STA303 Mini-portfolio

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

Xiaotang Zhou

2022-02-03

Contents

List of Figures

Introduction

In this mini-portfolio, transferable skills and methods are showcased across the three main sections of Statistical Skills Sample, Writing Sample, and Reflection. In each of these sections, both statistical and communications skills are applied and demonstrated.

In the Statistical Skills Sample section of this mini-portfolio, focus is put on setting up and formatting a document that not only examines several statistical methods, but can also be reproduced and easily understood by both a statistically proficient and non-statistically proficient audience. These methods include loading the appropriate libraries and modules for use throughout the course of the document, examining the effects of changing parameters on a random variable, creating and properly interpreting 95% confidence intervals, and determining and testing whether evidence for an association between two variables exists.

Next, in the Writing Sample section, an analysis of soft skills and analytic skills, such as proficiency in the desired coding languages of R, Python, and SQL, ability to communicate with both statistical and non-statistical audiences through clean, reproducible code, and an understanding of experimental design and application of statistical tests and techniques, that would be suitable for a data scientist position at Yelp! is conducted.

Finally, in the Reflection section, the accomplishments, possible applications, and possible shortcomings and improvements that were realized during the creation of this mini-portfolio are discussed.

Statistical skills sample

Setting up libraries

```
library(tidyverse)
library(readxl)
```

Visualizing the variance of a Binomial random variable for varying proportions

```
n1 <- 7 # assign an appropriate value for n1
n2 <- 42 # assign an appropriate value for n2
props <- seq(0, 1, 0.01) # generate a sequence of proportions
p_1_p <- props * (1 - props) # calculate the p(1 - p) part for the variance
for_plot <- tibble(props, n1_var = n1 * p_1_p, n2_var = n2 * p_1_p) # create tibble

# In order from left to right, top to bottom: set x and y data, set plot as
# scatterplot, label x and y axes, add title, add caption
for_plot %>% ggplot(aes(x = props, y = n1_var)) + geom_point() +
  xlab("Proportion") + ylab("Variance") +
  ggtitle("The Effect of Different Proportions on the Variance of a Binomial
  Random Variable With a Fixed Sample Size of 7") + theme_minimal() +
  labs(caption="Created by Xiaotang Zhou in STA303/1002, Winter 2022")
```

