STA238 - Winter 2021

Assignment 3 Instructions

Samantha-Jo Caetano

Instructions

This is a group assignment. You are expected to work on this in a group of up to for 4. Your group-mates can consist of students in either lecture section. You are expected to work exclusively with your group-mates and not other groups. You are more than welcome to discuss ideas, code, concepts, etc. regarding this assignment with your class mates. Please do not share your code or your written text with peers outside of your group. It is expected that all code and written work should be written by members of your group (unless they are taken from the materials provided in this course or are from a credible source which you have cited). You have the option to work in a group of smaller than 4, but please note that we do not recommend this, as the workload of this assignment is for groups of size 4. Please note, this assignment is fairly open, so the context of most of the work completed here should not match that of other groups.

There is a starter Rmd file (called Assignment3.Rmd) available for you to use to start your code.

Submission Due: Friday March 5th at 11:59pm ET

Your submission will consist of three components:

1. .Rmd file (submitted as a Group)
2. .pdf file (submitted as a Group)
3. Completion of Assignment 3 - Group Work Survey (completed as an individual)

Group Submission

Your group will submit one .Rmd file that you created for this assignment AND the resulting pdf (i.e., the one you 'Knit to PDF' from your .Rmd file) must be uploaded into a Quercus assignment (link: https://q.utoronto.ca/courses/204754/assignments/561540) by 11:59PM ET, on Friday, March 5th.

Please note that only one group member needs to submit the .Rmd and .pdf files onto Quercus. We will be directly marking on the LATEST submission of the .pdf (submitted on/before the due date/time).

Late assignments are NOT accepted. Please consult the course syllabus for other inquiries.

Individual Item to Submit

In order to keep a record of group contributions there is a survey (link: https://q.utoronto.ca/courses/204754/quizzes/155783) . Please complete this survey between Thursday March 4th 12:01am ET and Monday March 8th at 11:59pm E.T. Completion of this survey will be worth 1% of this assignment. The survey should not take more than 5 minutes to complete. So NO LATE submissions will be accepted (i.e., no submissions of the survey beyond Monday March 8th at 11:59pm E.T.). There is no time limit on the survey, but you must submit it before March 8th at 11:59pm ET. You have one attempt/submission of this survey.

This survey must be completed by ALL STUDENTS (even if you are working on Assignment 3 as an individual).

Assignment grading

This is a Group assignment. All group members will receive the same grade on the Group submission for this assignment (i.e., the pdf and Rmd submission). The 1% completion point for the survey is awarded to individuals who complete the survey. If there is a lot of evidence of little to no contributions by individual group members then those individuals may receive a lower grade than their group-mates.

There are two parts to this assignment. One is theory-based and the other is data analysis and communication/writing focused. We recommend you spellcheck and proofread your written work. We will be directly marking the pdf files, thus please ensure that your final submission looks as you want it to look before submitting it. We will be marking the LATEST submission of the pdf files.

As mentioned above, this assignment will be marked based on the output in the pdf submission. You must submit both the Rmd and pdf files for this assignment to receive full marks in terms of reproducibility. Furthermore, this is a GROUP assignment. You are expected to work in a group. The workload level is higher than that of an individual assignment. Thus, it is recommended that you work in a group (i.e., not on your own).

This assignment will be graded based off the rubric available on the Assignment Quercus page (link: https://q.utoronto.ca/courses/204754/assignments/561540). TAs will look over each section (on the submitted pdf) and select the appropriate grade for that section based off a coarse overview (one-time read over) of that section. Your assignment should be well understood to the average university level student after reading it once. I would suggest you make sure your (pdf) document looks clean, aesthetically pleasing, and has been proofread. You will be able to see the rubric grade for each section. There may be some comments/feedback provided (by the TAs) if the same issue seems to be arising in multiple sections, but you will likely receive no comments/feedback (due to the scaling of the class and marking).

Part 1

A company is interested in implementing some A/B testing on its website in order to improve sales. In order to do this they first need to look at what their current website usage is. Let X_i represent the number of customers who visit the webpage in a given hour.

Assume $X_1, X_2, ..., X_n \overset{iid}{\sim} Poisson(\lambda)$.

Let's use a Bayesian approach to make some inference about λ. Use $\lambda \sim Exponential(\beta)$ as a prior distribution. Where β is the mean parameter. So $f(\lambda) = \frac{1}{\beta} e^{-\lambda/\beta}$ for $\lambda \geq 0$.

