristol in a single visit. Assume that the probability density function of X is

$$f_X(x) = \frac{\beta \alpha^\beta}{x^{\beta+1}}$$

for $x \geq \alpha$ and $f_X(x) = 0$ otherwise. Here α and β are parameters of the distribution, and $\beta > 2$.

Suppose that we have a sequence X_1, X_2, $\cdots$ X_n which are i.i.d. (independent and identically distributed) random variables having the same distribution as that of X. The mean of the sequence is then given by $\bar{X} := (X_1 + X_2 + \cdots + X_n)/n$.

1. Derive a formula for the cumulative distribution function and the quantile function for the random variable X.
2. Derive an expression of the expectation of X in terms of parameters α and β. Similarly, derive an expression of the variance of X.
3. Derive the expectation $\mathbb{E}(\bar{X})$ of the random variable $\bar{X}$ as an expression of α, β, and n. Similarly, derive the expectation of the variance $\mathrm{Var}(\bar{X})$ of the random variable $\bar{X}$. What are the limits of expectation and variance of $\bar{X}$ when n goes to infinity?
4. Create a function called `gen_X` which takes α, β and n as input and outputs a sample $X_1, X_2, \cdots, X_n$ of X.
5. Now assume that $\alpha = 5$ and $\beta = 3$ and $n = 5000$. Carry out an experiment to simulate the distribution function of the random variable $\bar{X}$. In the experiment, generate a sample of $\bar{X}$ that consists of 1000

random numbers $\{a_1, a_2, \cdots, a_n\}$ drawn from the distribution of $\bar{X}$ (to generate a random number from $\bar{X}$, you may generate a sample $X_1, \cdots, X_n$ and then compute its average).

Now associated with the computed $\{a_1, a_2, \cdots, a_{1000}\}$ (as a sample of $\bar{X}$), let $t := \mathrm{SQ}(q)$ be the q sample quantile (for any $q \in (0, 1)$). Here $\mathrm{SQ}(\cdot)$ is a function that maps q to the q sample quantile. Create a scatter plot to plot q as a function of t. Then append to the scatter plot a curve representing the distribution function of normal distribution $\mathcal{N}(\mathbb{E}(X), \mathrm{Var}(X)/n)$. Try to describe the relationship between the scatter plot and the curve, and explain the reason for the observed relationship. Justify your conclusions.

6. Suppose that α is known and β is unknown. We want to estimate the parameter β given the sample $X_1, X_2, \cdots, X_n$. Derive a formula for the likelihood function of β. Derive a formula for the maximum likelihood estimate β_{MLE} of β.

7. Now assume that $\alpha = 5$ and $\beta = 3$. Carry out a simulation study to investigate the bias and consistency of β_{MLE} for different values of n. For each value of $n \in \{10, 20, 30, 40, 50, 100, 200, 300, 400, 500, 1000, 2000, 3000, 4000, 5000\}$, conduct 5000 trials, and in each of these trials, generate a sample $X_1, X_2, \cdots, X_n$ of X. For each generated sample, compute the associated maximum likelihood estimate β_{MLE}. You may want to use visualisation to study the values of β_{MLE}. Based on the experiment, discuss the behaviours of β_{MLE} as n increases: Is β_{MLE} unbiased? Is β_{MLE} consistent? Clearly display the results of your simulations to support your conclusions.

Part II (50 marks)

In this part of the assessment, you are asked to complete a Data Science report which demonstrates your understanding of a statistical method. The goal here is to choose a topic that you find interesting and explore that topic in depth. You are free to choose a topic and data set that interests you.

There will be an opportunity to discuss and get advice on your chosen direction in the computer labs.

Below are two flexible example structures you can consider for this section of your report. If you are unsure what to do, choose one of the following. Note that you should *not* submit more than one of the example tasks below.

Example task 1

Investigate a particular hypothesis test e.g. a Binomial test, a paired Student's t test, an unpaired Student's t test, an F test for ANOVA, a Mann-Whitney U test, a Wilcoxon signed-rank test, a Kruskal Wallis test, or some other test you find interesting.

