STA303/1002 Mini-portfolio

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

Yu-Chun Chien

2022-02-03

Contents

List of Figures

Introduction

This is the mini-portfolio assignment for the method of data analysis course. In the course, we are taught the ability to wrangle and explore datasets, create appropriate data visualizations, write R code for data analysis and visualization, understand the assumption of different statistical models, and interpret the results of the analysis and communicate with different audiences. In this assignment, I have the opportunity to demonstrate my data wrangling, analysis, and visualization skills, programming skills using R, as well as my communication skills.

In the Statstical skills part of the assignment, I have the the chance to visualize the variance of a Binomial random variable for varing proportions, to demonstrate frequentist confidence intervals as long-run probabilities of capturing a population parameter, and to investigate whether there is an association between cGPA and STA303/1002 students correctly answering a question on global poverty rates. In these tasks, I have demonstrated my data wrangling skills by cleaning, subsetting data, and adding new variables to a dataset. I have also demonstrate my data visualization skills by plotting multiple plots to demonstrate my analysis results, for instance plotting the confidence interval of 100 samples of size 30. Further, I demonstrate my knowledge of assumptions and appropriate use case for hypothesis testing and performed a Mann-Whitney U test.

In both the Statistical Skills and Writing part of the assignment, I have demonstrate my ability to communicate with different audience. In the Statistical Skills part, I explained confidence interval to audience who do not know statistics well. In the Writing part, I communicate my technical and soft skills required for the data scientist job at Yelp, as well as the skills I am developing in my final year of study.

Statistical skills sample

Setting up libraries

```
# load required packages
library(tidyverse)
library(readxl)
```

Visualizing the variance of a Binomial random variable for varying proportions

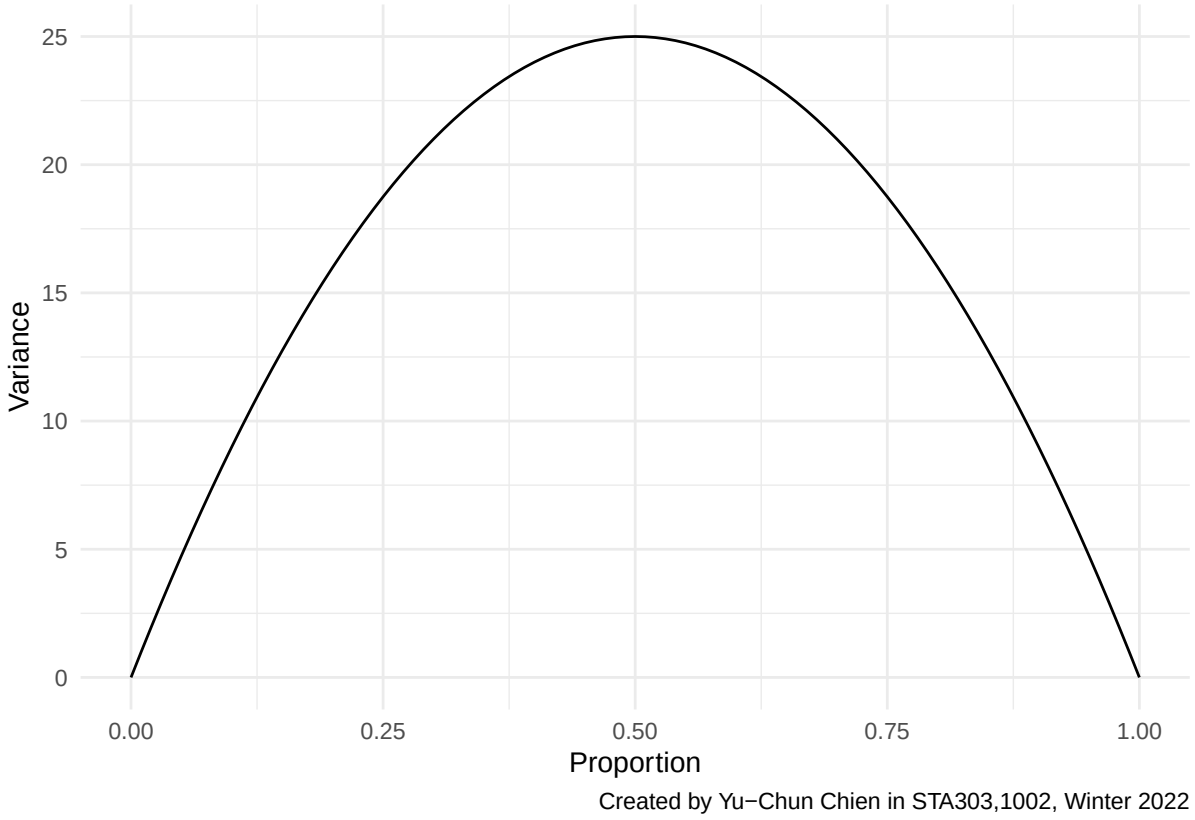
```
# pick two appropriate values of n > 0
n1 <- 100
n2 <- 300

# create a vector of proportions from 0 to 1 in steps of 0.01
props <- seq(0, 1, by = 0.01)

# calculate variance for n1 and n2
n1_var <- n1 * props * (1-props)
n2_var <- n2 * props * (1-props)

# create a tibble for_plot including three variables
for_plot <- tibble(props, n1_var, n2_var)
```

```
# creating plots for n1
ggplot2::ggplot(data = for_plot) + geom_line(aes(x = props, y = n1_var), stat =
↪  "identity") + theme_minimal() + labs(caption = "Created by Yu-Chun Chien in
↪  STA303,1002, Winter 2022") + xlab("Proportion") + ylab("Variance")
```



```
# creating plots for n2
ggplot2::ggplot(data = for_plot) + geom_line(aes(x = props, y = n2_var), stat =
↪  "identity") + theme_minimal() + labs(caption = "Created by Yu-Chun Chien in
↪  STA303,1002, Winter 2022") + xlab("Proportion") + ylab("Variance")
```

