STA303/1002 Mini-portfolio

An exploration of data wrangling, visualization, hypothesis testing and writing skills

Jiahao(Green) Bai

2022-02-03

Contents

List of Figures

Introduction

This mini-portfolio is a course project of STA303 with the usage of R.
STA303 is a statistic course at the University of Toronto. Its concepts include but are not limited to the following: wrangle and explore data, understand commonly used statistic models, data visualization, make proper conclusions based on data analysis, and present corresponding results to a range of audiences.

This mini-portfolio breaks into three sections: statistical skills sample, writing sample, and reflection.

The first part: statistical skill sample, mainly focused on data visualization, confidence interval, and statistical test selection. There are three primary tasks:
1.Visualizing the variance of a Binomial random variable for varying proportions
2.Demonstrating frequentist confidence intervals as long-run probabilities of capturing a population parameter
3.Investigating whether there is an association between cGPA and STA303/1002 students correctly answering a question on global poverty rates

Package tidyverse is included, functions of this package like ggplot(), geom_point() are heavily used to make different kinds of graphs. For example, discrete points plot for variances and confidence interval graph with confidence intervals that do not contain the population mean labeled in red. Also, the third task above involves checking assumptions for various statistical tests.

The other two sections focused on the ability to collect and summarize data. The writing sample section provides a job ad from Yelp. It requires finding soft skills (skills related to communicating and working with others) and analytic skills (skills related software use and performing data analysis) from the ad's content, two for each. The reflection section requires to go over the whole mini-portfolio again and write about good things done and something that could do differently for this project. Also, the application of things learned in future work and study.

Statistical skills sample

Setting up libraries

```
# Set up libraries
library(tidyverse)
library(readxl)
```

Visualizing the variance of a Binomial random variable for varying proportions

```
# Set up n1 and n2
n1 = 1
n2 = 50

# Set up proportions and create a tibble of it along with variances by using n1
# and n2.
props <- seq(0, 1, 0.01)
for_plot <- tibble(props, n1_var = n1 * props * (1 - props),
                          n2_var = n2 * props * (1 - props))
```

```
# Plot the proportions vs variance graphs
for_plot %>% ggplot(aes(x = props, y = n1_var)) + geom_point()+theme_minimal()+
         labs(caption = "Created by Jiahao Bai in STA303/1002,Winter 2022") +
         xlab("Proportions") +
         ylab("Variance with n = 1")
```

