STA303/1002 Mini-portfolio

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

Ruizhe Huang

2022-02-03

Contents

List of Figures

Introduction

STA303/1002 is a course about how to wrangle and explore a dataset; create appropriate data visualizations; describe ethical considerations in data analysis; understand the assumptions and appropriate use cases for linear mixed models, generalized linear models, generalized linear mixed models and generalized additive models; write and execute R code for the model types covered in this course; accurately and appropriately interpret the results of the model types.
In this mini-portfolio, I will first perform the statistical skills such as setting up libraries; visualizing the variance of a Binomial random variable for varying proportions; demonstrating frequentist confidence intervals as long-run probabilities of capturing a population parameter; investigating whether there is an association between cGPA and STA303/1002 students correctly answering a question on global poverty rates. Secondly, I will write down my soft skills and analytic skills to apply for a data scientist job, and some other skills I would like to develop during the remainder of my education. In the end, I will include a reflection section to discuss something specific I'm proud of in this mini-portfolio; apply what I've learned and demonstrated in this mini-portfolio in future work and study, after STA303/1002; something I'd do differently next time.

Statistical skills sample

Setting up libraries

```
# load the libraries
library(tidyverse)
library(readxl)
```

Visualizing the variance of a Binomial random variable for varying proportions

```
# choose n1
n1 <- 10
# choose n2
n2 <- 100
# create a vector of proportion
props <- seq(0, 1, 0.01)
# create a tibble
for_plot <- tibble(props,
                   n1_var = n1*props*(1-props),
                   n2_var = n2*props*(1-props))
# plot for n1
for_plot %>%
  ggplot(aes(x = props, y = n1_var)) +
  geom_line() +
  geom_vline(xintercept = 0.5, linetype = "dashed", colour = "red") +
  theme_minimal() +
  labs(caption = "Created by Ruizhe Huang in STA303/1002, Winter 2022",
       x = "proportions",
       y = "variance")
```

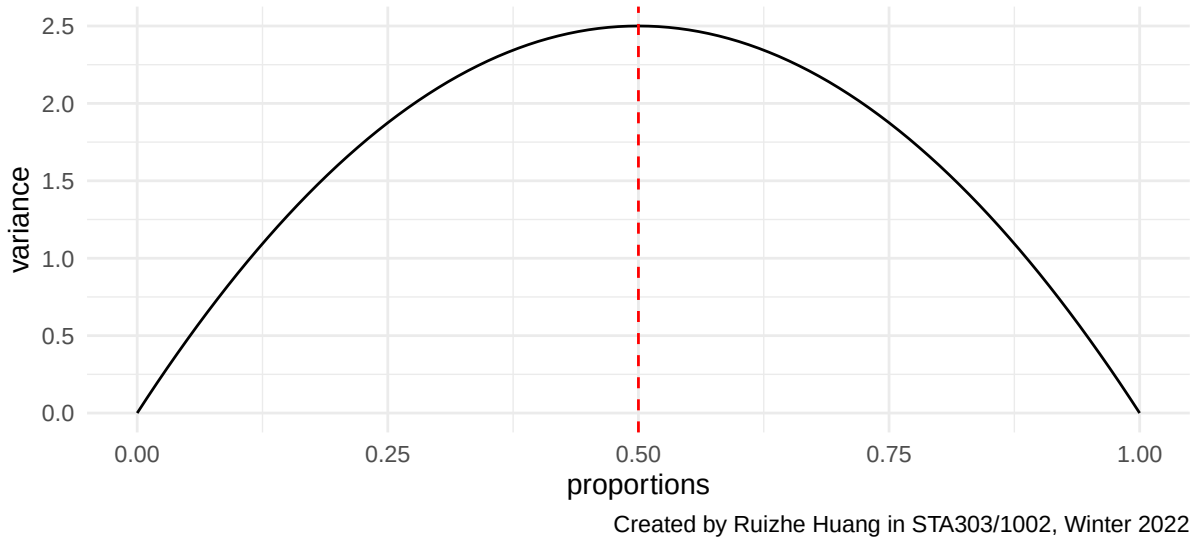


Figure 1: variance of a Binomial random variable for varying proportions with n1=10

```
# plot for n2
for_plot %>%
  ggplot(aes(x = props, y = n2_var)) +
  geom_line() +
  geom_vline(xintercept = 0.5, linetype = "dashed", colour = "red") +
  theme_minimal() +
  labs(caption = "Created by Ruizhe Huang in STA303/1002, Winter 2022",
       x = "proportions",
       y = "variance")
```