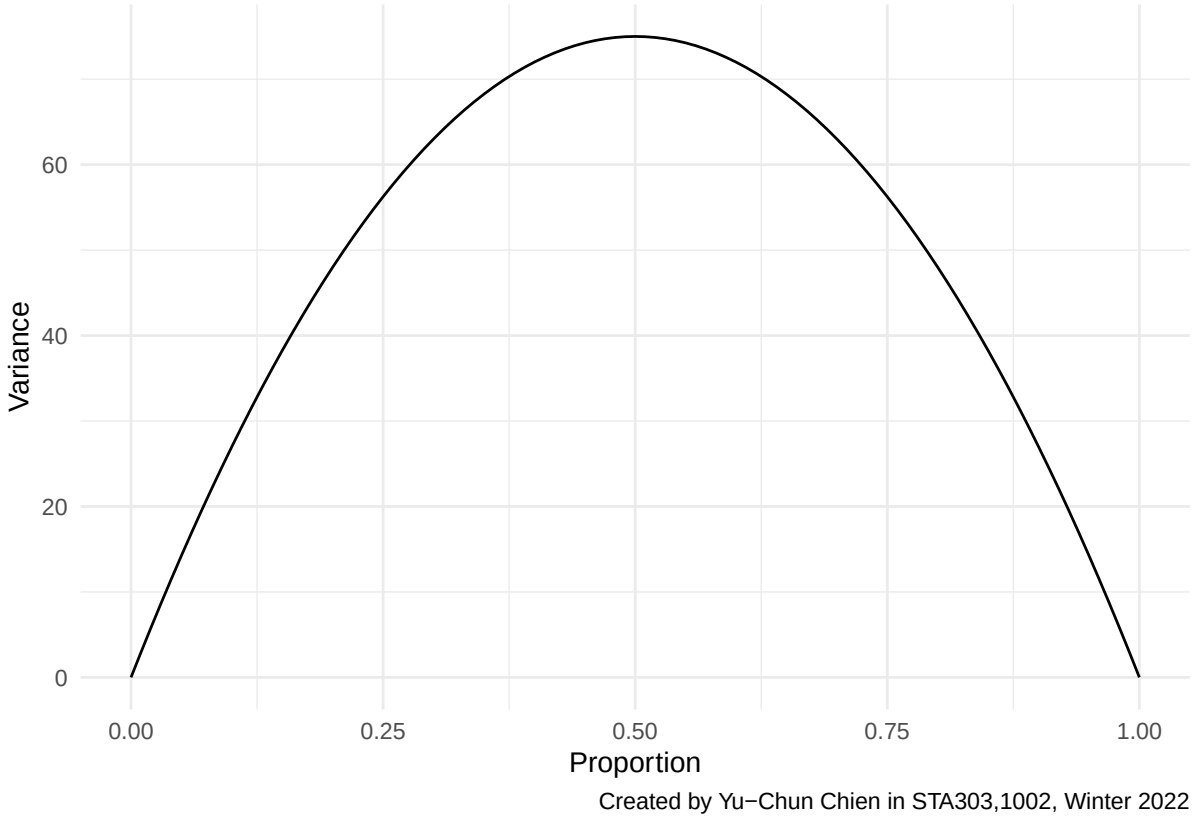


Figure 1: Variance of the Binomial Distribution as Proportion Varies (n = 300)

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

```r
# set up simulation parameters
sim_mean <- 10
sim_sd <- sqrt(2)
sample_size <- 30
number_of_samples <- 100

# calculate t-multiplier
tmult <- qt(0.975, 29)
```

```r
# set seed to last three digits of student ID
set.seed(380)
# create a simulated population
population <- rnorm(1000, mean = sim_mean, sd = sim_sd)

# actual true mean of 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 vector for labeling the 100 samples
group_id <- rep(1:number_of_samples, each = sample_size)
```

```r
# create a new tibble my_sim
my_sim <- tibble(group_id, sample_set)
```

```r
# create a new tibble ci_vals
ci_vals <- my_sim %>%
  group_by(group_id) %>%
  summarise(mean = mean(sample_set), sd = sd(sample_set))

# compute lower and upper bound
lower <- ci_vals$mean - (tmult * ci_vals$sd / sqrt(sample_size))
upper <- ci_vals$mean + (tmult * ci_vals$sd / sqrt(sample_size))
```

```r
# add them to the tibble
ci_vals <- tibble(ci_vals, lower, upper)

# create a temporary variable of a vector of population parameter for comparison
pop_param_temp <- rep(pop_param, 100)

# capture
capture <- ifelse(pop_param_temp >= ci_vals$lower & pop_param_temp <= ci_vals$upper,
↪  TRUE, FALSE)


# add capture to tibble
ci_vals <- tibble(ci_vals, capture)
```

```r
# proportion captured
proportion_capture <- sum(ci_vals$capture)/length(ci_vals$capture)
```

```r
ggplot() + geom_errorbar(data = ci_vals, mapping = aes(x = group_id, ymin=lower,ymax =
↪  upper, color = capture)) + geom_point(data = ci_vals, mapping=aes(x=group_id, y =
↪  mean, color = capture)) + coord_flip() + geom_hline(yintercept=sim_mean,
↪  linetype="dotted") + scale_colour_manual(name = "CI captures population
↪  parameter", values = c("#B80000", "#122451")) + theme_minimal() +
↪  theme(axis.title.x = element_blank(), axis.title.y =
↪  element_blank(),legend.position = "bottom") +  labs(caption = "Created by Yu-Chun
↪  Chien in STA303,1002, Winter 2022")
```

