STA303/1002 Mini-portfolio

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

Yuchen Zeng

2022-02-03

Contents

Introduction **3**

Statistical skills sample **4**

Setting up libraries 4
Visualizing the variance of a Binomial random variable for varying proportions 4
Demonstrating frequentist confidence intervals as long-run probabilities of capturing a population parameter 6
Investigating whether there is an association between cGPA and STA303/1002 students correctly answering a question on global poverty rates 9

Writing sample **14**

Reflection **16**

List of Figures

1 Proportion vs. Variance of a Binomial random variable with n=10 5
2 Proportion vs. Variance of a Binomial random variable with n=50 5
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). 8
4 Histogram of cGPA of students that correctly answered global poverty question and those who did not 11

Introduction

This mini-portfolio is an assignment of STA303 at the University of Toronto. This course discusses the exploration of data sets, data visualizations, data interpretation, and ethics in data analysis. It also studies various statistical models using R. In this mini-portfolio, I presented my statistical knowledge, data wrangling and visualization skills, writing skills, and self-reflection.

The first section is a presentation of statistical analysis and R skills. It shows how to set up libraries in R and visualize the spread of Binomial random values for varying proportions. I also simulated a population and took random samples. I demonstrated confidence intervals as long-run probabilities using these samples. I used statistical testing to analyze the relationship in cGPA between students who correctly answered a global poverty question and those who did not. I then identified and explained the test I chose. According to the result, there is no evidence of a difference in cGPA related to the question.

The second section is an article about my discussion on skills needed for a data scientist job position in Yelp. I mentioned the importance of soft skills such as communication and presentation skills. I also stated the importance of analytic skills such as fluency in R and statistical knowledge such as building a linear regression model. Moreover, I mentioned organization skill which is beneficial for similar jobs. Lastly, I examined how I possessed these skills and ways to improve.

The last section is a reflection of my performance in this mini-portfolio. I discussed the part I am proud of, what I have learned to apply in future work and study, and what I would do differently next time.

Statistical skills sample

Setting up libraries

```
# Load tidyverse library
library(tidyverse)
# Load readxl library
library(readxl)
```

Visualizing the variance of a Binomial random variable for varying proportions

```
# Define n1, n2 which is trial size for Binomial random variable
n1 <- c(10L)
n2 <- c(50L)
# Define Proportion
props <- seq(0, 1, by=0.01)
# Construct tibble, calculate variance with n1, n2 using np(1-p)
for_plot <- tibble(props,
                   n1_var=n1 * props * (1-props),
                   n2_var=n2 * props * (1-props))
# Plot Variance vs. Proportion with n1
ggplot(for_plot, aes(x = props, y = n1_var)) +
  geom_point() +
  labs(caption  =
         "Created by Yuchen Zeng in STA303/1002, Winter 2022.",
       x = "Proportion",
       y = "Variance") +
  theme_minimal()
```

