STA303/1002 Mini-portfolio

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

Shiqi Li

2022-02-03

Contents

List of Figures

Introduction

This mini-portfolio is written to exercise and demonstrate the skills I learned from the course STA303 which includes both statistical skills and presentation skills.

The first section concerns mainly exploratory data wrangling and visualization skills that is separated into 4 subparts. In the first subpart, I will present how to set up libraries for a data analysis project. The second subpart is where I create two graphs to present the relationship between the value of the proportion parameter and the variance of a binomial random variable. The third part I performed data simulation and sampling to visually present the statistical coverage of confidence intervals. The last part studies a data set from an in-class survey by concluding on the association between the two variables using a non-parametric test and a linear regression model.

The second section presents my profession writing skills by writing a cover letter sample based on a given job description. Specifically, I discussed how I am competent in terms of both soft and technical skills by relating to my experience as a university student.

At the end of the portfolio, I also presented how I view this mini-portfolio and how can I improve upon the skills demonstrated. I also discuss what are the possible next steps for future portfolio writing.

Statistical skills sample

Setting up libraries

```
# read the required package tidyverse and readxl
library(tidyverse)
library(readxl)
```

Visualizing the variance of a Binomial random variable for varying proportions

```
library(dplyr) # first load dplyr for tibble use
n1 = 100
n2 = 500 # simulated sample sizes
props = seq(0, 1, 0.01) # vector of proportion
for_plot = tibble(props=props, # 3 columns with p, var with n1, var with n2
                  n1_var=n1*props*(1-props), n2_var=n2*props*(1-props))
ggplot(for_plot,
       aes(x=props, y=n1_var)) +
  geom_line() + # line chart
  labs( # axis labels, title, and caption
    caption="Created by Shiqi Li in STA303, Winter 2022",
    x="Proportion",
    y="Variance"
    ) +
  theme_minimal(base_size=14,
                base_family="")
```

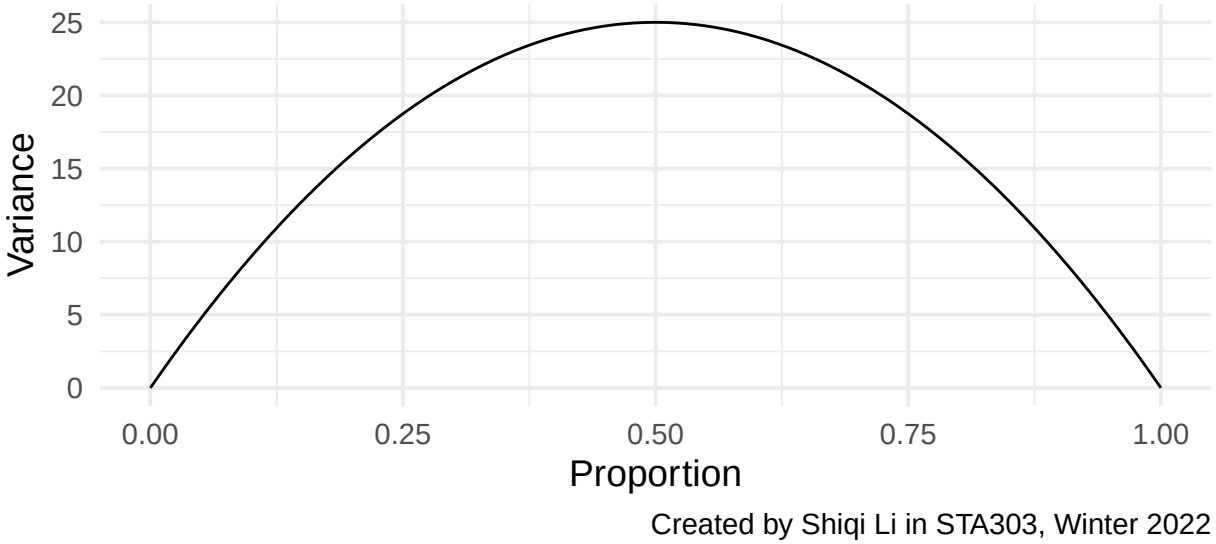


Figure 1: Binomial Variance vs Proportion Plot with n = 100

```
ggplot(for_plot,
       aes(x=props, y=n2_var)) +
  geom_line() + # line chart
  labs( # axis labels, title, and caption
    caption="Created by Shiqi Li in STA303, Winter 2022",
    x="Proportion",
    y="Variance"
    ) +
  theme_minimal(base_size=14,
                base_family="")
```