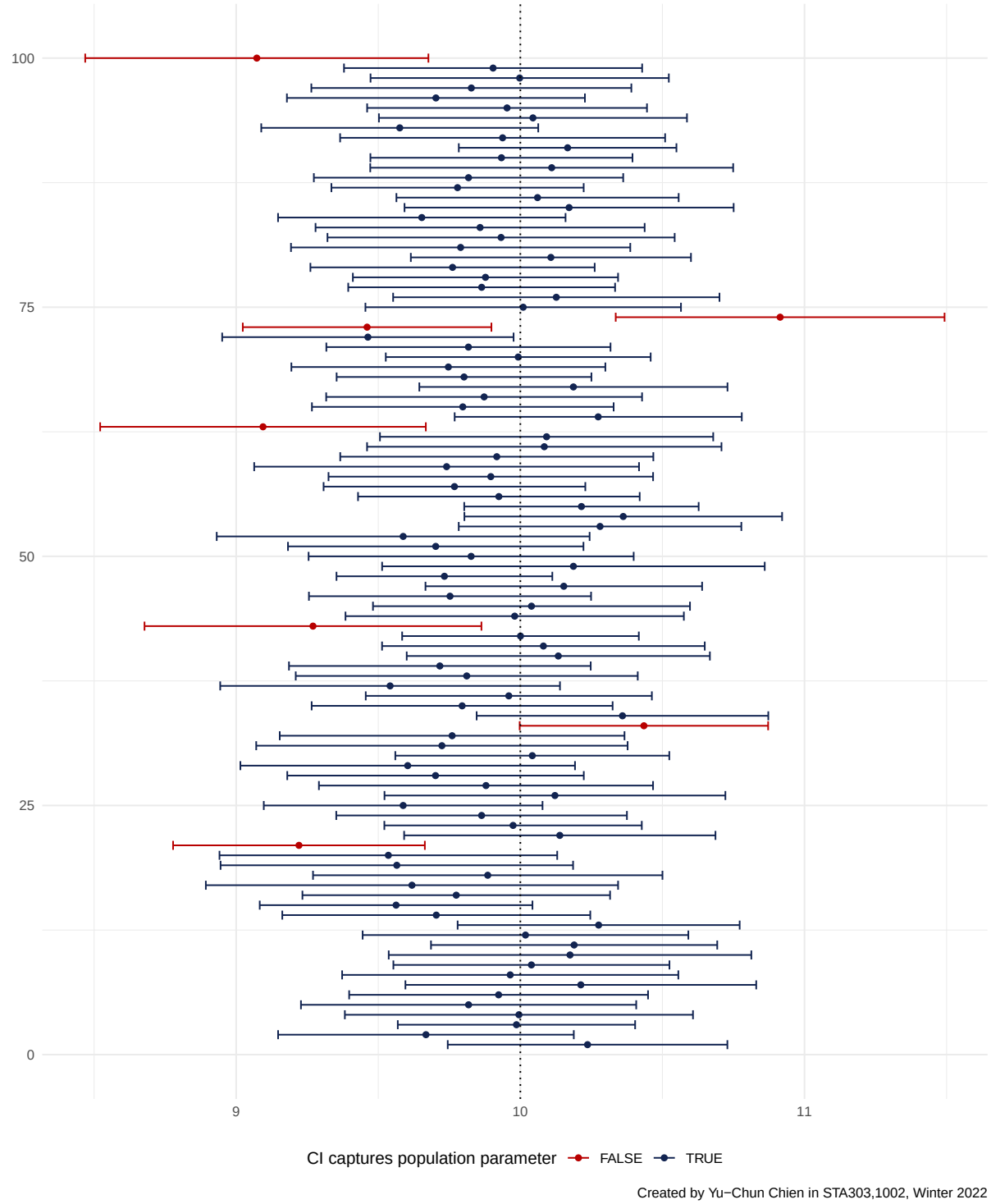


Figure 2: Exploring our long-run "confidence" in confidence intervals. This figure shows how often 95% confidence intervals from 100 simple random samples capture the population mean. The population was simulated from N(10, 2)

93 % of my intervals capture the the population parameter.