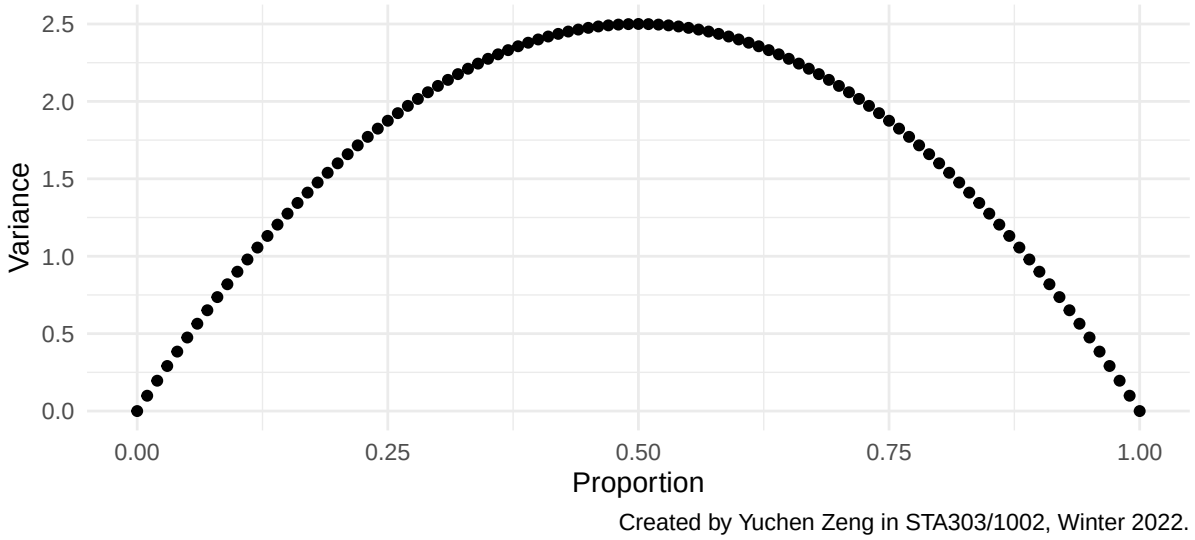

Figure 1: Proportion vs. Variance of a Binomial random variable with n=10

```
# Plot Variance vs. Proportion with n2
ggplot(for_plot, aes(x = props, y = n2_var)) +
  geom_point() +
  labs(caption  = "Created by Yuchen Zeng in STA303/1002, Winter 2022.",
       x = "Proportion",
       y = "Variance") +
  theme_minimal()
```

