Mini-portfolio instructions

STA303/1002, Winter 22

Contents

Information	**Note**
Name	Mini-portfolio
Type (Main, Mini or Basket)	Mini
Value	5% (Path A) 0% (Path B)
Due	Thursday, February 3, 2022 at 3:03 p.m. ET
Submission instruction	Submission: Via Markus
Accommodations and extension policy	In the case of a personal illness/emergency, a declaration can be made, but must be submitted no more than 3 days after the due date. Extensions may be requested through the same form up to 48 hours before the due date.

Portfolio assessments aim to help you demonstrate your technical coding, statistical thinking, communication and reflection skills. This mini-portfolio also aims to recap and refresh knowledge from your previous statistics courses as well as building your ability to create quality data visualizations.

General instructions

- Be very careful to follow instructions on variable naming. If you do not, your code won't pass auto-grading and you will not receive the grades. This will not be eligible for regrading requests.
- Comment your code! In an R code chunk comments start with a # (pound sign or hastag symbol). Don't confuse this with the use of # to denote different levels of headings in the text parts (governed by Markdown syntax) of an R Markdown document.
- You should neatly format your code. No strict style is required, but make sure it is easy to read your work.
- Include your code in the body of the PDF itself (don't set echo=FALSE, don't hide/suppress etc.). Note that this is different to what you will be asked to do in the final project or in professional reporting. This is a demonstration of your skills.
- KNIT EARLY AND OFTEN! Don't leave things till the last minute, your Rmd not knitting is not an emergency for which an extension will be granted.

Template

You can access the template for this assessment here.

There is currently a lot of 'filler text' and 'filler code' in the template that you will want to **delete.** Fun fact: All filler text sourced from Hipster Ipsum, which Katy Wang in the UG Stats office introduced me to.

Submission instructions

- Submit both your Rmd (must be called: sta303-w22-mini-portfolio.Rmd) and PDF (must be called: sta303-w22-mini-portfolio.pdf) on MarkUs.
- You do not need to submit any data or tex files.

Cover page

You don't have to use the provided template, but you DO need to write your mini-portfolio in RMarkdown and include a cover page.

The cover page must have:

- A title and subtitle (you can use my examples in the template or update them as you see fit, no points)
- Your name
- Date (assessment submission date is fine)

In the template, you can change the colour of this cover to any colour you would like by replacing 6C3082 in the YAML (line 11) to another hex code. You could use this tool to help you: https://htmlcolorcodes.com/color-picker/

Introduction

Write this section last or second to last (before the refection).

In the introduction section, write a brief summary of the skills you have demonstrated in this mini-portfolio, across the statistical skills sample, writing sample and reflection sections. Think of it like a **cover letter** for this document. It should be appropriate for a fairly general audience—imagine a future employer reading this. You may want to breifly explain the course context, as you understand it. What is STA303/1002 about? (Consider the learning objectives in the syllabus)

Your introduction should be **not be longer than 300 words** and must **fit on one page**.

Statistical skills sample

Task 1: Setting up libraries

Set up a chunk called `setup` where you load the `tidyverse` and `readxl` libraries. Set your chunk options to `message=FALSE` so all the package loading information isn't included in your our output. You will need to make sure you run this chunk each time you start a new session so you can sue many of the functions required.

Task 2: Visualizing the variance of a Binomial random variable for varying proportions

Show visually that for a fixed value of n, $p = 0.5$ will result in the largest variance for a Binomial random variable.

- Choose two appropriate values of $n > 0$ for your demonstration and save them as `n1` and `n2`
- Create a vector of proportions, `props`, from 0 to 1, in steps of 0.01 (Tip: use the `seq()` function).
  - If I suggest a function you haven't seen before, you can search its documentation in your console by typing a ? in front of the function name, e.g. ?seq.
- Create a tibble (a data type in R, like a dataframe), `for_plot`, with the vector of `props` as the first variable, and two additional variables calculating the variance for each of your two chosen n values. (Call these `n1_var` and `n2_var`)
- Create **two** plots, one for each of your values of n, using `ggplot` and apply `theme_minimal()` to each one. These should appear in your PDF and do not need to be saved with specific name.
  - Add an appropriate **figure caption** to each chart (use fig.cap="Your text here" in the R chunk settings.
    - An appropriate title should succinctly explain the chart and mention the chosen n value.
  - Add a **caption** *within* the ggplot that says "Created by STUDENT NAME in STA303/1002, Winter 2022".
  - Give the x and y labels appropriate **labels**.

