Assignment 2

AUTHOR
Statistical Thinking (ETC2420 / ETC5242)

PUBLISHED
Semester 2, 2025

Instructions

This assignment is due on Monday 13 October 2025 (Week 11) at 23:55.

This is a group assignment. Only one submission for each group is required.

See below for details of the tasks and the required format for the submission. Make sure you read all of the details carefully. All groups are to do all tasks in this assignment, regardless of your unit code.

Late submission: Because this is a group assignment, individual requests for extensions cannot be made via the Special Consideration process. We expect that a whole group would not require an extension, because there is plenty of time to complete the assignment and others in the group can cover if someone is sick at the last minute. However, if you think your group has a case for an extension, please send us a request to the **unit email address** (etc2420.clayton-x@monash.edu) and **copy in all of your group members**. You will need to provide a convincing case that takes into account the circumstances of your group. The penalty for late submission is 5% per working day or part thereof.

Effective strategies for group work

Communicate with each other early and organise times for when you will meet together, and how you will contact each other between meetings.

Our recommended strategy for tackling this assignment is to have each group member first attempt each task on their own, and then meet together to share your ideas. At that point, you can decide on the best set of ideas and approaches, including whether additional work is required. Be sure to leave enough time to combine your efforts and review the final report before submitting the assignment.

It is usually a poor strategy to allocate separate tasks to individuals and then just to present them together as if they were undertaken collectively by the group. You will lose the opportunity to get multiple perspectives and learn from each other.

Interacting beyond your group

You may *verbally discuss* ideas about the assignment with any of your classmates. However, all groups must separately prepare and write their own submission, which should accurately reflect the work and efforts of all students in your group.

We highly recommend that you *do not* share any files or written notes between groups, which will be treated as collusion.

Any discussion on the online forum should be limited to *seeking clarifications* about the instructions and tasks.

Preparing your submission

Your responses should be written in a Quarto file that compiles to produce the desired results.

This assignment consists of several Tasks. Each should be answered using a combination of written text and output from R, in the format of a mini-report. All answers need explanation and justification. Any recommendations and conclusions you draw should be supported with evidence. Your descriptions will need both a technical and a non-technical component (see the Tasks for more details).

Organise your submission carefully with headings and a clear section structure. We recommend that your response for each Task form its own section, with further sub-sections if/as required.

Presentation is important, as is conveying your results/recommendations/findings in a succinct and informative way. You need to include anything that is relevant to your specific analyses, and avoid including superfluous information.

All plots must be properly labelled and explained. For example, the definition of each axis should be clear and also what each visual element (points, lines, etc.) represents.

Show and evaluate all code chunks, and format the output so that it does not run off the page when rendered as a PDF. You may also suppress any inconsequential messages and warnings. Once you have finished drafting your work (and verified that are no messages and warnings that require attention), you can include the following as the first code chunk in your Quarto file:

```
knitr::opts_chunk$set(echo = TRUE, warning = FALSE,
                      eval = TRUE, message = FALSE)
```

Anything that is not part of your answer should **not** be included in the Quarto file, even if it is not evaluated or does not appear in the rendered PDF file.

Your Group Number (as allocated on Moodle), and all group members' names and student ID numbers must be stated in the YAML section of the Quarto file and on any other files submitted.

Submission

You need to upload **two (2) separate files** via Moodle for your group: the Quarto file and the corresponding PDF file. You may create the output in another format (such as HTML or Word) as an intermediate step, but you **must** convert it to PDF for submission, and it needs to be in a neat and

presentable format.

Your submitted files should be named according to the following convention:

```
GroupNumber_A2.qmd
GroupNumber_A2.pdf
```

Replace `GroupNumber` with your Group Number (as shown on Moodle). For example, if your Group Number is "Group92", your files should be named:

```
Group92_A2.qmd
Group92_A2.pdf
```

Submission of incorrect files or file names will lead to loss of marks.

Your files must be submitted **separately**, rather than combined together as a ZIP file or similar.

Peer evaluation

In addition to submitting the files, you will also need to complete a peer evaluation survey to provide feedback on the contribution of your team members. The survey link will be provided on Moodle.

Marking

There is a total of 100 marks available for this assignment.

There are five (5) broad Tasks, which together consist of 90 marks (specific mark allocations are given below). The remaining 10 marks will be awarded for your submission as a whole, as follows:

- Overall presentation [5 marks]. Whether the document is structured in a neat and clear manner, and is easy to navigate.
- Reproducibility [5 marks]. Whether the submitted Quarto file compiles as submitted without error.

We expect each group member to contribute equally to each group, and to complete the peer evaluation survey.

Marks may be deducted if any instructions are not followed, and responses from the peer evaluation survey may be used to adjust your individual mark.

Introduction

You are the statistician in the infrastructure department at the City of Melbourne. One of your department's responsibilities is ensuring that people can walk easily throughout the city using the footpaths and pedestrian crossings

footpaths and pedestrian crossings.