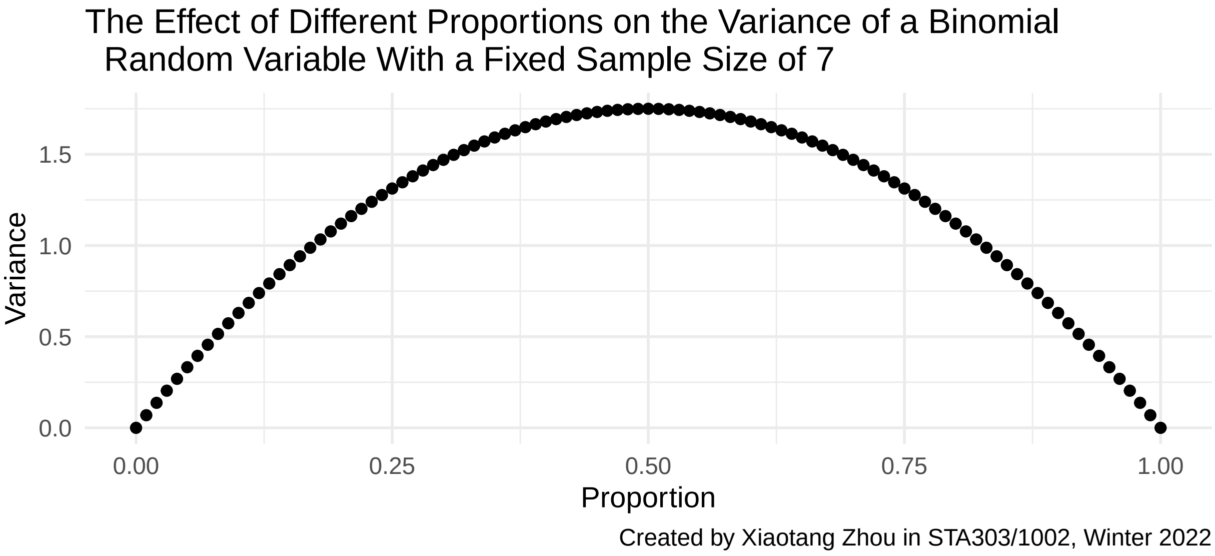


Figure 1: Relationship of Proportion and Variance of a Binomial Random Variable (n = 7)

```
# In order from top to bottom, left to right: set x and y data, set plot as
# scatterplot, label x and y axes, add title, add caption
for_plot %>% ggplot(aes(x = props, y = n2_var)) + geom_point() +
  xlab("Proportion") + ylab("Variance") + theme_minimal() +
  ggtitle("The Effect of Different Proportions on the Variance
  of a Binomial Random Variable With a Fixed Sample Size of 42") +
  labs(caption="Created by Xiaotang Zhou in STA303/1002, Winter 2022")
```