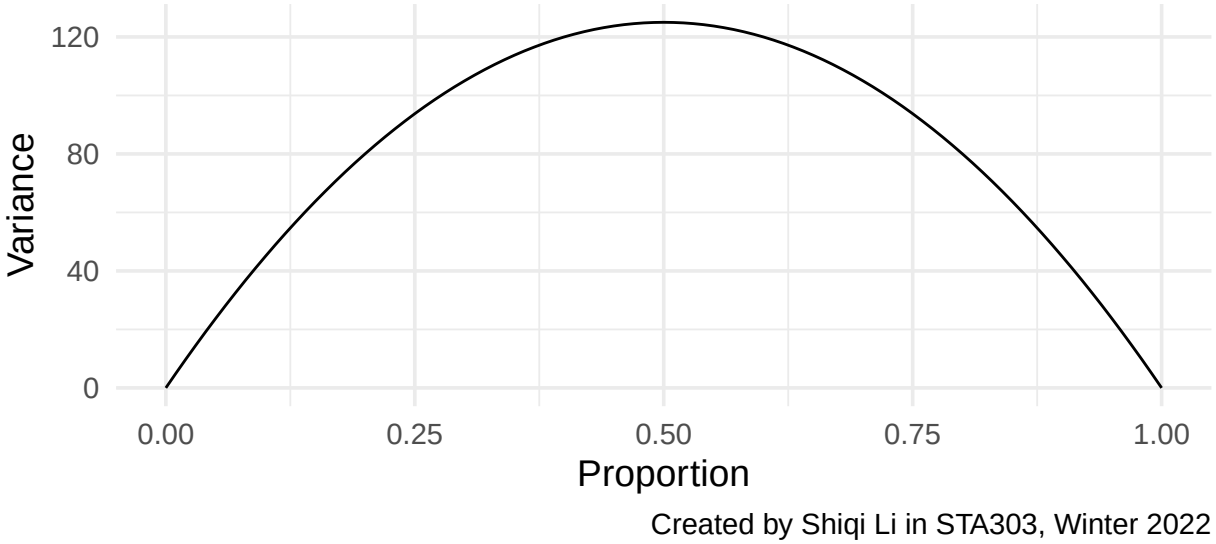


Figure 2: Binomial Variance vs Proportion Plot with n = 500

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

```
set.seed(512) # student id last 3 digits
sim_mean = 10; sim_sd = 2 # N(10, 2)
sample_size = 30; number_of_samples = 100 # sample size 30, 100 samples in total

# at 95% level, t multiplier is at the 97.5% quantile, with df = 30 - 1 = 29
tmult = qt(0.975, df=29)

population = rnorm(1000, sim_mean, sim_sd) # simulating population
pop_param = mean(population) # population mean

sample_set = unlist(lapply(1:number_of_samples,
    function (x) sample(population, size = sample_size))) # get 3000 samples
group_id = rep(seq(1, 100, 1), 1, each=30) # group id to pair 1 - 100
my_sim = tibble(group_id=group_id, sample_set=sample_set) # match sample with id

ci_vals = my_sim %>%
  group_by(group_id) %>% # group by sample id
  summarise(mean=mean(sample_set),
            sd=sd(sample_set)) %>% # mean and sd of each sample
  mutate(lower=mean-tmult*sd/sqrt(sample_size),
         upper=mean+tmult*sd/sqrt(sample_size), # column of interval bounds
         capture=ifelse(pop_param>=lower&pop_param<=upper, TRUE, FALSE))

# proportion of CI captured the true mean
proportion_capture = sum(ci_vals$capture==TRUE)/number_of_samples

# create the ggplot object, first put group id at x-axis
ggplot(ci_vals,
       aes(x=group_id)) +
  geom_point(aes(y=mean)) + # add the mean point
  # bar diagrams of intervals
  geom_errorbar(aes(ymin=lower, ymax=upper),
                col=ifelse(ci_vals$capture, "#122451", "#B80000")) +
  labs( # axis labels, title, and caption
    title = "CI captures population parameter",
    caption = "Created by Shiqi Li in STA303, Winter 2022",
    x = "Group ID",
    y = "Mean"
```

```
  ) +
coord_flip() + # now flip x and y axes
theme_minimal(base_size=14,
              base_family="")
```