Note that clarity of presentation is highly important. In addition, you should aim to demonstrate a depth of understanding. For this hypothesis test you are asked to do the following:

1. Give a clear description of the hypothesis test being considered, including the details of the test statistic and p-value, the underlying assumptions, the null hypothesis and the alternative hypothesis. Give an intuitive explanation for why the test statistic is useful in distinguishing between the null and the alternative.

2. Perform a simulation study to investigate the probability of type I error under the null hypothesis for your hypothesis test. Your simulation study should involve randomly generated data which conforms to the null hypothesis. Compare the proportion of rounds where a Type I error is made with the significance level of the test.

3. Apply this hypothesis test to a suitable real-world data set of your choice (some places to find data sets are described below). Ensure that your chosen data set is appropriate for your chosen hypothesis test. For example, if your chosen hypothesis test is an unpaired t-test then your chosen data set must have at least one continuous variable and contain at least two groups. It is recommended that your data set for this task not be too large. You should explain the source and the structure of your data set within your report.

4. Carefully discuss the appropriateness of your statistical test in this setting and how your hypotheses correspond to different aspects of the data set. You may want to use plots to demonstrate the validity of your underlying assumptions. Draw a statistical conclusion and report the value of your test statistic, the p-value and a suitable measure of effect size.

5. Discuss what scientific conclusions you can draw from your hypothesis test. Discuss how these would have differed if the result of your statistical test had differed. Discuss key experimental design considerations necessary for drawing any such scientific conclusion. For example, perhaps an alternative experimental design would have allowed one to draw a conclusion about cause and effect?

6. Demonstrate further understanding by investigating a particular aspect of the hypothesis test you choose. Raise a question/topic that you think is interesting about the hypothesis test and try to give an answer. As an example, you can discuss how the results are affected when an assumption of your hypothesis test does not hold. Or if there is a better option for the testing when we make a stronger assumption? You may also have a close inspection of the relationship on the significant level and power of the test based on examples/simulations $\cdots$. In summary, you can choose a direction/aspect that you are interested in, but it is important to think about how your findings are supported by your experiments or approaches.

Example task 2

Investigate a particular method for supervised learning. This could either be a method for regression or classification but should be a method with at least one tunable hyperparameter. You could choose one from ridge regression, k-nearest neighbour regression, a regression tree, regularized logistic regression, k-nearest neighbour classification, a decision tree, a random forest or another supervised learning technique you find interesting.

Note that clarity of presentation is highly important. In addition, you should aim to demonstrate a depth of understanding.

1. Give a clear description of the supervised learning technique you will use, including the underlying principles and any assumptions. Explain how the training algorithm works and how new predictions are made on test data. Discuss what type of problems this method is appropriate for.

2. Choose a suitable data set to apply this model to and perform a train, validation, and test split (some places to find data sets are described below). Be careful to ensure that your data set is appropriate for your chosen algorithm. For example, if you have chosen to investigate a classification algorithm then your chosen data set must contain at least one categorical variable. Your data set for this task does not need to be large to obtain good results. The size of your data set should not exceed 100MB and you should aim to use a data set well within this limit. Your report should carefully give the source for your data. In addition, describe your data set. How many features are there? How many examples? What type is each of the variables (e.g. Categorical, ordinal, continuous, binary etc.)?

3. Explore how the results of your model on the validation set vary when the model is trained on different random subsets of the training set (the size of the subsets could be fixed, such as at 5% or 10% of the size of the training data set). Given an input in the validation set, the output of your model may be different if it is trained on a different subset (randomly selected), hence given the input the variance of the output is not zero. Compute the variance averaged over the whole validation set. Does the variance become larger or smaller as you enlarge the size of the training subsets?

4. What is an appropriate metric for the performance of your model? Give a clear explanation of the metric. Explore how the performance of your model varies on *both* the train and the validation data as you vary a hyperparameter.