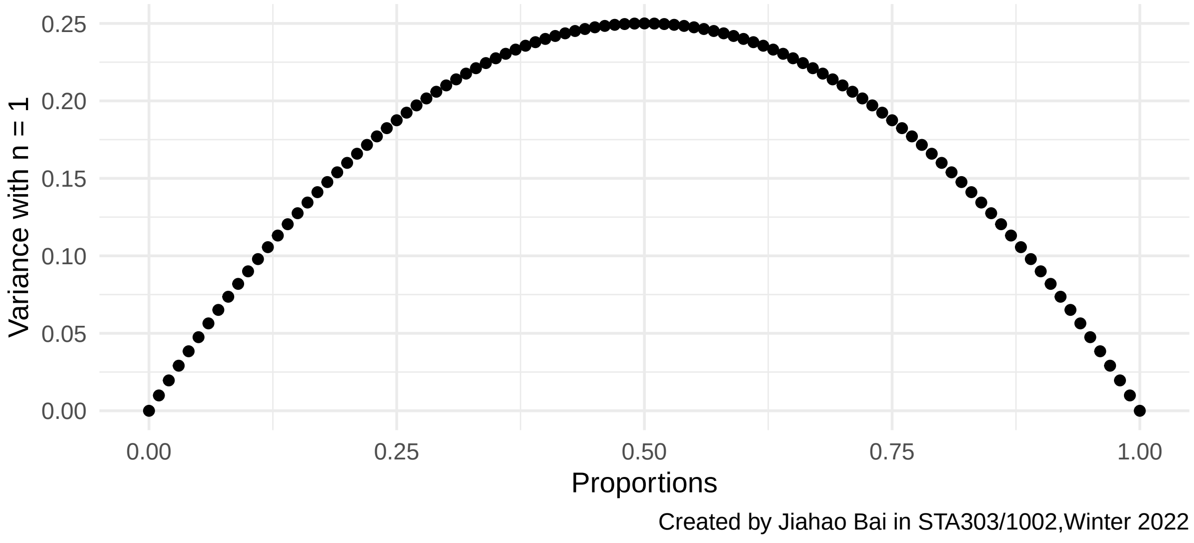

Figure 1: Variance vs Proportions Graph of a Binomial random variable with n1 = 1

```
# Plot the proportions vs variance graphs
for_plot %>% ggplot(aes(x = props, y = n2_var)) + geom_point()+theme_minimal()+
         labs(caption = "Created by Jiahao Bai in STA303/1002,Winter 2022") +
         xlab("Proportions") +
         ylab("Variance with n = 50")
```

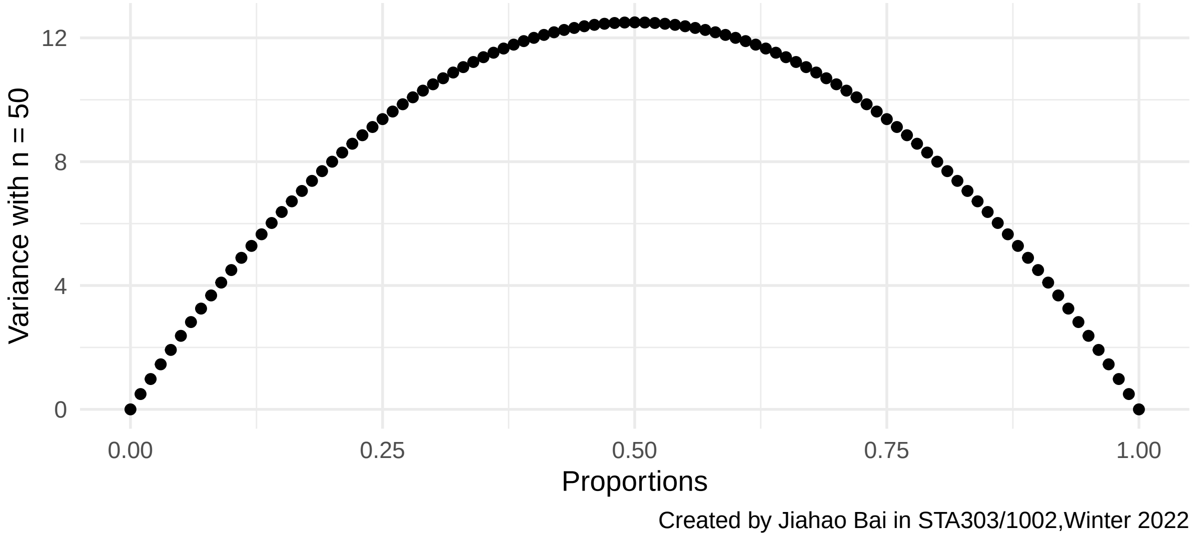

Figure 2: Variance vs Proportions Graph of a Binomial random variable with n2 = 50

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

```r
# Set seed by using last three digits of student number
set.seed(097)

# Initialize statistics: mean, variance, sample size, and number of samples
sim_mean = 10
sim_sd = sqrt(2)
sample_size = 30
number_of_samples = 100

# Calculate the t-multiplier (95% confidence interval)
tmult = qt(0.975, df = sample_size - 1)

# Generate/Simulate the population with sample size = 1000
population = rnorm(1000, sim_mean, sim_sd)

# Get actual true mean of population
pop_param = mean(population)

# Get 100 samples of size 30 from the population
sample_set <- unlist(lapply(1:number_of_samples,
      function (x) sample(population, size = sample_size)))

# Label the values from the 100 different samples above
group_id = rep(1:100, each = 30)

# Create a tibble with two cols: group_id and sample_set
my_sim = tibble(group_id, sample_set)

# Create a tibble with group_id, mean, and standard deviation, then add lower
# upper confidence intervals, and capture with T/F values.

ci_vals <- my_sim %>% group_by(group_id) %>%
           summarise(mean = mean(sample_set), sd = sd(sample_set)) %>%
           mutate(lower = mean - tmult * sd / sqrt(sample_size),
                  upper = mean + tmult * sd / sqrt(sample_size),
                  capture = (pop_param <= upper) & (pop_param >= lower))

# Calculate proportion of intervals that capture the population parameter
```

```r
proportion_capture = mean(ci_vals$capture)

# Plot the confidence intervals

ci_vals %>% ggplot(aes(x = group_id, y = mean, color = capture)) +
            geom_point() +
            scale_color_manual(values = c("#B80000", "#122451")) +
            geom_errorbar(aes(ymin = lower, ymax = upper)) +
            geom_hline(yintercept = pop_param, linetype = "dotted") +
            labs(caption = "Created by Jiahao Bai in STA303/1002, Winter 2022",
                 color = "CI captures population parameter") +
            coord_flip() + theme_minimal() +
            theme(legend.position = "bottom", axis.title.x = element_blank(),
                                              axis.title.y = element_blank(),
                                              axis.text.y = element_blank())
```