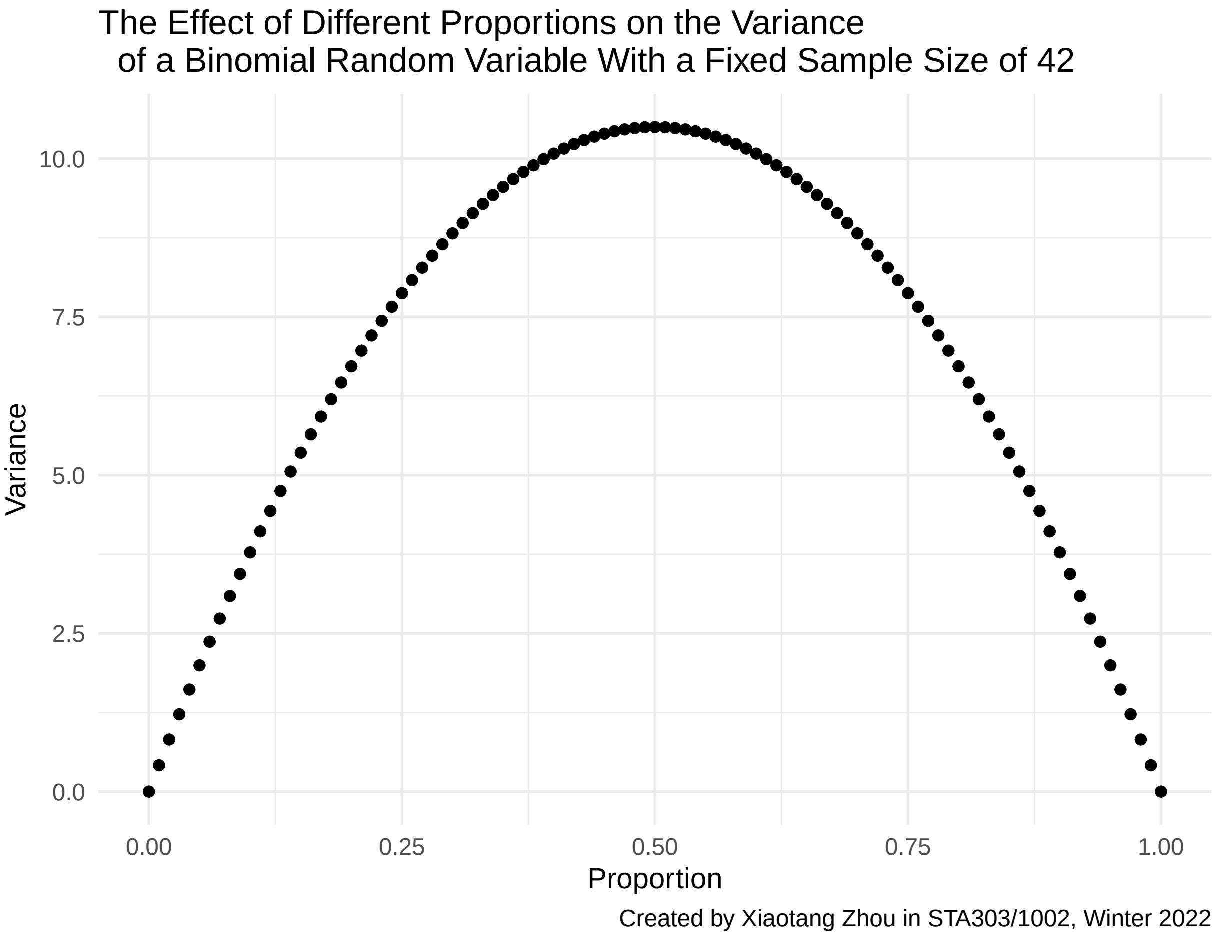


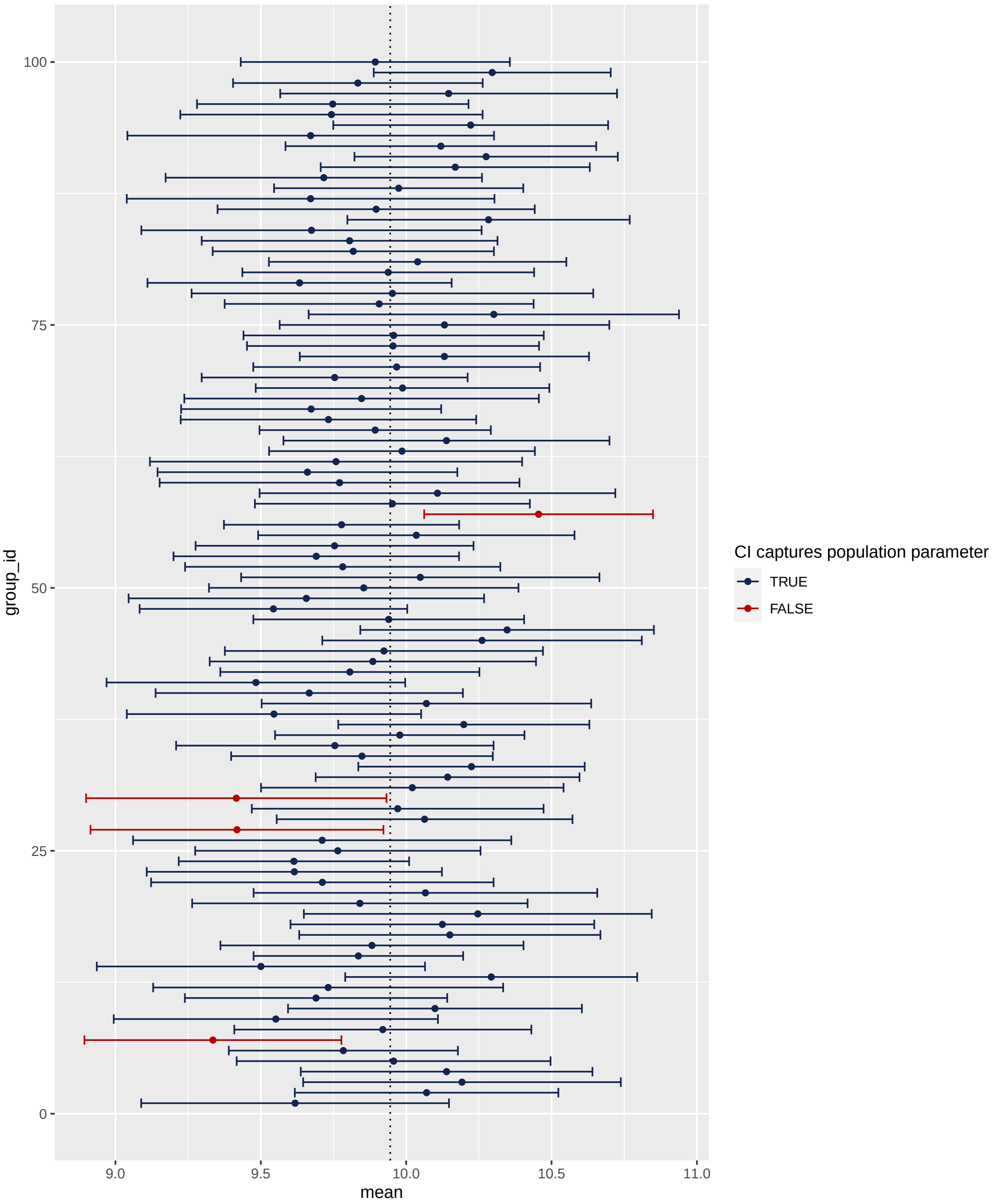
Figure 2: Relationship of Proportion and Variance of a Binomial Random Variable (n = 42)