Objects that must be carefully assigned to pass autograding

- `n1` and `n2` (both should be integer vectors of length 1)
- `props`, a numeric vector
- `for_plot`, a tibble with three correctly named columns

Task 3: Demonstrating frequentist confidence intervals as long-run probabilities of capturing a population parameter

Goal: Simulate a population of size 1000, using $N(10, 2)$, and take 100 independent, random samples of size 30 observations each from it. Calculate a Wald confidence interval (using am appropriate t-multiplier) for the population mean from each sample. Calculate what proportion of intervals contain the population mean and plot all these intervals, coloured by successful population mean capture or not.

Specific steps

- Set the seed to the last three digits of your student ID (this is your numeric student identifier number, NOT your UTORID).
- Set up the following objects with the appropriate simulation parameters, sample size and number of samples.
  - `sim_mean` and `sim_sd`
  - `sample_size`
  - `number_of_samples`
- Calculate the appropriate t-multiplier (function: `qt(...)`) for constructing a 95% confidence interval in this context. Make sure your degrees of freedom are appropriate. Save it as `tmult` for later use.
- Create a vector called `population`, a simulated population using `sim_mean` and `sim_sd` and with 1000 values (function to generate random numbers from a normal distribution: `rnorm(...)`).
- Find the *actual* true mean for your population and save it as `pop_param`. This should be a numeric vector of length 1.
- Get 100 samples of size 30 from your population and save them in a vector called `sample_set`. (This might be a little tricky/unfamiliar, so here is one way to do it. You can just copy and paste this code.)

```
sample_set <- unlist(lapply(1:number_of_samples,
function (x) sample(population, size = sample_size)))
```

- Create a new vector called `group_id` that will allow you to label the values from the 100 different samples above. Hint: `rep(...)` will be useful here, and has a great little argument called `each`. Take a look at the documentation (`?rep`) to compare the behaviour of `times` and `each`.
- Create a new `tibble` (a data type in R, like a dataframe), `my_sim`, that has two columns:`group_id` and `sample_set.`
- Create a new tibble, `ci_vals`, that starts with the dataset `my_sim` and then groups by `group_id`, and summarizes appropriately to create two new columns: `mean` and `sd`, that hold the means and standard deviations. There should be one row per group.
- Continue to change the tibble `ci_vals` by adding the following variables:
  - `lower` and `upper`, two columns that hold the lower and upper bound of a 95% confidence interval for the group. You will need to calculate this. Consider the equations for a confidence interval and remember that we are using a t-multiplier that you have already calculated.
  - `capture` which takes the values `TRUE` if the population parameter is in the 95% CI, and `FALSE` if not. These should be logical NOT character types.

- Create an object called `proportion_capture` that uses `ci_vals` and stores the proportion of intervals you created that capture the population parameter. This should be done using the object names, not 'hard coded', so that if you changed your set.seed or your sample size, etc., and run all the code again, it would update this value. It should be a vector of length 1.
- Plot these 100 confidence intervals in one plot, with the means indicated as points, as well as and a dotted line for the population parameter. `geom_errorbar()` will be very helpful.
  - **Colour** the confidence intervals by whether or not they contain the population parameter.
    - If the interval include the population parameter, colour it #122451 if it DOES contain the parameter (TRUE) and #B80000 if it DOES NOT (FALSE).
  - Set the **figure caption** to "Exploring our long-run 'confidence' in confidence intervals. This figure shows how often 95% confidence intervals from 100 simple random samples capture the population mean. The population was simulated from N(10, 2)".
  - Set the caption to "Created by STUDENT NAME in STA303/1002, Winter 2022". Replace STUDENT NAME with your name.
  - Set the **legend title** to "CI captures population parameter".
  - **Flip the coordinates** using `coord_flip()` so the intervals are horizontal across your chart.
- Add the following sentence to your Markdown (not in a code chunk):
  - 'r proportion_capture*100' % of my intervals capture the the population parameter
  - The parts in the backticks will be processed as inline R code.
- Briefly (1–2 sentences) describe why we can include the population parameter in this plot AND why we cannot usually compare the population parameter to our confidence interval in practice (e.g., when working with data that has not been simulated). Write this for a non-statistical audience.

