Statistical Thinking (ETC2420 / ETC5242)

Assignment 1

Semester 2, 2024

Instructions

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

All groups are to do all tasks in this assignment, regardless of your unit code.

Effective strategies for group work:

- Experience suggests it is a poor strategy for a group to allocate separate tasks to individuals and then just to present them together as if they were each undertaken collectively by the group.
- The best strategy is to have each group member attempt each task on their own, with the group meeting together after a few days to review and 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.

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 an R Markdown 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 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 using the `echo = TRUE` and `eval = TRUE` chunk options, and format the output so that it does not run off the page when printed. You can also suppress all other messages and warnings, as per the following command (to be included in the first code chunk of your R Markdown file):

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

Anything that is not part of your answer should **not** be included in the R Markdown 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 R Markdown file and on any other files submitted.

Submission

This assignment is due on Thursday 10 October 2024 at 23:55.

You need to upload **two (2) separate files** to the link on Moodle for your group: the Rmd file and the corresponding PDF file. You may create an HTML or Word document from your Rmd file, but you **must** convert this to PDF for submission.

Your files should be named according to the following templates:

`GroupNumber_A1.Rmd`
`GroupNumber_A1.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_A1.Rmd`
`Group92_A1.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.

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].
- Reproducibility [5 marks]. Whether the submitted Rmd files compile as submitted without error.

Peer evaluation

As part of the marking, you will also need to complete a peer evaluation survey to rate the contribution of each team member. The survey responses may be used to adjust your individual marks. We will provide more information about this process prior to the due date.

Introduction

A bank manager would like to get a better understanding of her current customers, of which there are 780,000. She has commissioned you to survey them to find out their incomes and education levels, and then analyse their responses. She has some specific questions the she wants you to answer, which her team will use to help inform their product design and marketing strategy.

The manager has some understanding of descriptive statistics, but is not very familiar with statistical modelling and inference. However, she has heard about various approaches, such as the bootstrap and Bayesian inference, and is happy for you to use any techniques and to explain the results to her in a non-technical way.

Data

Using a database of all of the bank's customers, you select 200 at random and send them a questionnaire to ask the following questions:

1. What was your annual gross income in the previous financial year?
2. What is the highest level of education you have received?
   Possible options include:
   - Doctorate
   - Master degree
   - Bachelor degree
   - High school or lower

The file called `banksurvey.csv` contains the customers' responses, with one row per customer.
The `income` column is in units of thousands of dollars.

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 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).

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]

The manager wants to understand the distribution of her customers' incomes. She would like some appropriate descriptive and visual summaries of the data. In addition, she would like a fitted model, which her team will use to help inform their product design and marketing strategy.

Required analyses and components:

- One or more appropriate plots to visualise the data.
- Some appropriate descriptive statistics.
- Explore fitting various models. Include at least the following distributions: normal, exponential, gamma.
- Your exploration should include using appropriate QQ plots and calculating maximum likelihood estimates.

Task 2 [18 marks]

The manager is particularly interested in estimating the 80th percentile of annual income of her customers. Use a variety of approaches and provide interval estimates for each.

Required analyses and components:

- Use estimators that are based on fitting a model. Provide at least one estimate for each model fitted in Task 1 (therefore, you need at least 3 such estimates).
- It is also possible to use an estimator that doesn't assume any specific model, by using a statistic known as a *sample quantile*. There are many ways to define such a statistic, most of which are implemented in the R function `quantile()`. The default choice for that function is what are known as 'Type 7' quantiles (as specified by the argument `type = 7`). Using this default choice, estimate the 80th percentile of annual income.
- Calculate 95% confidence intervals based on each of the estimators above. (Use any method that you think is appropriate, and justify your choice.)
- 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.
- Provide an overall conclusion, including recommendations about which of these estimates are likely to be more useful or reliable.

Task 3 [18 marks]

The manager wants to learn more about the different approaches to estimation that you have used in Task 2, including their strengths and weaknesses. Explain and illustrate the statistical properties of each estimator.

Required analyses and components:

- Simulate the sampling distribution of each estimator, assuming the (population) distribution of annual incomes follows an exponential distribution with mean $100,000.
- Compare these sampling distributions using appropriate plots and numerical summaries. For the latter, you should include at least the bias and standard deviation.
- Explain how to interpret these simulation results, and how you could use this type of simulation to select an appropriate estimator to use.
- Explain what you expect to see in your results if you simulated from a gamma distribution (with shape parameter not equal to 1) instead of an exponential distribution.

Task 4 [16 marks]

The manager's team wants to understand how their customers' income relates to their level of education. They are particularly interested in whether there is a large difference in average income between those who completed a university degree and those who haven't. Also, the team wants to know whether any such difference (if there is one) can be attributed to the fact that customers who more highly educated are able to find higher paid jobs.

Required analyses and components:

- Show appropriate plots that compare how the distribution of income varies across each of the four education groups.
- Calculate a 95% confidence interval for the difference in the mean annual incomes of customers who have completed a university degree and those who haven't.
- Comment on whether it would be appropriate to use the Central Limit Theorem when calculating this confidence interval.
- Explain whether these data would be able to provide evidence of a causal effect of higher education leading to higher incomes. Specifically, if the confidence interval were to show a large positive difference, does this survey result imply that providing greater education would lead to higher incomes? (Remember to give reasons that clearly justify your answer.)

Task 5 [18 marks]

The manager wants to know whether there is a big difference in the 80th percentile incomes of those with and without a university degree.

Required analyses and components:

- Set this up using a Bayesian inference approach, as follows:
  - Model each of the two education groups (with and without a university degree) separately.
  - For each group, assume an exponential distribution for the annual income. Each group will have its own rate parameter, λ_g, where $g = 0$ refers to the group with no university degree and $g = 1$ to the group that has a university degree.
  - For each group, use a conjugate prior. Select hyper-parameter values such that for each λ_g it gives a prior mean of 1 and prior variance of 1.
- State the posterior distribution for each λ_g and calculate 95% credible intervals.
- Simulate 10,000 values from the posterior distribution of the *difference* between the 80th percentile incomes between the two groups. (Note: you can't simulate these directly, so think carefully about what you *can* simulate from and what steps you need to take to achieve the desired simulated values.)
- Calculate an approximate 95% credible interval for the difference between the 80th percentile incomes between the two groups.