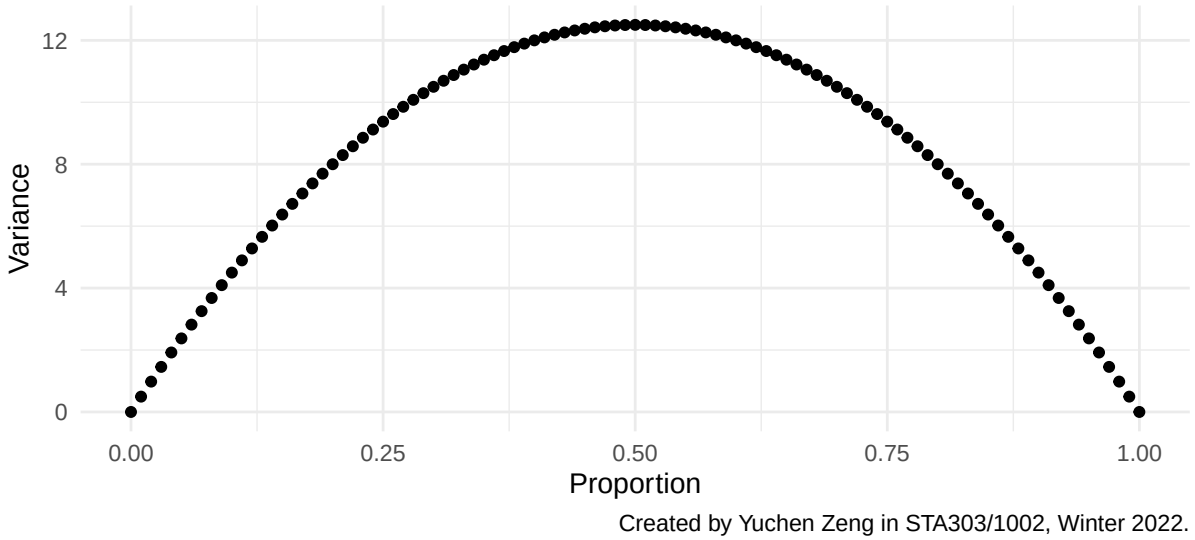

Figure 2: Proportion vs. Variance of a Binomial random variable with n=50

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

```
# Set up seed for randomization
set.seed(825)
# Define given variables
sim_mean <- 10
sim_sd <- sqrt(2)
sample_size <- 30
number_of_samples <- 100
# Define t multiplier for 95% confidence interval
tmult <- qt(0.025, df = sample_size - 1)
# Simulate population with 1000 values
population <- rnorm(1000, mean=sim_mean, sd=sim_sd)
# Find average of population values
pop_param <- mean(population)
# Take <number_of_samples> samples with size <sample_size> from population
sample_set <- unlist(lapply(1:number_of_samples,
    function (x) sample(population, size = sample_size)))
# Group each sample with size <sample_size>
group_id <- rep(c(1:number_of_samples), each = sample_size)
# Save sample values in a tibble
my_sim <- tibble(group_id, sample_set)
# Group samples and calculate some statistics
ci_vals <- tibble(my_sim %>%
                    group_by(group_id) %>%
                    summarise(mean = mean(sample_set),
                              sd = sd(sample_set)))
ci_vals <- ci_vals %>%
  mutate(lower = mean + tmult*sd/sqrt(sample_size),
         upper = mean - tmult*sd/sqrt(sample_size),
         capture = (lower <= pop_param & pop_param <= upper))
# Save proportion of samples that capture the population average
proportion_capture <- mean(ci_vals$capture)
```

```
# Plot confidence intervals of all <number_of_samples> samples
ggplot(data = ci_vals) +
  # Plot confidence interval
  geom_errorbar(aes(x = group_id, ymin = lower, ymax = upper, color = capture)) +
  # Set up colours
  scale_color_manual(values=c("#B80000", "#122451")) +
```

```
# Plot population mean
geom_hline(yintercept = pop_param, linetype = "dotted") +
# Plot sample mean
geom_point(aes(x = group_id, y = mean, color = capture)) +
coord_flip() +
theme_minimal() +
theme(legend.position = "bottom",
      axis.title.x = element_blank(),
      axis.text.y = element_blank(),
      axis.title.y = element_blank()) +
labs(color="CI captures population parameter.",
     caption  = "Created by Yuchen Zeng in STA303/1002, Winter 2022.")
```