Here, since we know our true population parameter, so we can compare it to confidence interval. However, usually we do not know the population parameter. We could only estimate the population parameter from taking sample from the population. By giving a 95% confidence interval, we are claiming that we are 95% certain that most of the samples contain the true population parameter.

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, students from STA303 were asked to provide their cGPA as well as the question "whether the proportion of people living below the global poverty line had halved, doubled or stay about the same in the last 20 years." The correct answer is that the proportion has halved. The goal of this analysis is to investigate whether student who got the answer correct and student who got the answer wrong have different cGPA or not.

Wrangling the data

```
# read data
cgpa_data <- read_excel("data/sta303-mini-portfolio-poverty.xlsx") %>%
  janitor::clean_names()

# renaming variables
cgpa_data$global_poverty_ans <-
↪  cgpa_data$in_the_last_20_years_the_proportion_of_the_world_population_living_in_extreme_poverty_

cgpa_data$cgpa <-
↪  cgpa_data$what_is_your_c_gpa_at_u_of_t_if_you_dont_want_to_answer_you_can_put_a_0

cgpa_data <- subset(cgpa_data, select = c(fake_student_id, global_poverty_ans, cgpa) )

# keep appropriate gpa
cgpa_data <- cgpa_data[!(cgpa_data$cgpa < 0 | cgpa_data$cgpa >4),]
cgpa_data <- na.omit(cgpa_data)
```

```
# new variable correct
cgpa_data$correct <- ifelse(cgpa_data$global_poverty_ans == "Halved", TRUE, FALSE)
```

Visualizing the data

```
# visualize correct vs. incorrect
ggplot(cgpa_data, aes(cgpa)) + geom_histogram(fill = "mediumpurple", color =
↪  "black")+facet_wrap(~correct, ncol = 1)+ theme_minimal() + scale_fill_brewer()
```