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

```r
# set the seed to preserve randomness
set.seed(859)

# create the sim_mean object and assign 10 (the mean of the desired Normal
# distribution) to it
sim_mean <- 10

# create the sim_sd object and assign sqrt(2) (the standard deviation of the desired
# Normal distribution) to it
sim_sd <- sqrt(2)

# create the sample_size object and assign 30 (the desired sample size) to it
sample_size <- 30

# create the number_of_samples object and assign 100 (the desired number of samples)
# to it
number_of_samples <- 100

# create the tmult object and assign the desired t-multiplier to it
tmult <- qt(p=0.975, df=sample_size - 1)

# create the population object and assign the vector of 1000 simulated values from
# the N(10, 2) distribution to it
population <- rnorm(n = 1000, mean = sim_mean, sd = sim_sd)

# create the pop_param object and assign the true mean of the population object
# to it
pop_param <- mean(population)

# create the sample_set object and assign all of the sampled data to it
sample_set <- unlist(lapply(1:number_of_samples, function (x) sample(population, size
↪   = sample_size)))

# create the group_id object, which will help us distinguish one sample from another
group_id <- rep(1:number_of_samples, each = sample_size)

# create the my_sim tibble with columns group_id and sample_set
my_sim <- tibble(group_id, sample_set)
```

```r
# create the ci_vals tibble where each sample is given 1 row and their sample mean
# and sample standard deviation is recorded
ci_vals <- my_sim %>%
  group_by(group_id) %>%
  summarise(mean = mean(sample_set), sd = sd(sample_set))

# create the lower column in ci_vals by assigning each entry its corresponding
# sample's derived lower bound in the confidence interval
ci_vals$lower <- ci_vals$mean - tmult * (ci_vals$sd / sqrt(sample_size))

# create the upper column in ci_vals by assigning each entry its corresponding
# sample's derived upper bound in the confidence interval
ci_vals$upper <- ci_vals$mean + tmult * (ci_vals$sd / sqrt(sample_size))

# create the capture column in ci_vals by assigning each entry its corresponding
# sample's status on whether the confidence interval derived through the sample
# captures pop_param or not
ci_vals$capture <- (ci_vals$lower < pop_param & pop_param < ci_vals$upper)

# create the proportion_capture variable by assigning the proportion of confidence
# intervals that capture pop_param to it
proportion_capture <- sum(ci_vals$capture) / number_of_samples
```

```r
# In order from top to bottom: set data and x and y axes, set plot as
# scatterplot, include error_bar aesthetic, flip plot horizontally, impose dashed
# line denoting true population parameter, adjust colour and legend schemes, add
# caption
ggplot(ci_vals, aes(group_id, mean, color = capture)) +
  geom_point() +
  geom_errorbar(aes(ymin = lower, ymax = upper)) +
  coord_flip() +
  geom_hline(yintercept=c(pop_param), linetype="dotted") +
  scale_color_manual(name ="CI captures population parameter", values = c("TRUE" =
  ↪  "#122451", "FALSE" = "#B80000")) +
  labs(caption="Created by Xiaotang Zhou in STA303/1002, Winter 2022")
```