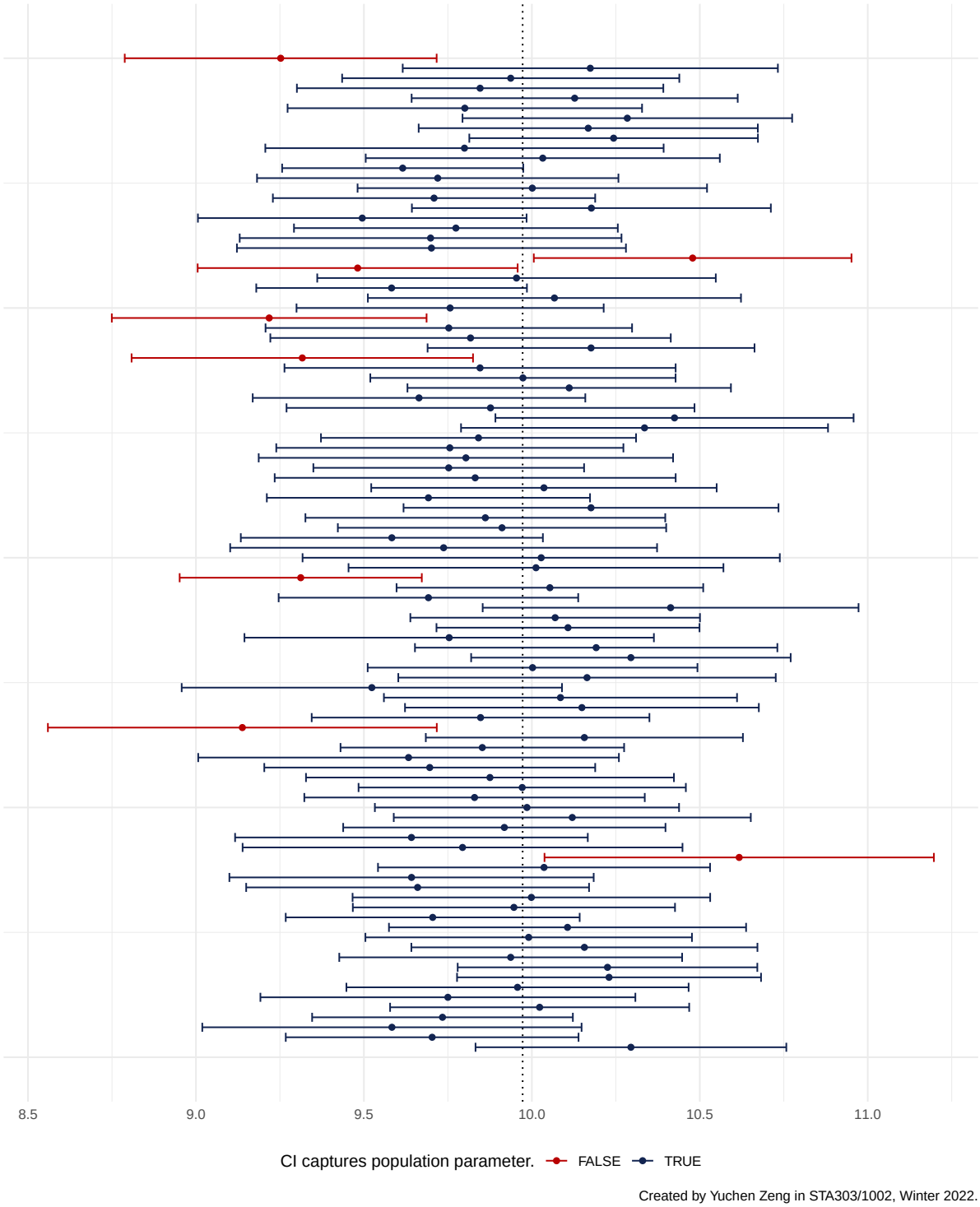

Figure 3: Exploring our long-run "confidence" in confidence intervals. This figure shows how often 95% confidence intervals from 100 simple random samples capture the population mean. The population was simulated from N(10, 2).

92 % of my intervals capture the the population parameter.

In this plot, we took our samples from a simulated population. Thus, we already knew the population average was 10. Therefore, we can draw the population average as a line and compare it to confidence intervals of samples on the plot. However, in practice, we cannot capture all the data in the field we are interested in. Usually, we only have some samples from the population. Since we do not have all the data, we do not know values that require calculating using all the data, such as the population average. Thus, we cannot compare population parameters such as average to our confidence interval.

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

Goal

In the "getting to know you" survey, STA303 students answered a question about global poverty. Given 200 STA303 students' answers to that question and their cGPA randomly sampled from the survey, the goal of this task is to find out if there is a relationship between cGPA of students that answered the question correctly and students that answered the question wrong.

Wrangling the data

```
# Load the data
cgpa_data <- read_xlsx("data/sta303-mini-portfolio-poverty.xlsx")
# Clean data names
cgpa_data <- janitor::clean_names(cgpa_data)
# Rename variables to cgpa and global_poverty_ans respectively
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_in_extreme_poverty_has)%>%
  # Remove cGPA that are not appropriate
  filter(cgpa > 0. & cgpa <= 4.) %>%
  # Create new variable
  mutate(correct = ifelse(global_poverty_ans=="Halved", TRUE, FALSE))
```

Visualizing the data

```
# Set up facet label names
answer.labs <- c("Answered Correctly", "Answered Incorrectly")
names(answer.labs) <- c(TRUE, FALSE)
# Plot histograms for students that answered correctly and students that did not
ggplot(data = cgpa_data, aes(cgpa)) +
  geom_histogram(binwidth=0.05) +
  labs(y = "Count of Students", x = "cGPA") +
  facet_wrap(~ correct, ncol = 1, labeller = labeller(correct = answer.labs))
```