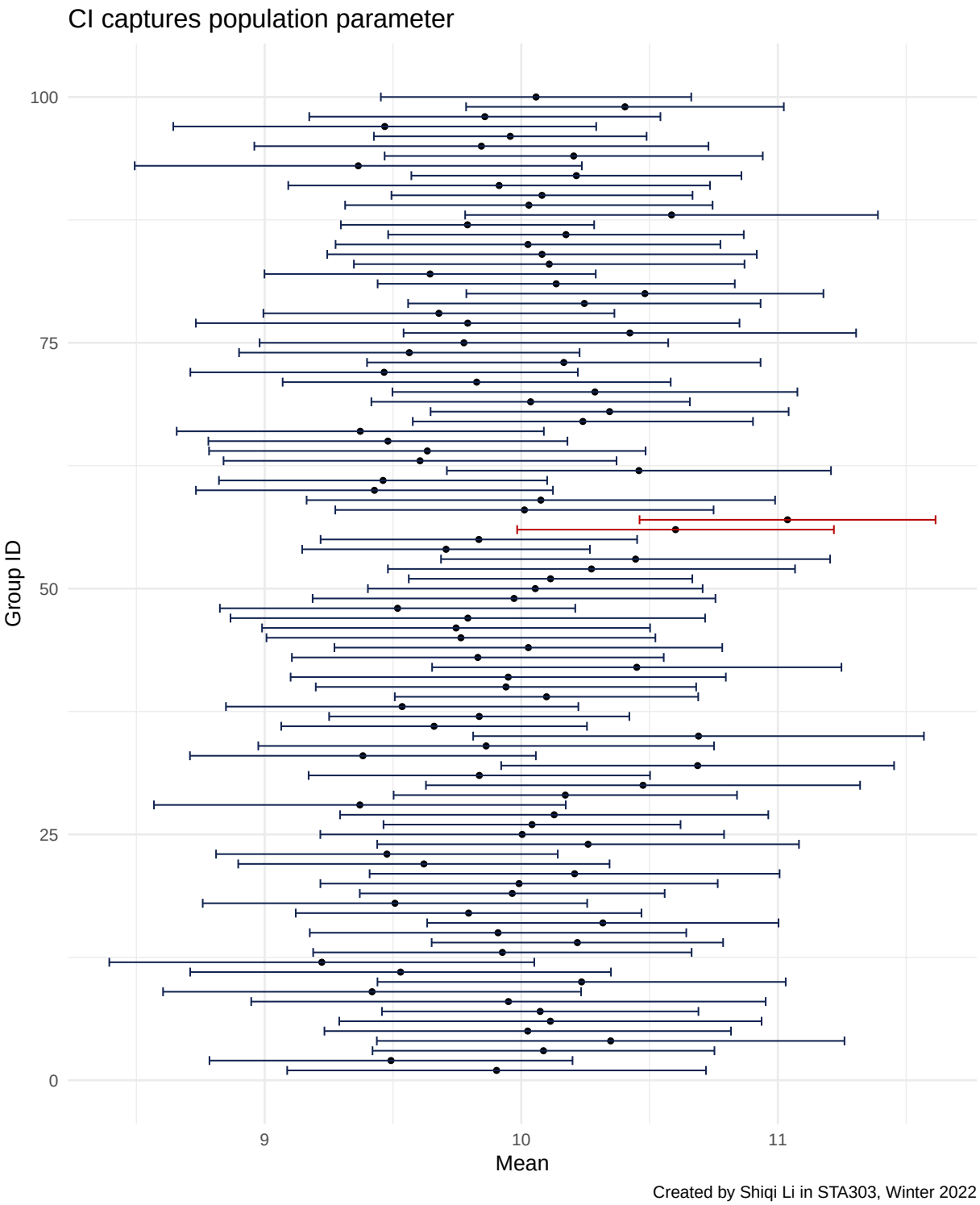


Figure 3: Exploring our long-run "confidence" in confidence intervals. This figure shows how often 95% confidence intervals from 100 simple random samples capture the population mean. The population was simulated from N(10, 2)

In this case the population is a simulated data set from a known distribution $N(mean = 10, sd = 2)$, so we can include the true mean in the confidence interval plot. In fact, this parameter allows us to see how many times out of the 100 samples the confidence interval is able to capture the true value. However we can't generate such a plot in real life because we never know the exact distribution that our sample came from, we can only inference on the true parameter value but have no idea of whether we actually get it right.

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