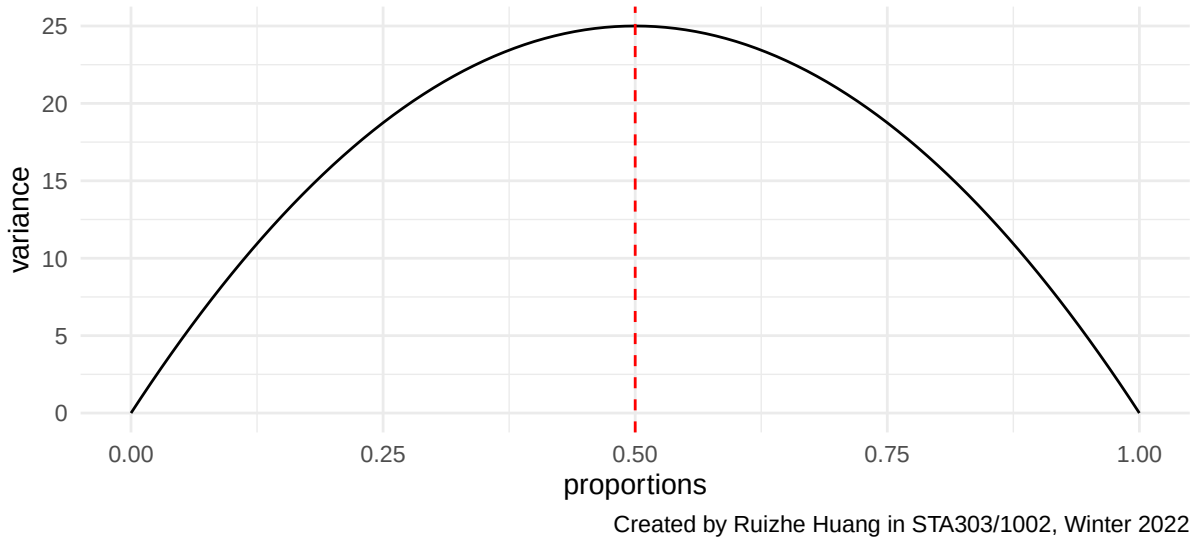


Figure 2: variance of a Binomial random variable for varying proportions with n2=100

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

```r
# set seed to the last three digits of student ID
set.seed(331)
# simulation mean
sim_mean <- 10
# simulation standard deviation
sim_sd <- sqrt(2)
# simulation sample size
sample_size <- 30
# simulation number of samples
number_of_samples <- 100
# Calculate the t-multiplier
tmult <- qt(0.975, df = sample_size - 1)
# simulate population
population <- rnorm(1000, sim_mean, sim_sd)
# actual true mean for population
pop_param <- mean(population)
# Get 100 samples of size 30 from population
sample_set <- unlist(lapply(1:number_of_samples,
    function (x) sample(population, size = sample_size)))
# Create a new vector that will allow we label the values from the 100 different
↪  samples above
group_id <- rep(1:100, each = 30)
# Create a new tibble
my_sim <- tibble(group_id, sample_set)
# Create a new tibble
ci_vals <- my_sim %>% group_by(group_id) %>%
  summarise(mean = mean(sample_set),
            sd = sd(sample_set))
# two columns that hold the lower and upper bound of a 95% confidence interval for the
↪  group
ci_vals <- ci_vals %>% mutate(lower = mean - tmult*sd/sqrt(sample_size),
         upper = mean + tmult*sd/sqrt(sample_size),
         capture = (pop_param >= lower) & (pop_param<=upper)) # capture which takes
         ↪  the values TRUE if the population parameter is in the 95% CI, and FALSE
         ↪  if not

# stores the proportion of intervals created that capture the population parameter
proportion_capture <- mean(ci_vals$capture)
```

```r
# Colour the confidence intervals by whether or not they contain the population
↪  parameter
ci_vals %>% ggplot(aes(x = group_id, y = mean, color = capture)) +
  scale_color_manual(values = c("#B80000", "#122451")) +
  geom_point() +
  geom_errorbar(aes(ymin = lower, ymax = upper)) +
  geom_hline(yintercept = pop_param, linetype = "dotted") +
  labs(caption = "Created by Ruizhe Huang in STA303/1002, Winter 2022",
       color = "CI captures population parameter") +
  # Flip the coordinates
  coord_flip() +
  theme_minimal()
```