Created by Xiaotang Zhou 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)

96% of my intervals capture the population parameter

In the previous plot (Figure 3), we can see that there is a vertical dash line that indicates where the population mean is. In this situation, we are able to draw this line because we actually knew what the mean of the population is since all the data was simulated, which in other words just means that we have the ability to put this line on the plot for the simple reason that we actually know where to draw it. This is obviously not applicable in practice since in most situations the population mean is unknown as it is not feasible or sometimes even impossible to gather every observation in the population in order to find the mean, which forces us to sample from the population and try to generate plausible estimates of the population mean using confidence intervals instead.

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 section is to find out whether there is evidence that the average cumulative grade point average (cGPA) of Winter 2022 STA303 students who answered the question on global poverty correctly is different from the average cGPA of Winter 2022 STA303 students who answered the question on global poverty INcorrectly, subject to the condition that both groups of students chose to disclose their cGPAs. In other words, we are interested in finding out if there is any association between a student's cGPA and whether they answered the question correctly or not.

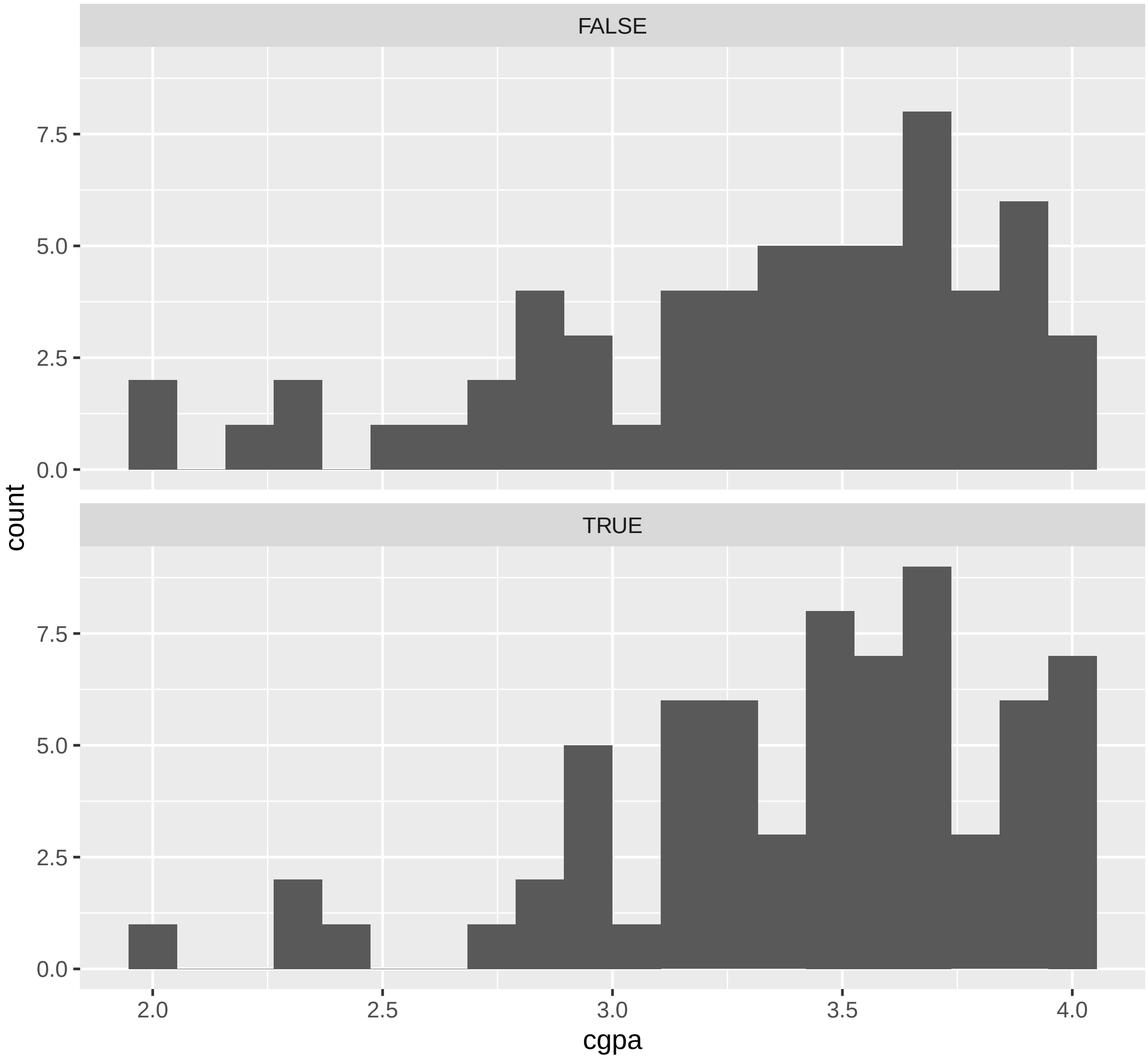
Wrangling the data

```r
# read in the data from the Excel file and save it to cgpa_data. Then, after
# calling the clean_names() function from the janitor library, rename the
# corresponding columns to cgpa and global_poverty_ans
cgpa_data <- read_excel("data/sta303-mini-portfolio-poverty.xlsx") %>%
  janitor::clean_names() %>%
  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)

# keep only the cgpa and global_poverty_ans columns as only they are needed to
# answer the desired question
```

