Part 1: short answer questions

The objective of this task is to apply your knowledge of M/M/1 queue diagnostics to answer some questions about a scenario involving an M/M/1 queue. You will need to interpret the questions and use correct formulas and mathematical notation to fully justify your answers.

A small business has recently purchased a specialised printing and binding machine that can read a PDF document and automatically print and bind the document. The business has designed an online interface where customers can upload documents, and then the specialised machine will automatically print and bind the document as soon as it is ready. Immediately after a customer uploads a document to the website, the document is added to a queue for processing and status updates will be made available for the customer. The status updates given to the customer are *Waiting*, *In Progress* and *Ready*. Upon being uploaded, the status will display *Waiting* until the specialised machine begins printing the document. Once it starts being printed, it will display *In Progress*, and when it is ready for the customer to pick up from the store, it will show *Ready*.

The store advertises a pickup time of 3 hours, suggesting to customers that, on average, they will be able to pick up their document approximately 3 hours after uploading it. A customer will be satisfied if they can pick up their document exactly after the advertised pickup time, and dissatisfied otherwise.

Assume that the machine runs every day, non-stop, without any breaks. Based on historical data, an average of 37 documents are uploaded per day, and the machine takes an average time of 29 minutes to print and bind the document. To analyse the system, the company has decided to model the service as an M/M/1 queue using these parameters.

For Part 1 of the project, you may only use facts and results that have been given in the subject materials, unless otherwise stated.

1. State the arrival rate, service rate and traffic intensity for this queue. Include appropriate units.
2. Determine the average time taken for the status to change from *Waiting* to *In Progress*.
3. Determine the probability that a document takes 16 minutes to finish being printed and bound after it starts being printed.
4. Let N denote the number of documents in the queue. State the probability distribution that N follows, giving your answer in the form $N \sim$ _____. Then use this to determine, on average, how many hours per day the machine is inactive.
5. Is the store's advertised pickup time reasonable? Explain.

You are now given the following general fact that may be used for the remaining questions: for an M/M/1 queue with $\rho < 1$, if the random variable T denotes the total time an individual spends in the queue, then $T \sim \text{Exp}(\mu - \lambda)$.

6. Explain how the given fact results in the formula for the average time spent in the queue.
7. Using the given fact, determine the probability that a customer will be satisfied.
8. Assume that the store wants at least 95% of customers to be satisfied based on the advertised pickup time. Should they change their advertisement? Explain. If so, suggest a suitable change.
9. Suppose that exactly 40 customers will upload a document tomorrow. Let C denote the number of tomorrow's satisfied customers (using the currently advertised pickup time). What probability distribution does C follow? Give your answer in the form $C \sim$ _____.
10. The store would like to understand the effects of increased demand on the system. Provide some insight to the store that would help them understand the level of demand that would make the system unmanagable.

You must give the result of any calculations to at least 3 decimal places.

What to submit

1. Filename: `part1.pdf` or `part1.docx`:

   A PDF or Word document with your answers, neatly typed, with all working shown.

Mark allocation

A total of 20 marks is allocated to Part 1:

- 15 marks for correctness of solutions.
  - The marks for each question are indicated above.
  - The majority of your marks will be awarded for correct working out. An incorrect final answer with otherwise correct working will result in a minor deduction, whereas answers without working will be awarded zero marks (unless otherwise stated).
  - The keyword "state" indicates that working is not required, for example, "State the arrival rate..." means you can write down the arrival rate without any working out.
  - Your answers will be marked consequentially, so if you make an error in an earlier part but use that error correctly in a later part, you can still get full marks for the later part.
- 5 marks for presentation and communication.
  - Marks will be allocated for layout, clarity, readability, appropriate use of vocabulary and language, and use of correct notation and symbols.