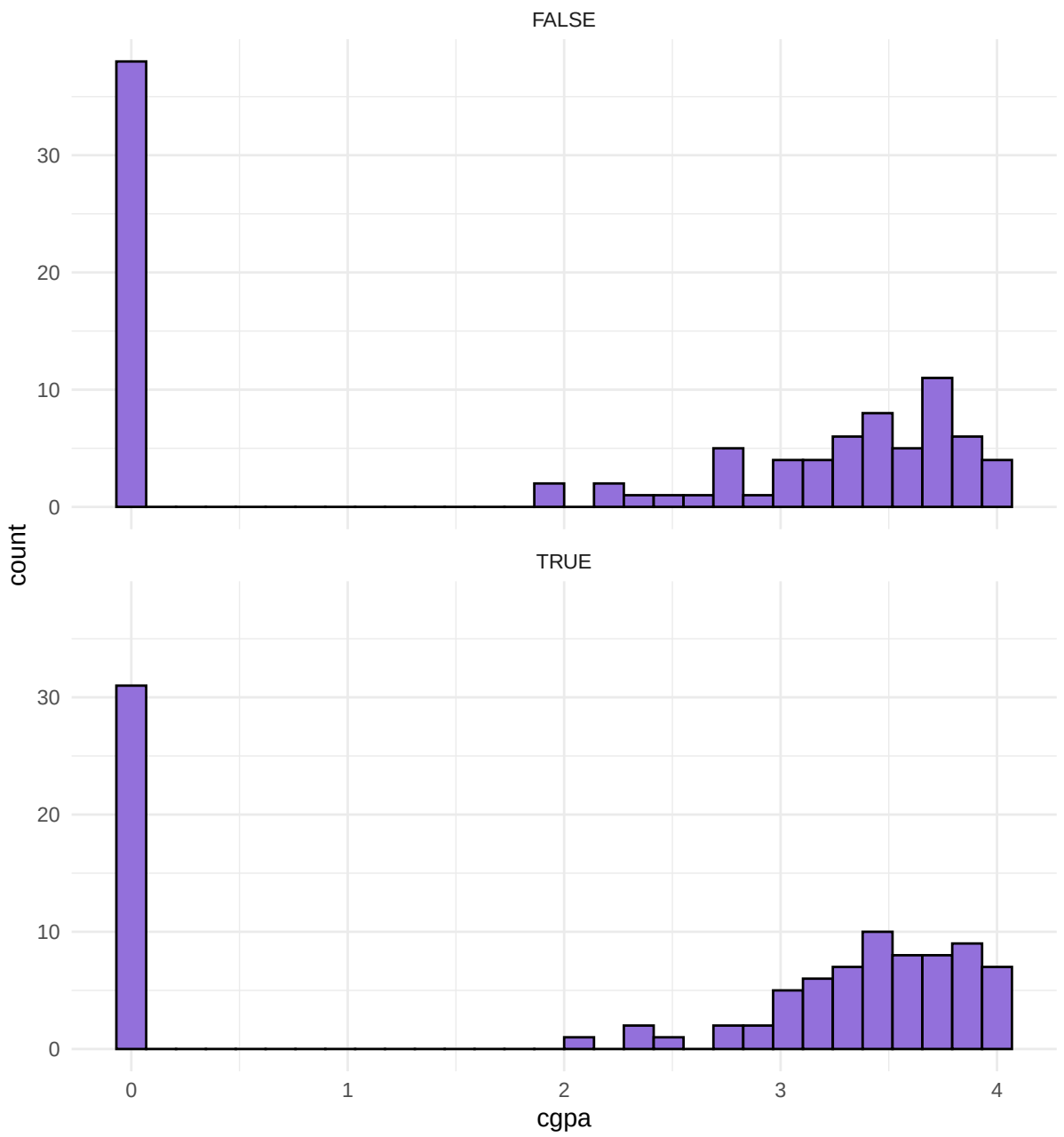


Figure 3: Distribution of cGPA for students who got the answer incorrect and correct

Testing

The Mann-Whitney U test is a non-parametric test of comparing two independent samples and do not need the normality assumption. Here, since the data is not normally distributed and is left skewed, Mann-Whitney U test is the most appropriate in this context. In addition, the

test will be two sided, with the null hypothesis being that the cGPA of students who got the right answer is the same as the cGPA of students who got the wrong answer. The alternative hypothesis is that the two groups of students did not have the same cGPA.

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

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  cgpa by correct
## W = 4355.5, p-value = 0.1675
## alternative hypothesis: true location shift is not equal to 0
```