Goal

Using the class survey data, we wish to determine whether there is an association between cGPA and STA303/1002 students correctly answering a question on global poverty rates. We will test this statement first by conducting the non-parametric Kruskal-Wallis Rank Sum test because we do not know the distributional form of CGPA. We will then fit a linear regression model with CGPA being the response to see if the estimated coefficient has similar p-value to justify the conclusion.

Wrangling the data

```
library(janitor) #ffirst load janitor for clean_names
```

```
##
## Attaching package: 'janitor'
```

```
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

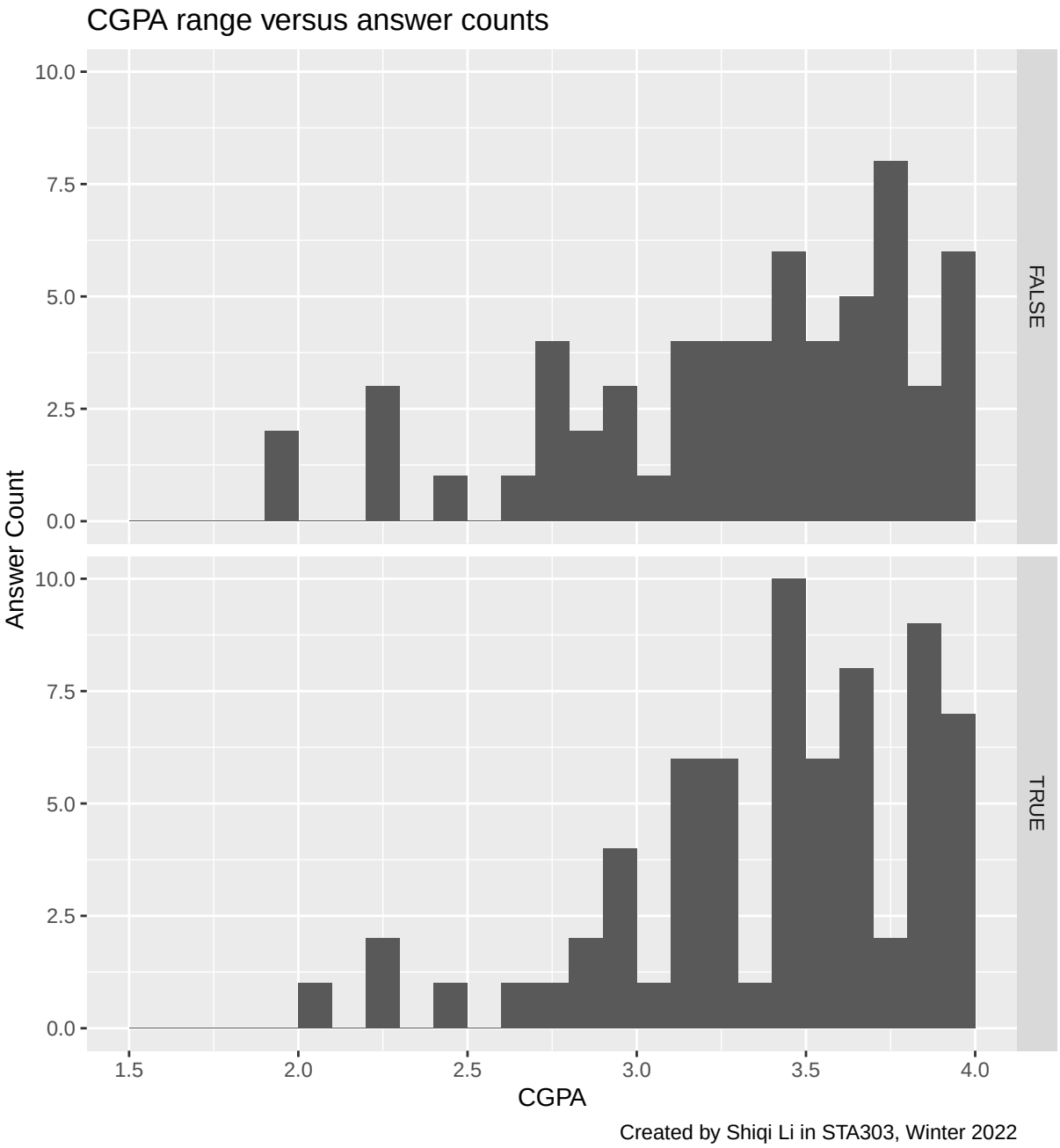
```
cgpa_data = readxl::read_xlsx("~/Downloads/sta303-mini-portfolio-poverty.xlsx") # read
↪  xlsx file
cgpa_data = clean_names(cgpa_data)