Some advice

- This part is based on the content learned in Week 3.
- Use words and sentences, as well as mathematical notation, to clearly communicate your answers.
- You will need to type your solutions to this part, inluding your working. It is possible to do this using the equation editor in Microsoft Word, but you may instead like to write this part using the mathematical typesetting language LaTeX. See here for a document about LaTeX written especially for STM4PSD: https://www.overleaf.com/read/cqyqpcwxtvts

Part 2: investigation of probability distributions and the central limit theorem

The objective of this task is to investigate and implement a continuous probability distribution of your choice. You will start by finding a continuous probability distribution that meets some specific novelty and suitability requirements outlined below. Once you have identified a suitable distribution, you will write code in R to implement various functions for your chosen distribution. Finally, you will write a concise report that includes a short summary of the distribution, the results of some calculations using your distribution, and a demonstration of the central limit theorem using your distribution.

The first part of the task is to decide the continuous probability distribution you will use. It is worth noting that there are many probability distributions meeting these requirements that you can choose from; the best way to start may be by searching the internet.

The following two **novelty requirements** must be met:

- The probability distribution you choose must have at least two parameters.
- The density function for your distribution cannot be one that has been mentioned in any of the subject materials.

The following two **suitability requirements** must also be met:

- The density function for your distribution must equal zero everywhere except for an interval of finite length.
- The density function for your distribution must be continuous on that interval.

The next part of the task is to write some code using R. You have been provided with a file (`part2_main.R`) that can be used to generate random numbers according to a probability distribution that meets the suitability requirements above. For the random number generation to work, there are two functions that must be defined: the `nonzero.interval` function and `distribution` function. You also also required to define some other functions that are not needed for random number generation, as specified below. The functions you must write in the `part2_main.R` file are:

- A function called `nonzero.interval` that takes the parameters of your distribution as arguments and returns a vector containing two elements, where the first element of the vector is the lower bound of the interval needed in the suitability requirements, and the second element is the upper bound.
- A function called `density` that takes an x-value and the parameters of your distribution as arguments and calculates the probability density function for your distribution at x using those parameters.
- A function called `distribution` that takes an x-value and the parameters of your distribution as arguments and calculates the probability distribution function for your distribution at x using those parameters.
- A function called `expected.value` that takes the parameters of your distribution as arguments and calculates the expected value of the distribution using those parameters.
- A function called `variance` that takes the parameters of your distribution as arguments and calculates the variance of the distribution using those parameters.

The last part of the task is to write a short report (no more than 400 words) that demonstrates your chosen distribution and illustrates the central limit theorem using your probability distribution. In your report, you must:

- Give the name of the distribution, state its parameters (and any constraints on those parameters), the notation used for the distribution, and give the formula for its density function. To ensure the accuracy of the distribution's name, parameters, and formula, you must provide a reference from a reliable textbook or journal article (a reference to a website will not be accepted).
- Choose some specific values of the parameters and then:
  - Provide plots of the the probability density function and probability distribution function using those parameters, and explain how these plots demonstrate the suitability requirements.
  - Choose some calculations to perform using the functions you defined, and then list the results using correct notation.
  - Explain the implications of the central limit theorem as sample sizes increases. To aid your explanation, use the provided `generate` function to generate suitable random samples using the same parameters and use them to construct some plots.

  All plots must be generated using R.
- Include any code used to supplement your report in a separate file for submission. Your report must not include detailed code or calculations in the main body; instead, be descriptive about what you are doing with the code and why. For example, starting sentences with phrases like "By using R to calculate..." would be suitable.

What to submit

1. Filename: `part2_main.R`

   Your modified version of the `part2_main.R` file.

2. Filename: `part2_extra.R`:

   An R script file containing the code used to supplement your report (code used to generate plots, samples, etc.). The code must be commented so that the purpose of each part is clear.

3. Filename: `part2_report.pdf` or `part2_report.docx`:

   A PDF or Word document of your report (no more than 400 words).

Mark allocation

A total of 40 marks is allocated to Part 2:

- 20 marks for correctly implementing the required functions.
  - The implementation must be mathematically correct, coded correctly, and demonstrate your understanding of the underlying probability concepts.
- 15 marks for the report.
  - The report must effectively cover all requirements listed above, with an accurate description of your probability distribution, clear density and distribution plots, relevant explanations, and an insightful explanation of the central limit theorem. Any code used to supplement the report must be accurate.
- 5 marks for presentation and communication.
  - The report must be clear and well organised, employ suitable probability and statistics terminology, be well written and engaging and be free of spelling and grammar errors.