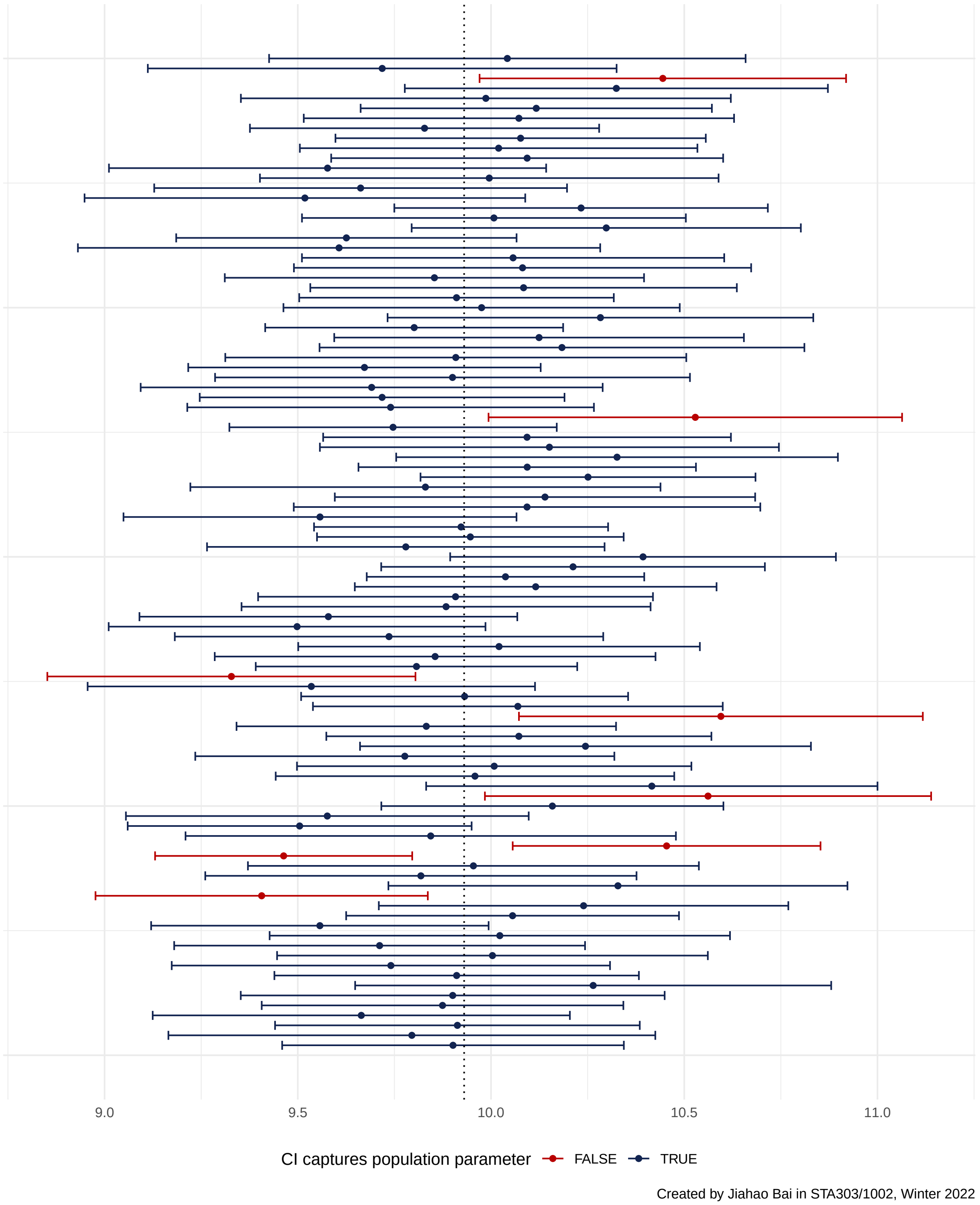

Figure 3: Exploring our long-run "confidence" in confidence intervals. This figure shows how often 95% confidence intervals from 100 simple random samples capture the population mean. The population was simulated from N(10, 2)