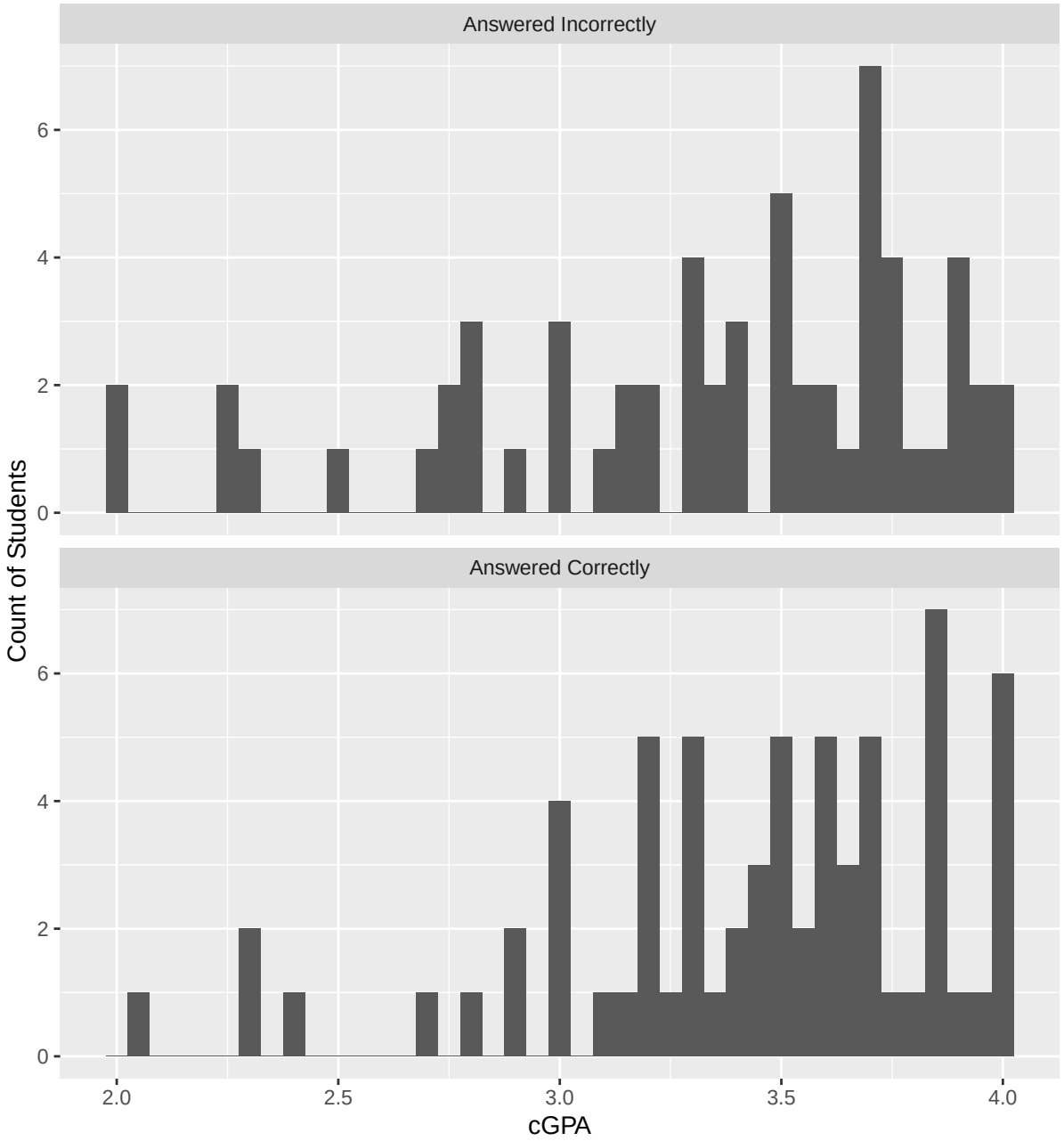

Figure 4: Histogram of cGPA of students that correctly answered global poverty question and those who did not

Testing

I chose to use the Mann-Whitney U test.
We divide our data into two sample, one contains students that answered the question correctly,

one contains students that answered the question incorrectly. We want to use statistical tests to see whether there is a difference of cGPA between the two sample.
The histogram I plotted for students cGPA shows that the distribution of student cGPA in STA303 is left-skewed. Thus, I think it is likely that the distribution of student cGPA is not normally distributed. Since parametric tests such as the t-test and the ANOVA summary assumes the distribution of the population is normal, it is not a good idea to use them because they will be less accurate.
Although Wilcoxon signed-rank test is a non-parametric test, it is used when we want to know the population mean using one sample. Thus, it is not appropriate for this task.
Kruskal-Wallis Rank test and Mann-Whitney U test are both non-parametric tests and can be used to compare two independent samples. Since the Kruskal-Wallis Rank test is an extension to Mann-Whitney U test for more than two samples, I can just use the Mann-Whitney U test. They make the assumption that the samples are independent and randomly sampled. I expect that both assumptions are satisfied, because the data is a random sample of students who responded to the survey, and students' cGPA is probably not related to other students.

```r
# Linear Regression test
summary(lm(data = cgpa_data, rank(cgpa) ~ correct))
```

```
##
## 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
```

```r
# 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
```

The p-value for the Mann-Whitney U test and the equivalent test is 0.35 which are both larger than 0.05. With a significance level of 5%, we do not have enough evidence to reject the null hypothesis that there is no difference in cGPA between students who correctly answered the global poverty question and those who did not.