5. Choose a hyper-parameter and report your performance based on the test data. Can you get a better understanding by using cross-validation?

6. Demonstrate further understanding by investigating a particular aspect of the supervised learning method that you choose. Raise a question/topic that you think is interesting about the supervised learning method and try to give an answer. As an example, you can discuss how the variance and bias are defined for supervised learning models and their relationship and how they behave as you vary the hyperparameters or the size of the training set. Or you can discuss the scalability/robustness of the supervised method. Another example is to explore the possible situations where the method fails and give counterexamples. You can also have a closer inspection of the performance of a similar learning method and understand the difference between the two methods $\cdots$. In summary, you can choose a direction/aspect that you are interested in, but it is important to think about how your findings are supported by your experiments or approaches.

Alternative tasks

You could also choose an alternative task in which you explore a statistical method or methods which interest you.

A couple of elements to bear in mind here:

1. Demonstrate a solid level of understanding of the technique or techniques you consider. We expect to see that you understand the basic principles of the techniques, how they work, the problems they address, any assumptions being made, any underlying theory underpinning the techniques, and how to

apply the techniques for solving the problems. Use experiments or simulation studies to demonstrate and investigate the techniques you consider with appropriate reasoning.

2. Apply your chosen method or technique to a real data set. This data must be publicly available and should not exceed 100MB in size.
3. Where appropriate, use simulated data to explore and demonstrate the properties of your chosen method.
4. The subject of your report should be statistical methods or techniques and their performance and behaviour. Whilst you can consider techniques motivated by a particular application, the application itself should not become the focus of your report.

Further instruction for Part II.

Note:

1. Do not complete and submit more than one of the above tasks. These are example tasks and you should only choose one. The goal here is to explore a topic in detail.
2. You will be graded on *the level of understanding of the key concepts* demonstrated within your report. Additional marks will be given for more advanced methods, provided that a very strong level of understanding is displayed. However, you should avoid choosing complex methods without properly demonstrating your understanding. The main focus here is a clear understanding and you should not sacrifice understanding for the sake of complexity. A clear understanding of the basic concepts is paramount.
3. You do not need to use large data sets. The dataset you choose should not be larger than 100MB. This is an upper bound. You should aim to use a data set well within this limit.
4. We expect that your approach should be visual and clear within the report itself. Therefore it is highly recommended to include pieces of clear and well-written code along with necessary comments and explanations within the report itself.
5. We expect that you interpret and make sense of the experiment results obtained, instead of displaying a list of the results without explanation or analysis. A high quality report should be able to use the experimental results to support its conclusions and findings in a consistent manner.
6. We do not have a page limit for the report. A *rough* guideline is that your report should ideally be no more than 10 pages, *if* all figures, tables, and large pieces of code were removed. However, this is not a strict constraint. Again, clarity is highly important, and you should include sufficient details to demonstrate your approach and the level of understanding of the key concepts.

Data sets

There are a vast number of freely available data sets across the internet. Below is a few example sources. You are also welcome to use data sets from other sources. Any data you use should be freely available and accessible. The source of your data and the steps required to retrieve it should also be described within your main report.

You should also explain its structure e.g. the number of rows and the number of columns, and what the data in each column of interest represent for, $\cdots$. You are encouraged to use tabular data throughout.

https://www.kdnuggets.com/datasets/index.html

https://r-dir.com/reference/datasets.html

http://archive.ics.uci.edu/ml/datasets.php

http://lib.stat.cmu.edu/datasets/

http://inforumweb.umd.edu/econdata/econdata.html

https://lionbridge.ai/datasets/the-50-best-free-datasets-for-machine-learning/

https://www.kaggle.com/

https://www.ukdataservice.ac.uk/

https://data.worldbank.org/

https://www.imf.org/en/Data

Final remarks

Throughout your report you should emphasise:

- Reproducible analysis (be careful with randomised procedures).
- Clear and informative visualisations of your results.
- Demonstrate a depth of understanding.
- A clear writing style.