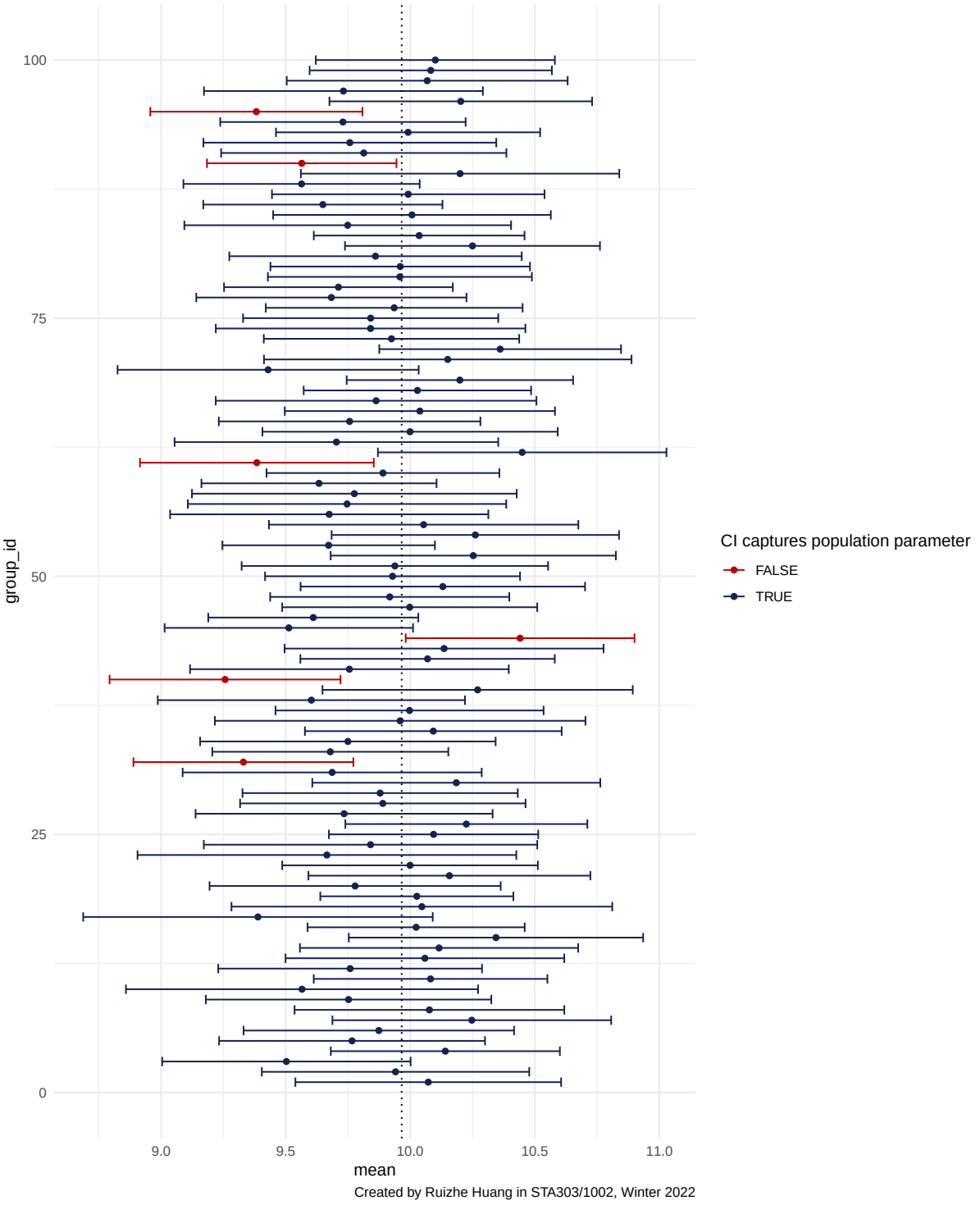


Figure 3: Exploring our long-run 'confidence' in confidence intervals. This figure shows how often 95% confidence intervals from 100 simple random samples capture the population mean. The population was simulated from N(10, 2)

94% of my intervals capture the the population parameter.
We can include the population parameter in this plot, since we simulate the population manually then we have the population mean. However, in the real life the population mean is unable to be simulated, it is unknown since population is too extensive in practice.