92% of my intervals capture the the population parameter.

Real-world data is often more complicated and unpredictable. We could include the population parameter in the plot for this mini-portfolio because we only have a population of size 1000. 1000 is a small population if we compare it to real-life situations, so it's possible to calculate the mean, which is the average of the data we are interested in. In practice, real-world's population is much more greater. It's really hard to gather the data of population and calculate its average. For example, it's almost impossible to collect weight of everybody on Earth and calculate its average. Therefore, it's this particular special setting of population of this task allows us to present further more information (graph) of confidence interval.

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

Goal

The goal of this task is to study the relationship between correctly answering the global poverty question and STA303/1002 students' cGPA.

Wrangling the data

```
# Import data
path = "data/sta303-mini-portfolio-poverty.xlsx"
cgpa_data = read_xlsx(path)

# Rename the cGPA variable to cgpa and the poverty question answer to
# global_poverty_ans. Then, clean the data and create variable correct
cgpa_data <- cgpa_data %>% janitor::clean_names() %>%
             rename(cgpa = 3, global_poverty_ans = 2) %>%
             filter(!is.na(cgpa) & cgpa > 0 & cgpa <= 4) %>%
             mutate(correct = (global_poverty_ans == "Halved"))
```

Visualizing the data

```
# Plot the histograms of cga_data with two groups of correct specified: true and false
cgpa_data %>% ggplot(aes(x = cgpa, fill = correct)) +
              geom_histogram(binwidth = 0.2) +
              facet_wrap(~correct, nrow = 2, ncol = 1) +
              labs(caption = "Created by Jiahao Bai in STA303/1002,Winter 2022")
```