Some advice

- This part is based on the content learned in Weeks 3 and 4.
- An example of a distribution that meets the suitability requirements is the uniform distribution, since it has two parameters, a and b, and it is non-zero and continuous on the interval $[a, b]$. The uniform distribution does not meet the novelty requirements, because we have studied it in this subject.
- Examples of distributions that do not meet the suitability requirements are the normal distribution and the exponential distribution. This is because the normal distribution is non-zero on the interval $(-\infty, \infty)$, and the exponential distribution is non-zero on the interval $[0, \infty)$, both of which are infinite.
- If you have found a probability distribution you would like to use, but you are not sure if it meets the novelty and suitability requirements, please feel free to consult the teaching staff.
- A partially completed example to illustrate how the `generate` function works using the uniform distribution has been supplied to you in the file `part2_sample.R`.

Part 3: data analysis

The objective of this task is to perform analysis of a real life data set using the techniques we have learned in this subject. You have been supplied with a CSV file called `bikes.csv`, which contains 56 entries recording the number of users of the Capital Bikeshare bicycle-sharing system based in Washington D.C., USA, with 28 samples from the year 2011 and 28 from the year 2012. Each row includes the year, month and date of the record, and the number of casual users and the number of registered users on that day. For example, the first row of the data is:

year	month	date	casual	registered
2011	2	9	53	1552

This says that on February 9, 2011, there were 53 casual bikeshare users and 1552 registered bikeshare users.

Imagine that you have been approached by an executive with no background in statistics who is interested in the usage patterns of the bikeshare system. Specifically, they would like information about:

- The average number of casual users, registered users, and total users per day.
- The proportion of registered users amongst all users.
- Differences in usage by casual users versus registered users.
- Whether there are yearly differences in bikeshare usage.

Drawing on your expertise, they have requested that you analyse the data to provide meaningful insights on these facets of the bikesharing system. Your task is to use R to investigate the data and write a short summary for the executive. In your investigation, you must:

- Apply the techniques learned in Weeks 4 and 5 to construct *interval estimates* that address the areas of interest for the executive.
- Decide when it is appropriate to use a confidence interval for a single mean, for a proportion, for a paired difference in two means, and for an unpaired difference in two means.
- Use each of the four types of interval listed in the previous bullet point at least once.
- Use only the built-in R functions that have been studied in the workshops and reading materials.
- Write your code in a generic way, so that someone who has a CSV with the same column labels but different data could run your code and get the correct results.
- Write a short summary of the results of your investigation for the executive (no more than 300 words) with limited use of technical language.

What to submit

1. Filename: `part3_code.R`

   A script file containing the code you used to conduct the analysis.

2. Filename: `part3_summary.pdf` or `part3_summary.docx`:

   A PDF or Word document containing the summary for the executive (no more than 300 words).

Mark allocation

A total of 40 marks is allocated to Part 3:

- 20 marks for the analysis.
  - This will be based on the code you submit. The analysis must be comprehensive, covering all of the details outlined above, and performed correctly, with each technique used appropriately.
- 15 marks for the summary.
  - The summary must be clear and concise, capturing the key insights from the analysis. It must also be written with limited technical language, as the intended audience has no background in statistics.
- 5 marks for presentation and communication.
  - The summary must be clear and well organised, employ limited technical language, be well written and engaging, and be free of spelling and grammar errors.

Some advice

- This part is based on the content learned in Weeks 4 and 5.
- You do not need to design any functions yourself: we have learned how to use R's built in functions to calculate confidence intervals.
- You can filter for entries meeting certain conditions using the following syntax:

```
bikes.march <- bikes[bikes$month == 3, ]
```

  The command above will extract the rows where the month is March, and stores the resulting data frame in the `bikes.march` variable. **Note that the final comma is an essential part of the syntax** (indicating that all columns should be retrieved).