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

Goal

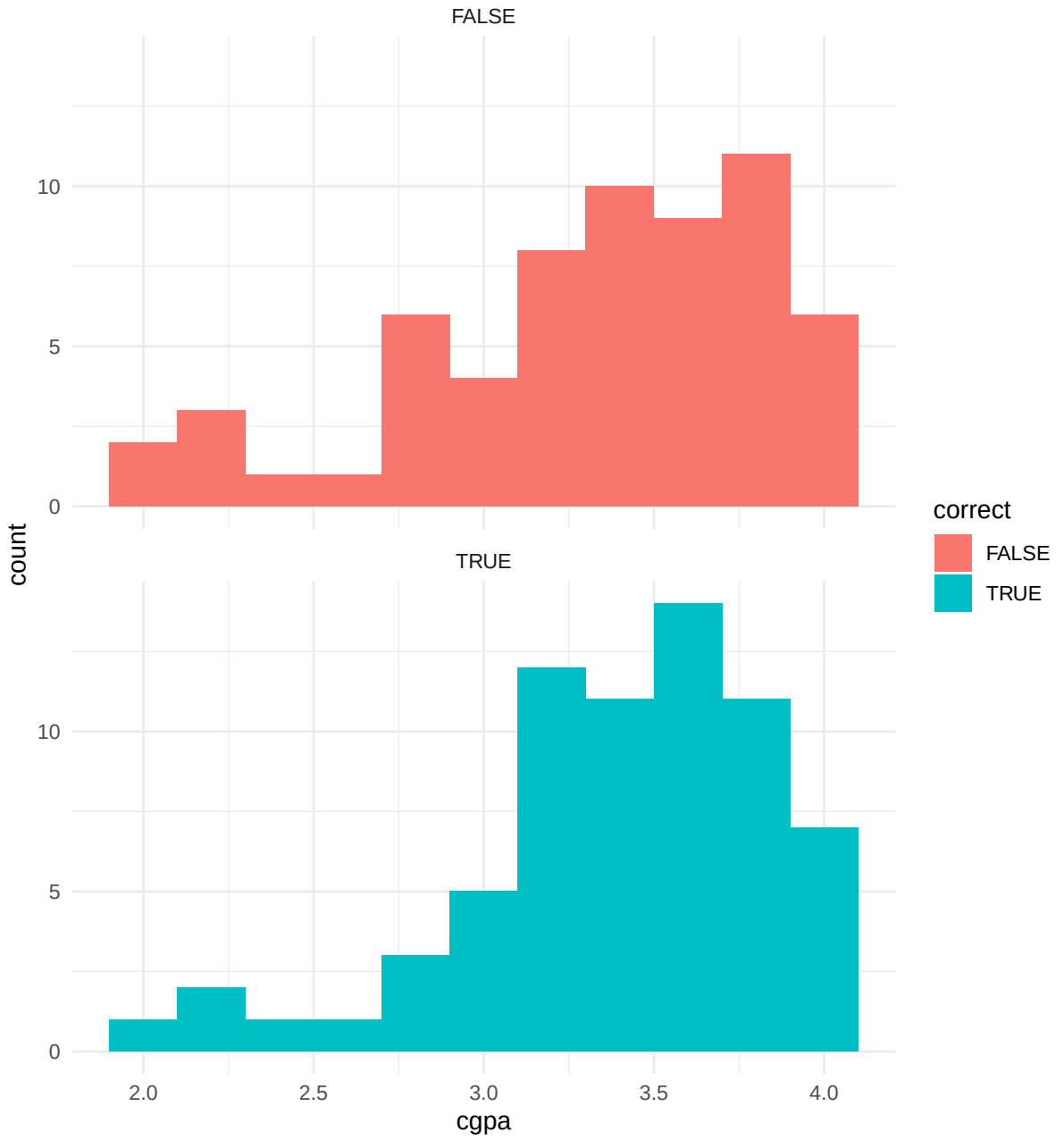
Among the students who correctly answered that the proportion of the global population living below the poverty line has halved and those who did not, compare the students' cGPA to test that whether there is a difference. If the cGPA among them are the same, then there is no association between cGPA and STA303/1002 students correctly answering a question on global poverty rates.