To help you assess how many people travel through some key points in the city, you have commissioned a pedestrian traffic survey. You need to use the data collected from these surveys to help your manager with some operational decisions (described across the various tasks, below).

Your manager has some understanding of descriptive statistics, but is not very familiar with statistical modelling and inference. However, she is happy for you to use any techniques and to explain the results to her in a non-technical way.

Data

At 12pm–1pm each day, trained traffic surveyors count the number of people that travel through three designated pedestrian crossings outside the following places:

- Flinders Street Station
- Southern Cross Station
- QV Melbourne.

The surveys are each done at the same time, across several weekdays chosen randomly throughout the year.

The file called `pedestrians.csv` contains the counts, with one row for each day.

(While the surveys are only carried out in the middle of the day, this was deemed adequate for the purposes of informing the questions and decisions posed below.)

Tasks

For each of Tasks 1–5, there is a set of required analyses and components that you should carry out and include in your response. For each Task, write a short report that includes:

- Explanations of what analyses you are carrying out.
- The results of your analyses, including any plots, numerical results and conclusions.
- Enough details about your analysis choices and assumptions so that someone with technical knowledge (e.g., another statistician) would understand them and could replicate your work.
- A summary of your methods and findings, and any key assumptions, in a form that would be understandable to the manager (e.g., using non-technical language). Include this as the final paragraph or subsection of your response to each Task.

You should be able to cover what you need in about 3 or 4 paragraphs of text for each Task. Note that some Tasks may require more explanation than others.

Task 1 [20 marks]

Your manager wants to understand the distribution of the number of people using the three pedestrian crossings. She would like some appropriate descriptive and visual summaries of the data. In addition, she would like fitted models for each crossing, which she wants to share with the engineering team to help with planning and design.

Required analyses and components:

- One or more appropriate plots to visualise the data.
- Some appropriate descriptive statistics.
- Fit appropriate models to the data, and report the resulting estimates.
- Assess the adequacy of the model fits, using appropriate statistical techniques.

Task 2 [20 marks]

The traffic regulations require that crossings are designed to have smooth pedestrian flow 90% of the time. To help the engineering team assess whether the crossings are large enough to meet the regulations, they would like an estimate of 90th percentile amount of traffic for each crossing. Use two approaches (as outlined below) and provide interval estimates for each.

Required analyses and components:

- As the first approach, use the *sample quantile* as the estimator. This is implemented in R by the function `quantile()`.
- As the second approach, use estimators based on the models you fitted in Task 1.
- Calculate 95% confidence intervals based on each of the estimators above. Use any method that you think is appropriate, and justify your choice.
- Provide an overall conclusion, including recommendations about which of these estimates are likely to be more useful or reliable.
- In your report, explain the key statistical ideas to the manager, including what the parameter of interest is, how it relates to the data, and what assumptions you are making for each estimate.

Task 3 [16 marks]

The engineers have identified that they need to make some upgrades to the crossings outside of the two railway stations (Flinders Street and Southern Cross). To save money, they are hoping they can use the same design and materials for both crossings. They can do so only if the average number of people using each crossing is similar. Specifically, the difference between the averages needs to be no more than 80 people per hour.

needs to be no more than 30 people per hour.

Can the engineers safely save money in this case? Is the difference small enough and how certain can we be?

Required analyses and components:

- Calculate an appropriate 95% confidence interval for the difference in the mean number of people crossing outside the two railway stations.
- Comment on whether it is valid to use the Central Limit Theorem when calculating this confidence interval.
- Based on your confidence interval, what conclusions can you provide to the engineering team?
- Illustrate your analysis with any appropriate plots.

Task 4 [16 marks]

The marketing team learns about your survey and wishes to use it to help raise revenue by selling billboard space near each pedestrian crossing. They can use the collected data to quantify the traffic at each location and convince advertisers to pay more for high-traffic locations.

To start with, the team wants a report about the potential value of each location. This can be assessed by the average number of people (per hour) that pass through each intersection.

Required analyses and components:

- Calculate an appropriate interval estimate for the mean number of people crossing at each of the three locations.
- Comment on the relative value of each location.
- Present your report using an appropriate visualisation.

Task 5 [18 marks]

The marketing team is in discussion with a large bank who is interested in buying some advertising space on the billboards. The bank is willing to spend more money on billboards that will be seen by more people. They offer the following terms: they will pay a minimum of \$10,000 for a billboard, and will offer a bonus payment of \$5,000 $\times \theta$ where θ is the proportion of the time (i.e., the proportion of days) that the number of people that walk past the billboard exceeds 1000 people per hour. For each location, estimate the potential revenue under this offer.

Required analyses and components:

- Use an appropriate method to estimate the expected revenue for each location. (Note: there is more than one way of doing this.)

more than one way of doing this.)

- Calculate corresponding interval estimates for each of these quantities.
- Include an appropriate visualisation of your results.