Objects that must be carefully assigned to pass autograding

- `sim_mean` and `sim_sd`
- `sample_size`
- `number_of_samples`
- `tmult`
- `population`
- `pop_param`
- `group_id`
- `sample_set`
- `my_sim` , tibble with two columns
- `ci_vals`, tibble with 5 columns (after all steps completed)
- `proportion_capture`

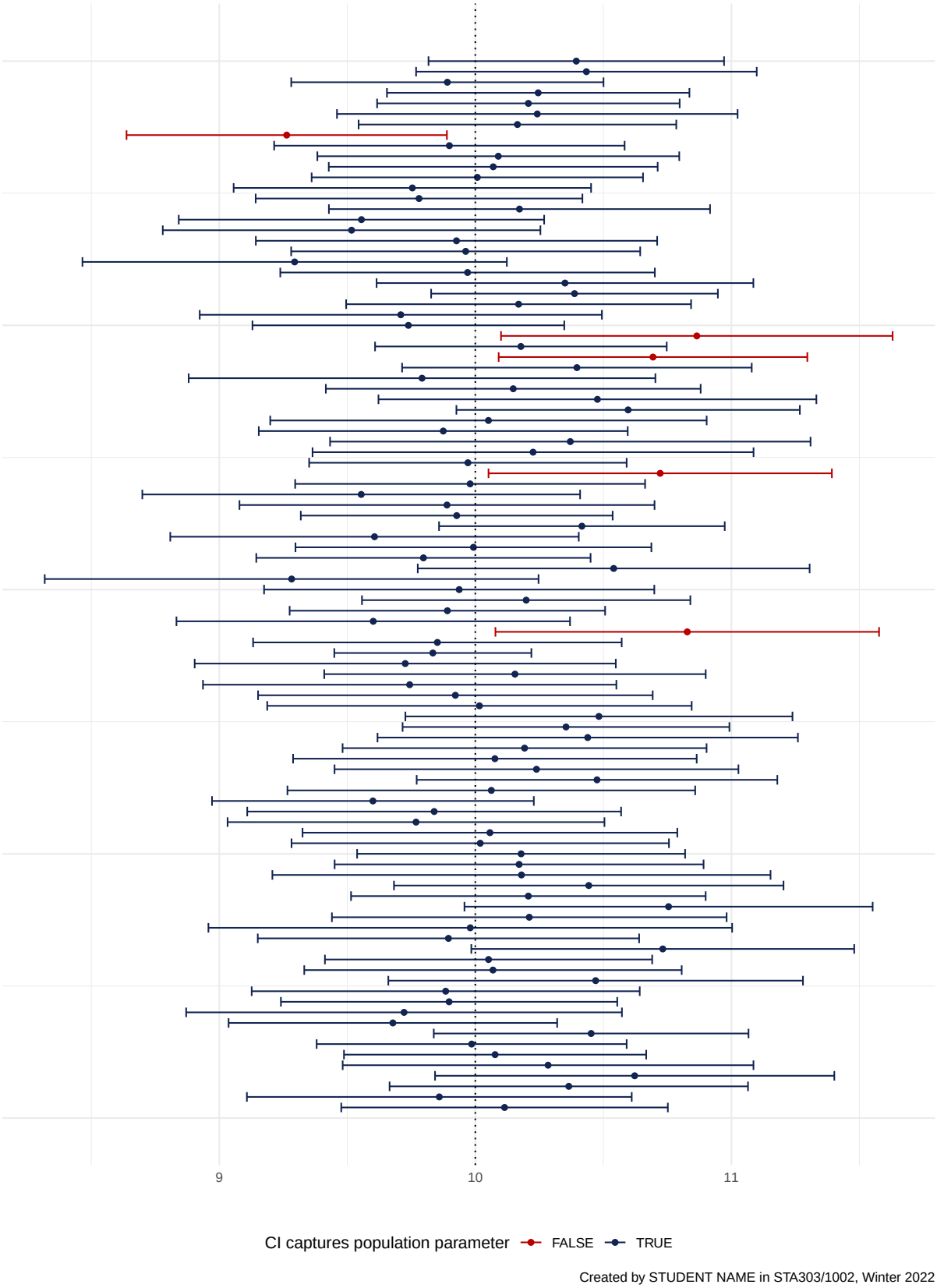

Figure 1: This is generally what your output should look like for the confidence interval task. Note: It won't be exactly the same.

Task 4: Investigating whether there is an association between cGPA and STA303/1002 students correctly answering a question on global poverty rates

In the 'getting to know you' survey at the beginning of STA303, students who participated in the survey were asked:

- What their current cumulative grade point average (CGPA) was at U of T.
- Whether the proportion of people living below the global poverty line had halved, doubled or stayed about the same in the last 20 years.

So, what was the correct answer to the question about global poverty? The proportion of the global population living below the poverty line has HALVED!

In the .xlsx file called `sta303-mini-portfolio-poverty.xlsx` there are 200 observations that represent the patterns in our class but from which is is not possible to identify individual students. The final goal of this task is to test whether there is a difference in cGPA between students who correctly answered this question and those who di not.

Goal

1. Briefly describe the goal of this task to someone who has not read these instructions. **Put it in your own words.**

Wrangling the data

2. Load the data, `sta303-mini-portfolio-poverty.xlsx`, into an object called `cgpa_data`, and apply the `clean_names()` function from the `janitor` package. Pay attention to the path to the file when providing it to R. It won't know to look in the data folder if you don't tell it.

3. Rename the cGPA variable to `cgpa` and the poverty question answer to `global_poverty_ans`.

4. Clean the data so only appropriate cGPA variables are included.

5. Create a new variable called `correct` that takes the value TRUE (logical, not character) if the respondent answered 'Halved' and FALSE if they answered 'Doubled' or "stayed about the same'.