Wrangling the data

```
# import data
cgpa_data <- read_excel("data/sta303-mini-portfolio-poverty.xlsx") %>%
  janitor::clean_names() %>%
  # rename the variables
  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_in_extreme_poverty_ha
         ↪  %>%
  # remove the missing values
  filter(!is.na(cgpa)) %>%
  # keep only appropriate cGPA
  filter(cgpa > 0 & cgpa <= 4) %>%
  # Create a new variable that takes the value TRUE if the respondent answered
  ↪  'Halved' and FALSE if they answered 'Doubled' or "stayed about the same'.
  mutate(correct = (global_poverty_ans == "Halved"))
```

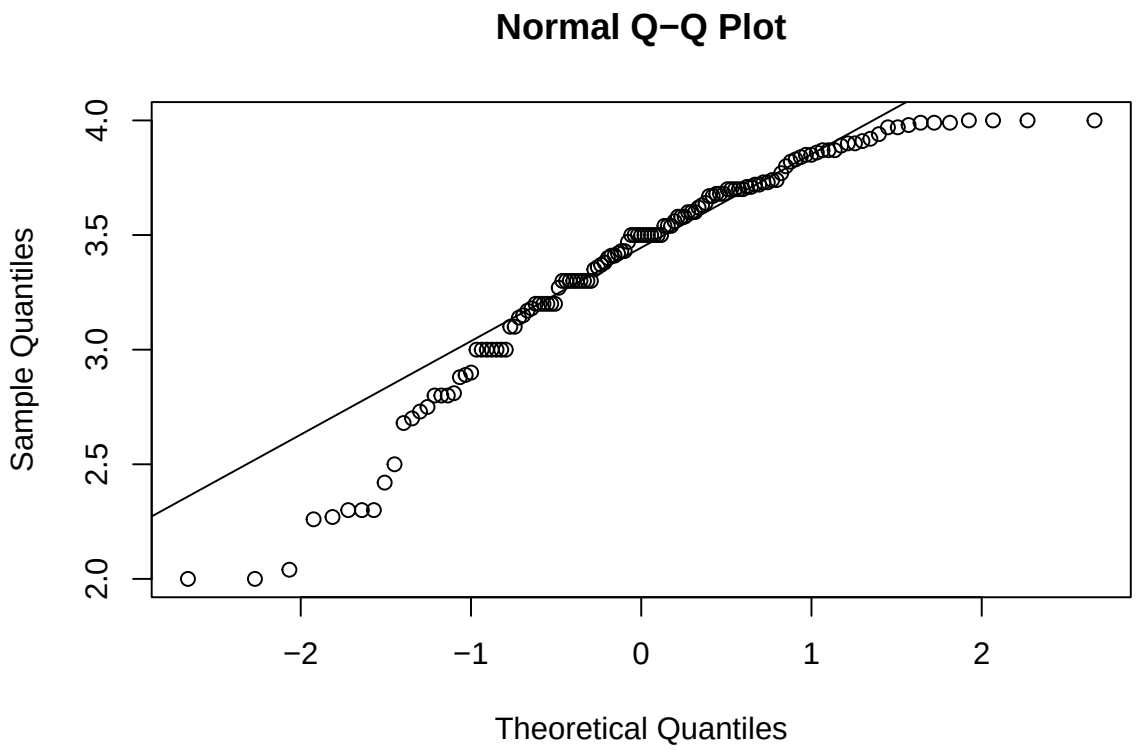
Visualizing the data

```
# Create the histograms positioned on top of each other
cgpa_data %>%
  ggplot(aes(x = cgpa, fill = correct)) +
  geom_histogram(binwidth = 0.2) +
  facet_wrap(~correct, ncol = 1) +
  theme_minimal()
```