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

# keep only rows where cgpa is positive (we don't include rows where the cgpa
# value is 0 because these rows represent people who did not want to disclose
# their cGPA) and where cgpa is 4 (the maximum possible value) or less
cgpa_data <- subset(cgpa_data, cgpa > 0 & cgpa <= 4)

# create the correct column in cgpa_data by assigning each entry its corresponding
# global_poverty_ans value's status on whether the question on global poverty was
# answered correctly or not
cgpa_data$correct <- ifelse(cgpa_data$global_poverty_ans == "Halved", TRUE, FALSE)
```

Visualizing the data

```
# In order from top to bottom: feed cgpa_data to ggplot command, set the aesthetic
# (the horizontal axis) to the cgpa variable, set plot as histogram, call facet_wrap
# in order to layer histograms showing the distributions of the cgpa variable for
# both correct = TRUE and correct = FALSE on top of one another
ggplot(cgpa_data, aes(cgpa)) +
  geom_histogram(bins = 20) +
  facet_wrap(~correct, ncol = 1)
```



Testing

Looking at the above distributions of the `cgpa` variable for both values of the binary `correct` variable, we can see that neither distribution roughly follows a Normal distribution. This means we cannot use the usual parametric two-sample t-test, and we must resort to its non-parametric alternative, the Mann-Whitney U test, instead.

```
# First, we use the wilcoxon.test function to execute the 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
```