Writing sample

Introduction

When looking for a job, it is important to have an accurate understanding of the required skills. In this writing assignment, I will discuss soft skills and analytic skills for a remote Data Scientist job from Yelp, and other skills I can develop as a student for a similar job.

Soft skills

In Yelp, data scientists are working with partners on other teams. Thus, Yelp seeks communicating skills from employees for group working success. I should be a good listener for other teams' opinions, giving them feedback and negotiating on time. I have been working in a group for several projects, and when a teammate wants to share their idea, I can quickly recognize it and let him speak. I will also consider the advantages and disadvantages of new ideas and share them with others.

Also, presentation skills are critical for team working. A data scientist should not only be able to perform analysis but also present it. I am good at paraphrasing statistical terms so that others that are not professionals can also understand them. I have experience in presenting and explaining statistic studies in STA130. I always write a summary of my work for my teammates, and they can understand it faster.

Analytic skills

Other than soft skills, some analytic skills are crucial for this position. For example, I need to be fluent with SQL and Python or R. They are useful software tools for efficient data analysis. I have been using R and Python for data analysis for three years, and I have taken courses related to SQL in school. Since I am already fluent in Python and R, I should practice my SQL programming skill by working on a personal project using SQL.

Also, data scientists in Yelp need to design, execute, and analyze experiments and evaluate models. Thus, I need a solid understanding of statistical inference and experimental design and analysis. I have taken courses and learned thoroughly about statistical inference. Also, I completed analysis and research using linear regression. To process my experimental design skills, I am currently taking STA305, which is about experimental design, and I expect to learn more about statistical practices by taking STA355.

Connection to studies

Furthermore, I can possess organizational skills during the remainder of my education to get ready for similar jobs. It includes managing time and using resources efficiently. Currently, I am working on a research project that requires weekly meetings. I record key messages and take notes in every meeting so that my group knows the current goal.

Conclusion

In conclusion, I have completed several statistical types of research that I have adequate knowledge of statistical inference and fluency in software. I also worked in a group on many projects and practiced my communication and presentation skills. In the future, I can also practice my organization skills which are beneficial for similar jobs.

Word count: 469 words

Reflection

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

I am proud of my explanation for my test choice in task 4. When deciding on a suitable statistical test, I found that I should practice more on statistical analysis. I reviewed all statistical tests we learned from Module 1 and did some extra research on the Kruskal-Wallis Rank test. It is beneficial because they are all powerful tools in statistical analysis, and I also need to apply them in other courses. In that question, I wrote a detailed explanation for why each test is appropriate for this question or not. Now, I had a much more thorough understanding of different statistical tests.

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

The statistical skills sample can be posted on my personal website as proof of my skills for future employers. I will also need to apply them in statistical analysis in the future or other courses, and this mini-portfolio reminds me to review materials that I am not familiar with. In the writing sample, I explored a lot of different types of skills needed for statistical job positions. I now have a better understanding of my advantages and which skills I need to keep working on. Furthermore, I think in the future job-seeking process, the writing sample is a good reference for writing my CV letter.

What is something I'd do differently next time?

When working on the mini-portfolio, I found that I have forgotten the usage of some statistical tests. I have to check my note several times in Task 4 to understand what to implement. I learned that I need to review them periodically to reinforce my memory and understand them more thoroughly. Next time, I would like to review what I learned from previous modules and past courses. Also, I should check other classmates' questions regularly, such as in Piazza. Professor Bolton's hints are also very helpful. This way, I can use my time more efficiently when working on the mini-portfolio.