All of the above changes should be saved into the dataset called `cgpa_data`. I recommend using pipes `%>%`. There is a keyboard shortcut too: Cmd+Shift+M (Mac) or Ctrl+Shift+M (Windows).

Visualizing the data

6. Create a set of histograms, one on top of the other, the will let you examine the data in a useful way.

Testing

1. Choose an appropriate test to test whether there is an association between cGPA and if a student in STA303/1002 answered this question correctly. JUSTIFY your choice appropriately.

2. Conduct the test AND the equivalent version using `lm()` **interpret the result of the test appropriately** in a sentence or two.

Recall the types of tests you have now encountered, in this and previous courses:

- T-test `t.test(x, data = my_data)` or `t.test(x~y, data = my_data)`
- ANOVA ‘summary(aov(x~y, data = my_data))
- Wilcoxon test `wilcox.test(x, data = my_data)`
- Mann-Whitney U `wilcox.test(x~y, data = my_data)` (note different use of wilcox function)
- Kruskal-Wallis Rank Sum test `kruskal.test(x~y, data = my_data)`

Make sure you're considering the assumptions for the tests as well as some of the other criteria about which tests we choose to reach for. You may need ask some questions in office hours or find relevant readings to help you consider the non-parametric tests. Your strategies here might be something to consider in your reflection.

Writing sample

Prompt

Read the below job ad and write about the skills you would need to apply. Write on at least 2 soft skills and 2 analytic skills. The ad targets those with a MSc/PhD or those with 2+ years of experience, but imagine this requirement is not there.

1. **Soft skills**. What soft skills relating to communicating and working with others does the company seek?
    1. In what way do you already possess two of these skills?
    2. What evidence do you have of possessing one or two of these skills?
2. **Analytic skills.** What analytic skills relating to software use and performing data analysis does the company seek?
    1. In what way do you already possess two of these skills?
    2. What evidence do you have of possessing one or two of these skills?
3. **Connection to studies.** What other skills can you develop and what evidence can you accumulate during the remainder of your education to be ready for a similar job ad?