Step 1 (Mathematical Justification)

Derive the posterior distribution of λ. Identify what well-known distribution the posterior follows, and be sure to identify it's parameters. Note: the parameters of the posterior should be expressed as a function of the sample mean, sample size, β and numbers (exclusively).

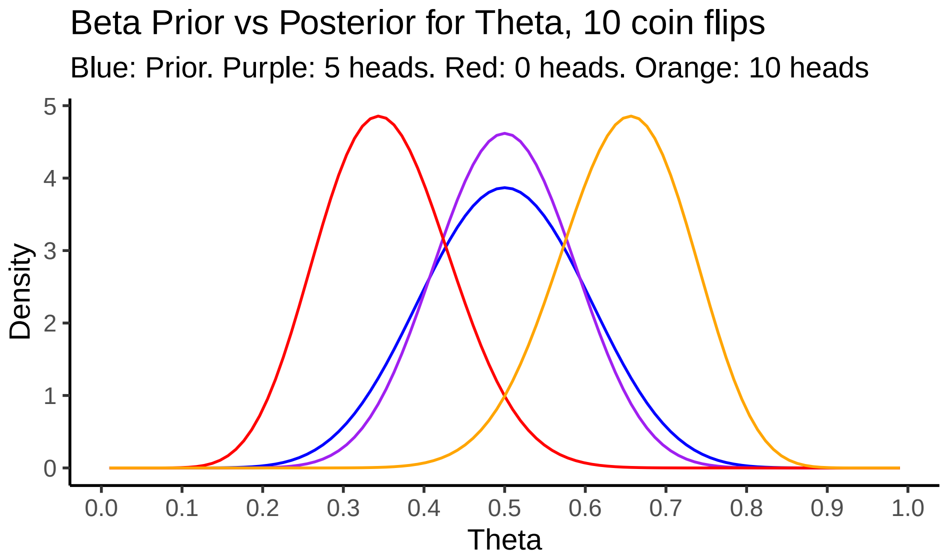
Step 2 (Simulation/Graphical Investigation of the Posterior)

Here you will comment on how the posterior distribution changes based off the prior and the data.

For the purposes of this section assume β is some fixed value (you can select whichever value you'd like - so long as it's appropriate. E.g., 0.5 or 10). Then assume you collected some data, what happens if you collected only a little data vs. a lot of data (i.e., small n vs. large n). What happens if your sample mean is close to β vs far from β?

Here you should create some type of visualization to comment on how the posterior is affected/changes based on the prior, (changes in the) sample size and (changes in) the sample mean. I would recommend plots similar to the Visualization in 11.1.2 of the Supplementary Materials. And comment on the relationship between the posterior and prior for different n and $\bar{x}$. You should have at least 2 plots, each with 4 comparisons. Each plot will include 4 curves (the prior and 3 posteriors - based on different sample means). One plot with 4 curves for small n and another plot for large n.

Here is the 11.1.2 example. The idea is to produce two plots like this (one for small n, one for large n):



Please include some text describing the relationship between posterior and the prior, sample size and sample mean. Your text should relate back to the original context of the problem (i.e., visits on a webpage). Be sure that you know what n and $\bar{X}$ represent in the context of the problem.

General Notes (for Part 1):

- This question is open book, so you can use outside sources (i.e., textbooks, academic papers, credible websites, etc.), to derive the posterior distribution. Just make sure you properly credit any outside sources.
- You will likely need to use LaTeX code in your Rmd file. Please have a look at our course Resources page, as well as the synchronous lecture in Week 4.
- Grammar is *not* the main focus of the assessment, but it is important that you communicate in a clear and professional manner. I.e., no slang or emojis should appear.
- You may want to include a bibliography in this section. If it is clear that you (or the reader) looked up something that is not common knowledge (and it was not cited) then you will lose points.
- Use inline referencing.

Part 2

Description:

In this question you will write a report on a data analysis in which your main methodology will be to derive at least two confidence intervals via bootstrapping. One bootstrap confidence interval should be for a proportion, one bootstrap confidence interval should be for a mean/median. Both bootstrap confidence intervals should be meaningful/appropriate based on the data. The report will consist of 5 sections: Introduction, Data, Methods, Results, and Conclusions.

There should be no evidence that Part 2 is an assignment, I should be able to take a screenshot of this section and paste it into a newspaper/blog. There should be no raw code. All output, tables, figures, etc. should be nicely formatted.