```
# using lm()
summary(lm(rank(cgpa) ~ correct, data = cgpa_data))
```

```
##
## Call:
## lm(formula = rank(cgpa) ~ correct, data = cgpa_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -70.00 -58.99   0.75  48.00 102.50
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   93.995      5.622  16.720   <2e-16 ***
## correctTRUE   11.010      7.950   1.385    0.168
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 55.94 on 196 degrees of freedom
## Multiple R-squared:  0.00969,    Adjusted R-squared:  0.004637
## F-statistic: 1.918 on 1 and 196 DF,  p-value: 0.1677
```

The p-value is 0.1677, which means that under the null hypothesis that students with the right and wrong answer have similar cGPA, there is a 16.7% probability that we can observe our data.

Using a significance of 0.05, we could not reject the null hypothesis; students who got the correct answer have similar cGPA with students who got the wrong answer.

Writing sample

Introduction

I believe that I am an excellent candidate for the data scientist job that Yelp has posted. I have excellent communication skills communicating with people from different fields demonstrated in past internships and project experiences. In addition, I have over three years of data analysis and visualization using R and Python and one year of experience using SQL. I already have some basic statistical inferencing skills and am currently strengthening my inferencing skill through taking statistics courses.

Soft skills

The company seeks to find a person that not only knows how to communicate with others but also works with people from different teams including colleagues from engineering, product, and business teams.

I have had multiple experiences communicating with interdisciplinary teams in internships and in group projects. For instance, as a Research Analyst Intern at Mthree, I have worked with colleagues studying Human Resource Management, Law, and Social Sciences. My team and I were complimented by the manager for our excellent communication skills when discussing our research findings and acknowledging that all of us are from different backgrounds.

Analytic skills

Yelp would like a person that knows how to use SQL and Python or R for data analysis and also data visualization. They also are looking for a candidate that understands statistical inference, experimental design and analysis well.

I have approximately three years of experience using Python and R, and 1 year of experience using SQL. In most of the statistics courses that I have taken in my undergraduate studies such as the data visualization course and data analysis course, I utilized R to do the course projects and received A or A- for the courses. I have also taken courses from Computer Science Department such as Machine Learning Course and have utilized Python libraries including Matplotlib to visualize the results and received A- for the course. Moreover, I have learned SQL across multiple Coursera courses and have done a few projects in the courses, which makes me familiar with SQL.

Connection to studies

I would like to further develop my data analysis skills and statistical inference skills during the remainder of my education. This semester, I am taking the method of data analysis course as well as the applied bayesian statistics course. By taking the method of data analysis course, I could learn more types of models and analysis methods such as logistic regression in order to complement my previous knowledge in linear regression. In addition, the applied Bayesian statistics course allowed me to have a different approach to analyzing data and making statistical inferences, which can complement my knowledge in the frequentist approach.

Conclusion

I believe my previous experience, data analysis and visualization skills, as well as strong communication skills, would make me a strong fit as a Data Scientist at Yelp. As a Data Scientist at Yelp, I would like to apply my communication and data analysis skills, while continue developing my statistical inferencing skills.

Word count: 479 words

Reflection

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

One thing that I'm proud of is that I became more familiar with plotting graphs using ggplot2. I have plotted graphs using ggplot2 in my Data Visualization course and have explored multiple simple yet creative ways of visualizing data, and this assignment extends my knowledge by giving me the opportunity to know how to have more customization on the caption and legend. It also provides me the opportunity to know a new graph, which is the error bar. Furthermore, I do not know that I can add caption by adding fig.title to the r chunck, and this method is really useful to me as an applied statistician wanna-be in order to write paper in rmd file. I enjoyed applying my ggplot2 skills leared in data visualization course and keep extending my knowledge.

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

I might apply my data wrangling skills, data visualization skills, and communication skills in future work. I will apply effective methods to wrangle and clean the data in order to analyze it. As Professor Bolton mentioned, many data analyst or data scientists spent 70-80% of their time preparing and cleaning data for analysis, so I definitely think that knowing and applying effective way to visualize data is very important. In addition, since I would like to be a data scientist or a biostatistician, talking to people that do not have a solid statisitcal background might be common in my daily work. Thus, I might have to visualize my data so that it is easily interpretable for audience without statistics background. In addition to data visualization, I would also apply my communication skills by explaining hard concepts or terms in an easy to know fashion so that even audience without statistics background will understand.

What is something I'd do differently next time?

Something that I would do differently next time is to make sure I am familiar with the course content up to date before starting my assignment. I did finish going through all of Module 1 and 2 and attend all lectures before starting, but I am not that familiar with the student grades case study. In the case study, there are some useful methods to clean the data. Since I am not familiar enough with it, I often found myself performing a task in a rather clumsy way. Then I will start going back to the case study and found out that only one or two lines of code is needed to accomplish the task. This is a waste of time, and I definitely should be more familiar with all materials before starting the assignment.

All filler text sourced from: Hipster Ipsum