Structure your answer under three headings: 'Soft skills', 'Analytics skills' and 'Connection to studies'. Write your answers in full sentences, with appropriate paragraphing. There should be a brief introduction and conclusion. Imagine a future employer or graduate school admission officer was reading this, you should explain what you are going to do, do it (the three headings) and then sum up what you did.

Word count: 300–500 words. **Please add a statement of your word count at the end of the passage.**

Figure 2: Yelp logo

Job add

Data Scientist (Remote) **Category** Data Science & Analytics

Location Toronto, Ontario, Canada

Department Engineering and Product

At Yelp, it's our mission to connect people with great local businesses. Yelp's unique dataset contains billions of interactions between users and local business around the globe, from a user reviewing a neighborhood

coffee shop to requesting a repair quote with a photo of a leaky faucet. Data Scientists at Yelp work to make sense of these interactions to deliver impactful analyses and products to our users, business partners and the general public.

The Data Science team performs analyses, builds models, and designs experiments that directly impact Yelp's business and users. Our centralized team is the most wide-ranging consumer of data at Yelp, adept at tasks from modeling content growth and user behavior to sharing insights about the health of local economies. With varied backgrounds and expertise, we strive for learning and growth in a collaborative environment.

We'd love to have you apply, even if you don't feel you meet every single requirement in this posting. At Yelp, we're looking for great people, not just those who simply check off all the boxes.

This opportunity is fully remote and does not require you to be located in any particular region. We welcome applicants from throughout Canada.

We Are Looking For:

- 3+ years of experience as a data scientist or MS/PhD and 2+ years of industry experience in a quantitative role.
- Fluency with SQL and Python or R for data analysis.
- Solid understanding of statistical inference, experimental design and analysis.
- Enthusiasm for clean code and sharing reproducible results.
- Communication skills to work with partners on engineering, product and business teams.
- An eye for great data visualization with Matplotlib, Plotly, ggplot, or Tableau.
- If you don't have 2+ years of industry experience in a quantitative role, please take a look at our College Data Scientist roles instead!

Where You Come In:

- Define key metrics to track Yelp's performance and inform product decisions.
- Assess and frame questions from partners into actionable deliverables.
- Design, execute, and analyze complex experiments impacting millions of users.
- Devise and evaluate models for diverse business needs, such as identifying growth opportunities, personalizing user experience, and matching consumers to businesses.
- Own analyses start-to-finish and communicate key insights to stakeholders.
- Share your technical skills to develop and maintain high-quality, reusable analysis tools.

LI-Remote

At Yelp, we believe that diversity is an expression of all the unique characteristics that make us human: race, age, sexual orientation, gender identity, religion, disability, and education — and those are just a few. We recognize that diverse backgrounds and perspectives strengthen our teams and our product. The foundation of our diversity efforts are closely tied to our core values, which include "Playing Well With Others" and "Authenticity."

We're proud to be an equal opportunity employer and consider qualified applicants without regard to race, color, religion, sex, national origin, ancestry, age, genetic information, sexual orientation, gender identity, marital or family status, veteran status, medical condition or disability.

We are committed to providing reasonable accommodations for individuals with disabilities in our job application process. If you need assistance or an accommodation due to a disability, you may contact us at accommodations-recruiting@yelp.com or 415-969-8488.

Note: Yelp does not accept agency resumes. Please do not forward resumes to any recruiting alias or employee. Yelp is not responsible for any fees related to unsolicited resumes.

Accessed from **Yelp®** on Feb 28, 2021, 15:15 PM

Large logo from Wikipedia https://en.wikipedia.org/wiki/Yelp#/media/File:Yelp_Logo.svg

Reflection

Briefly, 100 to 200 words each, answer the following questions:

- What is something specific I'm proud of in this mini-portfolio?
- How might I apply what I've learned and demonstrated in this mini-portfolio in future work and study, after STA303/1002?
- What is something I'd do differently next time?