This will allow you to look at some interesting aspects of the data. Please find some data through either:

1. The Toronto Open Data Portal (https://open.toronto.ca/) and download it via the opendatatoronto R package (https://sharlagelfand.github.io/opendatatoronto/).
2. Some survey data from the 2019 Canadian Election Study (http://www.ces-eec.ca/) and download it via the cesR R package (https://hodgettsp.github.io/cesR/).

Feel free to use the same data as you did on Assignment 1 and/or Assignment 2, just **make sure that the data is appropriate for your methods**. Pick something that is interesting to investigate and has variables appropriate for the methodology you are going to perform. NOTE: If your data is not appropriate in performing two bootstrap CIs (one for a proportion and one for a mean/median) then you will not be eligible for full marks on this assignment. Please visit office hours, post on Piazza or email our teaching team at sta238@utoronto.ca for clarification on appropriate data.

Introduction

The goal of the Introduction section is to introduce the overall "problem" to the reader.

Your **Introduction** section should include the following:

- Describe the data and the problem in 2-3 clear sentences.
- Should introduce the importance of the analysis.
- Get the reader interested/excited about analysis.
- Provide some background/context explaining the global relevance of the problem/data/analysis.
- Introduce terminology and prep the reader for the following sections.
- Introduce hypotheses.

Data

The goal of the Data section is to introduce the reader to the data set, showcase some meaningful aspects of the data, and get them thinking about potential hypotheses/findings.

Your **Data** section should include the following:

- A description of the data collection process.
- A summary of the cleaning process (if you cleaned the data).
- A description of the important variables.
- Some appropriate numerical summaries (at minimum center and spread, but something else may be more appropriate). If there are a lot, please put them in a well formatted and labelled table.

- At least 1 aesthetically pleasing plot/graph/figure (No more than 4 plots).
- Text explaining/highlighting each table or figure.
- Some text (and perhaps graphical summaries) of the variables you will perform the bootstrap on (don't do the bootstrap here - just prep the reader for what is coming in later sections). This should help prep the reader in understanding why the CI is important/interesting and whether it is appropriate.
- In line referencing/text if needed.
- Reference the programming language/software used to complete this section.

Methods

The goal of the Methods section is to introduce the reader to the statistical methods that you will be using to analyze the data.

Your **Methods** section should include the following:

- A complete explanation of what a bootstrap sampling is.
- You should clarify if you are using Empirical or Parametric bootstrap sampling (and be sure to appropriately describe your selected method).
- An explanation of the confidence intervals which you will use to analyze the data.
- An explanation of the parameters of interest (i.e., proportion and mean/median).
- Specify all components of the bootstrap (i.e., number of iterations, type of bootstrap, etc.).
- Reference any packages/software used to complete this section.

Results

The goal of the Results section is to present the results of the bootstrap CIs to the reader.

Your **Results** section should include the following:

- The results of the two (or more) bootstrap CIs (I would recommend putting them in a table if there are a lot).
- An explanation/interpretation of the results.
- Some commentary on whether or not the results seem reasonable.
- At least two plots/tables to help describe the variables of interest, used in the bootstrap.
- Text explaining/highlighting each table or figure.
- In line referencing
- In line R code (if needed).

Conclusions

The goal of the Conclusions section is to present the story of your analysis to the reader.

Your **Conclusions** section should include the following:

- A brief recap of the hypotheses, methods, and results.
- State (or re-iterate) your key results.
- State any reasonable conclusions drawn from the results.
- An explanation/interpretation of the results.
- Some commentary on any drawbacks/limitations.
- Recommendations for Next Steps for future analyses/reports.

General Notes (for Part 2):

- It is expected that you include two CIs in your report, but you can include more (i.e., maybe you want to look at more variables).
- All tables/figures should be well labelled and clean.
- Everything in Part 2 should be written in full sentences/paragraphs.
- There should be no evidence that Part 2 is an assignment, I should be able to take a screenshot of this section and paste it into a newspaper/blog.
- There should be no raw code. Any output should be nicely formatted.
- You will also need a reference section. You should reference the data, any outside code/documentation and any ideas/concepts that are taken outside of the course.
- Note, we are not marking grammar, but we are looking for clarity. If you need help with writing there are resources posted on the Course Info>Resources page of Quercus. It is important that you communicate in a clear and professional manner. I.e., no slang or emojis should appear.
- Use full sentences.
- Be specific. Remember, you are selecting this data and the reader/marker may not be familiar with it. A good principle is to assume that your audience is not aware of the subject matter.
- Remember to end with a conclusion. This means reiterating the key points from your writing.