```
# Next, we use the lm function to get the same p-value/result:
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
```

From the two blocks of output above, we can see that the p-value is 0.35, and since $0.35 > 0.05$ we fail to reject the null hypothesis that there is no association between a Winter 2022 STA303 student's cGPA (on the condition that they provided their cGPA) and whether they answered the question on global poverty correctly. This means there is NOT statistically significant evidence that there is an association between these two variables.

Writing sample

Introduction

In this job advertisement for a data scientist position at Yelp!, having some of the analytic and soft skills that are listed under the “We Are Looking For” section would strongly bolster one’s application for this role. In particular, the soft skills of being an effective group member in a team and proficiency in communicating seemingly complicated ideas to a lay audience and the analytic skills of knowledge of a wide range of statistical practices and methods and proficiency in the programming languages of R, SQL, and Python would help an application to this role substantially.

Soft skills

For me, two soft skills that I possess include the ability to contribute effectively to a team and the ability to communicate complex ideas to a lay audience. Evidence of this comes both from courses I have taken in the past and many of the projects that I have done both individually and in a team. For example, my previous experience in creating reproducible documents in R, Python, and SQL from courses I have taken highlights my ability to take a complex idea and distill it into something that someone not well-versed on it can understand, while group projects such as the creation of a computer game and a conference-managing app demonstrate my ability to work in and effectively contribute to a team effort.

Analytic skills

Next, two analytic skills that I possess include the knowledge of a diverse range of statistical practices and methods and the ability to code in R, SQL, and Python. Evidence of this comes again from courses I have taken previously and many of the projects that I have done both individually and in a team. For example, my previous experience in creating reproducible documents in R, Python, and SQL from previous courses shows my knowledge of several different statistical practices and methods while also, in addition to projects such as an NHL database analysis in SQL, highlighting my proficiency in these languages.

Connection to studies

In addition to the previously mentioned soft and analytic skills, another very important skill that could be developed to further strengthen an application to this role is the ability to think creatively to solve unique and unusual problems because in a data-driven environment, it is often important to know how to adapt to rapid change. One of the ways to bolster this and the previously mentioned skills during education is to participate in research opportunities, as the experience gained during these opportunities will be extremely valuable in setting one up for success in the future.

Conclusion

All in all, the soft skills of being an effective group member in a team and having the ability to communicate complicated ideas to a lay audience and the analytic skills of knowledge of a wide range of statistical practices and proficiency in R, SQL, and Python are very important to someone looking to apply to the posted role. In addition, creative thinking is also an important skill that will set up an applicant to this job and others like it for success.

Word count: 500 words

Reflection

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

One thing that I am proud of in this mini-portfolio is the dramatic improvement I have personally seen in my ability to write clean and readable R code compared to work that I have done in the past. As someone who usually prefers to code in Python as I feel it is syntactically cleaner, I think working on this mini-portfolio has helped me see how R code can also be written in a clean and interpretable way. One of the main ways in which I feel I improved my skills in writing R code that does what is required but is also easy for others to read is in my improvement in my ability to explain exactly what is going on in the code through the use of comments. By writing comments that explain to the reader what the code is doing, I feel that I implicitly coerced myself into making the code as clean and concise as possible so that it would be easily explainable.

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

There are many ways in which I feel I can apply what I've learned and demonstrated while working on this mini-portfolio in future work and study. For example, the skill of creating and formatting reproducible documents that can be easily read by both statistically versed and non-statistically versed people will be extremely important to demonstrating experience in statistical and computational literacy. Also, the skill of being able to communicate statistically-drawn conclusions to a lay audience will be essential to ensuring that the ideas obtained through any statistical analysis can be easily understood by those who may not understand all of the technical jargon.

What is something I'd do differently next time?

If I could do this mini-portfolio all over again, one thing I would do differently is to completely finish one task before moving on to another one, as over the course of working on this mini-portfolio I found myself constantly flipping between multiple tasks after getting bogged down while trying to focus my attention on a specific one, which at the time seemed like a good idea but in retrospect probably did more harm than good. The primary reason I feel this way about this approach is that constantly changing which task I was working on started to create a lot of confusion around what I needed to code and write to the point where I was making extremely egregious errors, such as running a simulation where one wasn't needed or assigning a variable in the wrong code chunk, which probably wasted more time than actually spending all the time necessary to figure out the problem surrounding a task and figuring out how to solve it before moving on to another task.