Testing

```
# plot the normal qq plot
qqnorm(cgpa_data$cgpa)
qqline(cgpa_data$cgpa)
```



From the normal Q-Q plot of the response variable cGPA in the linear model, the plot is somewhat deviated thus the normality assumption is not well satisfied. Therefore, I will use the non-parametric tests as it does not depends on the Normality assumption.

```
# linear model
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
```

Since the p-value is 0.351, which is greater than 0.05, we fail to reject the null hypothesis at 5% threshold, the true mean cGPA between whether or not correctly answer the question is the same. Therefore, there is no association between cGPA and STA303/1002 students correctly answering a question on global poverty rates.

```
# 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
```

Since the p-value is 0.35, which is greater than 0.05, we fail to reject the null hypothesis at 5% threshold, the true mean cGPA between whether or not correctly answer the question is the same. Therefore, there is no association between cGPA and STA303/1002 students correctly answering a question on global poverty rates.

Writing sample

Introduction

I am going to apply for a remote Data Scientist job in the department of Engineering and Product at Yelp. The Data Science team performs analyses, builds models, and designs experiments that directly impact Yelp's business and users. In the following part, I would write on the soft skills and analytic skills that Yelp is looking for.

Soft skills

I think I possess the communication skills to work with partners on engineering, product and business teams. I am an active listener, I will pay attention to the person I am speaking with. I can give and receive feedback, I will accept constructive criticism so that I can improve myself to be a better colleague. I also can give constructive feedback to others.
Moreover, I believe that I am enthusiastic about clean code and sharing reproducible results. I usually make comments while coding, I filter the missing values as needed and I do not do "hardcoded".

Analytic skills

I think I am fluent with R for data analysis. I used R and R libraries such as tidyverse did many projects. I can conduct the hypothesis test or confidence interval for the population mean, variance and proportion. I can create a model for a given dataset, and check the assumptions of the model by residual plot, normal Q-Q plot and so on.
Besides, I can visualize the data with ggplot. I can plot the different kinds of histogram, boxplots, barplot and scatterplots. For example, use two-sided boxplots to compare the median values for a categorical variable and a numerical variable.

Connection to studies

I would like to develop my leadership especially in the case of conflict management. During the remainder of my education, I will do more teamwork while during team projects and practice empathy sees the situation from other viewpoints in the conflict.
In addition, I would develop Python, SQL and Tableau skills. During the remainder of my education, I will be actively involved in learning the courses related to Python, SQL and Tableau and able to use them in the future.

Conclusion

In summary, I have good communication skills and I am enthusiastic about clean code and sharing reproducible results. Moreover, I am fluent with R for data analysis and I can visualize the data with ggplot. During the remainder of my education, I will develop my leadership in case of conflict management and develop Python, SQL and Tableau skills.

Word count: 409 words

Reflection

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

I am really proud of the visualization especially the confidence intervals plot looks aesthetic, the legend is clear and the results of the confidence interval is coloured so it is easier to identify which confidence interval captures the population mean. Moreover, I am proud of I develop my writing skills during this mini-portfolio and examine myself towards a data scientist job to discover my soft skills and analytic skills. It motivates me to develop more necessary skills such as fluent use of python, SQL and Tableau to qualify for the desired job. I am also proud of the reflection session, I am able to have an overall idea throughout this mini-portfolio and reflect critically.

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

I have learned how to make an aesthetic cover page and format the report more pleasing. I would apply these skills in future courses and projects. If I am going to do a statistical report, I would make each plot with the necessary legend, caption, and apply the theme to the plots. I also learned how to make a catalogue, I would add a catalogue to my future report so that it is easier to find the sections. In addition, I have learned how to examine myself by writing down the soft skills and analytic skills I possess in terms of the data scientist job. I would apply these skills once I have similar job opportunities to be well prepared.

What is something I'd do differently next time?

I am not very familiar with how to design a cover page and how to make the plot looks more aesthetically pleasing this time, so I have to use the template provided. I would do this mini-portfolio without a template to practice what I've learned this time and try to be creative on the formatting next time. What's more, I would try to develop and write down more soft skills and analytic skills to better fit the job requirement next time. Especially, in terms of soft skills, there is a lot more needed to be improved. I want to develop my public speaking skill and write it down next time.