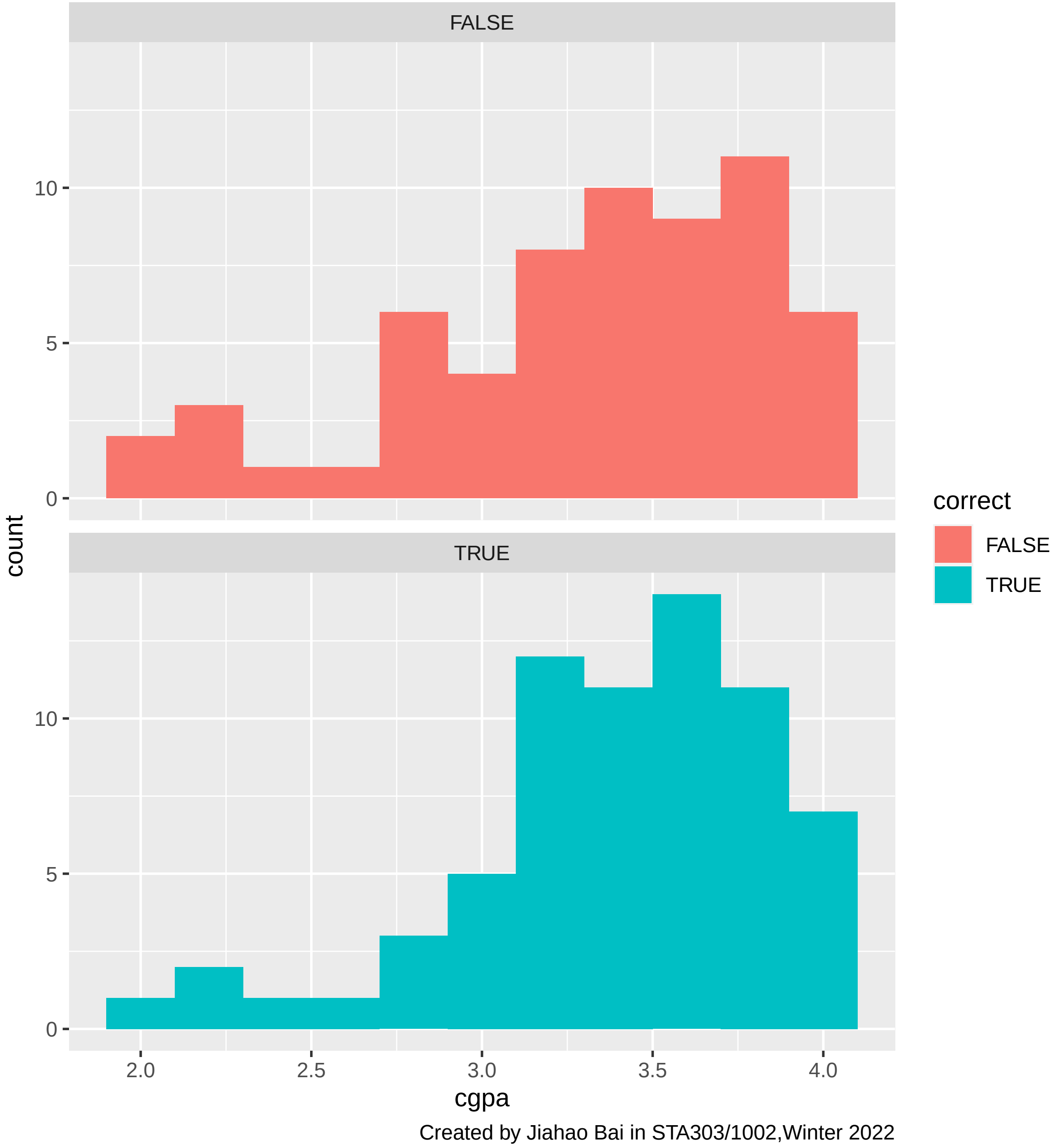

Figure 4: cGPA vs Counts of Correctness of Answering Poverty Question

Testing

I chose to test whether there is an association between cGPA and if a student in STA303/1002 answered this question correctly with the Mann-Whitney U test. Since the data is independent and variable: correct's value has already been classified into two groups: True and False, so two-sampled t-test (parametric) and Mann-Whitney U test (non-parametric) should be considered. However, the standard deviations of these two groups are different. Approximately the standard deviation of the False group is 0.5129, and that of the True group is 0.4400. Also, the QQ plot of the linear model between cGPA and counts of correctness (Value: True or False) shows a parabola instead of a line. Refer to the histograms drew in last section, they are left skewed which proves the data is not normal as well. Therefore, the normality and constant variance assumptions are violated. Since the Mann-Whitney U test allows ignoring the distribution and constant variance, so it's used.

```
# Apply Mann-Whitney U test
wilcox.test(cgpa~correct, data = cgpa_data)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  cgpa by correct
## W = 1875.5, p-value = 0.35
## alternative hypothesis: true location shift is not equal to 0
```

```
# Apply the linear model which does the same thing with Mann-Whitney U test
summary(lm(rank(cgpa)~correct, data = cgpa_data))
```

```
##
## Call:
## lm(formula = rank(cgpa) ~ correct, data = cgpa_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -64.919 -30.419  -1.419  33.754  65.754
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   61.746      4.786  12.902   <2e-16 ***
## correctTRUE    6.173      6.592   0.937    0.351
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37.38 on 127 degrees of freedom
## Multiple R-squared:  0.006859,   Adjusted R-squared:  -0.0009611
## F-statistic: 0.8771 on 1 and 127 DF,  p-value: 0.3508
```

Based on the p-values of the linear model and Mann-Whitney U test: 0.3508 and 0.35, respectively. The two p-values meet. The conclusion is: we have no evidence that STA303/1002 students' cGPA would affect the correctness of answering the global poverty question. The "correct" variable is insignificant in the linear model, and 0.35 indicates no significant difference between the two groups (True and False).

Writing sample

Introduction

I fell in love with statistics. Not only because the statistic is fascinating, but it's also a way to understand this complicated world. Building relevant skills are also essential to accomplish the dream of interpreting the world. Through Yelp's job advertisement, there are two kinds of skills that need attention: soft skills for communicating and analysis skills for analyzing data.