# create the required tibble
cgpa_data = cgpa_data %>%
  rename(cgpa=what_is_your_c_gpa_at_u_of_t_if_you_dont_want_to_answer_you_can_put_a_0,
↪
```

```
        ↪  global_poverty_ans=in_the_last_20_years_the_proportion_of_the_world_population_living_i
        ↪  %>% # rename columns
  filter(cgpa != 0) %>% # get rid of NA and 0 values of cgpa
  # create a column indicating whether the answer is correct or not
  mutate(correct = ifelse(global_poverty_ans=="Halved", TRUE, FALSE))
```

Visualizing the data

```
# create the ggplot object for histogram
ggplot(cgpa_data,
       aes(cgpa)) +
  geom_histogram(breaks=seq(1.5, 4, 0.1)) + # histogram column width 0.1
  facet_wrap(~correct) + # two plots for each level of `correct`
  facet_grid(rows=vars(correct)) + # make two plots vertical
  labs( # axis labels, title, and caption
    title = "CGPA range versus answer counts",
    caption = "Created by Shiqi Li in STA303, Winter 2022",
    x = "CGPA",
    y = "Answer Count"
  )
```



Testing

As we want to determine the association between a binary variable and a continuous variable, then the t-test procedure is not applicable. Instead we can use a ANOVA test or a Kruskal-Wallis Rank Sum test. However, ANOVA test relies on the assumptions of linear models, which means it requires CGPA variable to be normally distributed. However from the above histogram we see that it is no where near a normal distribution. Hence the non-parametric Kruskal-Wallis Rank

Sum test is more suitable.

```
kruskal.test(cgpa~correct, data = cgpa_data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  cgpa by correct
## Kruskal-Wallis chi-squared = 0.58867, df = 1, p-value = 0.4429
```

We see the test has p-value of 0.4429, which means there is no evidence suggesting any association between CGPA and correct answers. We then fit a linear model with correct as the only predictor, the p-value below is 0.297 also means that correctness of this question does not contribute to explaining CGPA, which means there is no association between them, this conclusion is the same as the Kruskal-Wallis Rank Sum test.