Soft skills

To begin with, one of the soft I have is the passion for sharing reproducible results and clean coding. As a programmer, sometimes real people would assess my works. Whenever I'm writing codes, I always try to explain what I just wrote to myself. In practice, in my mini-portfolio project of STA303 (a statistic course), I comment on the significant step of my R codes and format the lines in a parallel pattern so that no line is lengthy. Thus, whoever checks my work would easily track my codes and reproduce them to get the same statistics. The other soft skill I have is the communications skills for working with other departments. This skill is connected with teamwork which is the hardcore of a company. I was involved in a Java(a programming language) month-long group project. I was responsible for back-end coding, but I understood the terms related to front-end coding. Thus, I was able to exchange ideas between two departments. Then, we finished the assignment in 4 days. Also, the feedback of this project is very pleasing: 98/100.

Analytic skills

To finish the work assigned, analytic skills are required as well. I'm familiar with data analysis and data visualization skills. The former is related to Python, R, and SQL. I have experience in regressions and machine learning algorithms, such as linear regression with R and K-mean clustering with Python. Also, I've completed several projects with these two topics and got good marks. For example, I used linear regression to determine are possible reasons for high blood pressure in STA302 (a university course I've taken). Moving to data visualization skills, I frequently use ggplot in R to create tables and graphs to demonstrate my data. In the mini-portfolio I mentioned in the soft skills section, I used ggplot to plot histograms of two different data groups. There were two different colours with a legend and one on top of the other to make comparison easy.

Connection to studies

I still need to practice my experimental design skills because I haven't studied this so far. Good experimental design skills would allow me to collect data more effectively, and I would create some surveys and present some reports based on them. These surveys and corresponding reports would be a great experience to put on my resume.

Conclusion

Overall, soft skills like clean coding, sharing reproducible works, communication skills between different departments, and analytic skills like data analysis and visualization programming languages are critical skills for the data scientist job, and I'm already equipped with these skills. However, there is still more space for me to improve.

Word count: 489 words

Reflection

What is something specific that I am proud of in this mini-portfolio?

I didn't have a chance to study STA130, but Liza provided me with this mini-portfolio project to catch up and make some applications. Firstly, I'm very proud that I could finally have a place to demonstrate newly learned R coding skills, for example, the function I used to plot my statistics/data: ggplot(). These graphing functions give me a better chance to gather information virtually and perform my studies on the data. I successfully reproduced the beautiful confidence intervals graph, which makes me also feel proud. Reproducibility is a critical concept in the statistic field. I'm glad that I could figure out the functions that needed to be used and explore them a little by reading their supporting documents. Lastly, I'm very proud that I found out why we actually can't use the two-sample t-test on task 4. I often forget about checking assumptions, but I dug into the data and realized that normality and constant variance assumptions are violated this time. Overall, these are the three things that 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?

Data cleaning and data visualization are the topics never get outdated in the statistic field. I learned data cleaning techniques from this mini-portfolio, such as using the filter() function to obtain the data in the dataset with specific conditions, which is handy in practice. In future work or studies, the data collected from real life is messy or contains too much extra information at first most of the time. Therefore, this filter() function is an excellent choice for picking the data I am interested in. Also, the function rename() could let me rename variable names, which are lengthy and unintuitive. Then, I could better work with shorter and meaningful variable names. Moreover, I could also deploy the data visualization techniques I used on this mini-portfolio in future work. If the results need to be compared in my future career, I could use functions like facet_warp() to demonstrate my data in a parallel format. If there are some unexpected values among all results, I could do the same for confidence intervals that do not enclose the population mean in task 3, which is to use an eye-catching label. Overall, this project helps me build a solid foundation for upcoming data analysis.

What is something I'd do differently next time?

We should always maintain the passion for exploring different things, same for this mini-portfolio. The first thing I would do next time is to change the colour of the cover page. I kept it purple for this assignment because I was unfamiliar with the beginning section of the markdown files, where

I could make many adjustments here. Therefore, I would study it for the upcoming weekend and apply what I researched for the next assignment. The second thing I would try for a different approach is the labelling of the graphs. For task 2 of this portfolio, I only did the minimal work the portfolio instruction wanted us to do. Therefore, I think there should be more I could do to make the graph better looking and contain more information, for example, colour the highest variance.