```
reg_model = lm(cgpa~correct, data=cgpa_data) # save linear model object
summary(reg_model) # results
```

```
##
## Call:
## lm(formula = cgpa ~ correct, data = cgpa_data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -16129034 -16129032        -1         0 983870964
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  16129036   11134436   1.449    0.150
## correctTRUE -16129032   15395204  -1.048    0.297
##
## Residual standard error: 87670000 on 128 degrees of freedom
## Multiple R-squared:  0.008502,   Adjusted R-squared:  0.000756
## F-statistic: 1.098 on 1 and 128 DF,  p-value: 0.2968
```

Writing sample

Introduction

Dear Hiring Manager,

I am a third-year statistics student at the University of Toronto who holds a keen interest in the data industry. I would like to join your team to grow my career with your innovative ideas, and therefore would like to request an opportunity for interview. In the below, I would like to demonstrate my skill set trained from my academic and community experience that meets your requirement profiles.

Soft skills

As an outgoing and driven student, I always like to seek challenges even outside of academics. During my university years, I consistently participated in different local volunteering opportunities such as the yearly Waterfront Marathon event and the Toronto TIFF Film Festivals. My responsibilities range from organizing university students to form volunteer teams, communicating with the event host to understand demand, handled logistics of volunteers during the events. These experience improved my interpersonal skills to communicate effectively under different settings.

I am also the director of the Events Department at the U of T Chinese Music Club where I was in charge of designing, campaigning, and executing on-campus events such as singing competitions, live house performances, and freshmen welcome parties. From this experience, my leadership skill was exercised and improved.

Analytic skills

By studying in a competitive discipline that puts high demand on intellectual challenges, both of my theoretical and applicable skills are trained well. During the program, I built a solid background in statistical theories such as hypothesis testing, probability distributions, and observational and experiment study design. I also had hands-on experience in modeling tools such as linear regression, time series analysis, and casual inference techniques. As a course project, I studied the potential causal relationship between heart attack and rheumatoid arthritis using propensity score matching.

These theoretical knowledge were all learnt with code examples in R, hence my programming skills were also improved during the process. Besides, I also spent time self-learning Python with popular packages such as `numpy` and `pandas`.

Connection to studies

Besides the skills demonstrated above, I am also working on presentation methods for technical projects such as typesetting documents in formats such as LaTex, Markdown, and Jupyter Notebook both in my programs and potentially in professional environments. In my statistical courses, I always tried to typeset my project and assignments using RMarkdown to make it look more professional. I also plan to spend time experimenting with Notebook for Python projects.

Conclusion

As presented above, with the comprehensive skill set I am developing, I think I am on the right track of meeting the rapidly changing industry demands. In order to actually become a competent candidate in this urging industry, I think I will still need to put more effort into building a stronger portfolio.

Word count: 444 words

Reflection

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

I am proud in this mini-portfolio that I can quickly think of the solution to the data wrangling and visualization tasks, and that I am able to write some cover letter paragraphs following a specific job description. This mini-portfolio improves my confidence in the stage of exploratory analysis in a data science project, as well as the ability to convey ideas and demonstrate my own strengths and weaknesses. The final typeset output file also looks nice and professional, which I think I am comfortable to showcase to my potential employers.

How might I apply what I've learned and demonstrated in this mini-portfolio in future work and study, after STA303/1002?

Since this mini-portfolio mainly focuses on exploratory analysis, the skills I exercised in the skill sample part would be very helpful as I move on to future portfolios and projects. I will continue to improve my proficiency in the uses of `ggplot` and `tidyverse` in this course. Also, the writing sample provides a basic structure for my cover letter writing which I can put into real uses when I apply for related jobs. While the content is still yet to improve, I think I already have a rough idea of how to structure a competent cover letter.

What is something I'd do differently next time?

I think overall this mini-portfolio demonstrated different skills in pieces, so for the next time, I would like to add more structures to the write-up so that the demonstration can appear more logical for readers to follow. For examples, I can design a study project that utilizes the models I have learned which integrates stages from exploratory analysis to interpreting and discussing the final results.

Also for the writing sample part, next time I would like to imagine it to be a real cover letter. This means I would need to do some industry and company research to persuade the hiring manager that I would be a good fit to the team.