STA303 Mini-portfolio

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

[Purumidha Sharma]

2022-02-03

Contents

List of Figures

Introduction

In this mini-portfolio, I have learnt to thoughtfully communicate my understanding of statistical concepts and make decisions on using the best hypothesis testing for a given scenario. I have learnt new R coding functions in this STA303 class that optimized my time spent on making R coding efficient and neat. To give a little context about STA303, this course is a Data Analysis II course for third year students which focuses heavily on technical coding and refining writing and verbal skills. This is helpful in demonstrating skills such as collaboration, and communication. The learning objectives of this course is to teach students about data exploratory analysis, not limited to, wrangling, data visualization, and data cleansing and investigating ethical considerations in data analysis. Theoretically speaking now, the need to understand assumptions and the usage cases for linear mixed models, generalized linear mixed models, and additive models. Although these words are too technical, to sum it up, the crux of this course is to well-verse with various scenarios to be able to R code, and apply data analytical tools learnt in this course for future settings. This is so we can accurately and appropriately interpret results of different models, and being able to communicate to a larger general audience.

Statistical skills sample

Setting up libraries

```
# setup libraries
library(readxl) # readxl is for reading .xlsx data file

library(tidyverse) # tidyverse is a very useful package for data analysis in R (covers
↪  almost all you need to analyze a dataset)
```

Visualizing the variance of a Binomial random variable for varying proportions

```
library(ggplot2) # set up library for ggplot (useful for data visualization,
↪  incorporates all types of visual graphs)

n1 <- 10
n2 <- 100
props <- (seq(0,1, by = 0.01)) #  makes regular sequences of numbers (remember c() is
↪  not the only way to make a vector, even seq() does the same thing but also takes
↪  in a parameter for counting by (increment of the sequence))

# Create a dataframe using tibble()
for_plot <- tibble(props, n1_var = (n1*props)*(1- props) , n2_var = (n2*props)*(1-
↪  props))


# create a line plot using geom_line() (geom_line connects observations of data from
↪  left to right based on the variable on x-axis)
for_plot %>% ggplot(aes(x = props, y = n1_var)) + geom_line() + labs(title = "For n1
↪  =10, Distribution of proportions \n for variance", caption = "Created by Purumidha
↪  Sharma in STA303/1002, Winter 2022", x = "proportions", y = "variance where n1 =
↪  10") + theme_minimal()
```

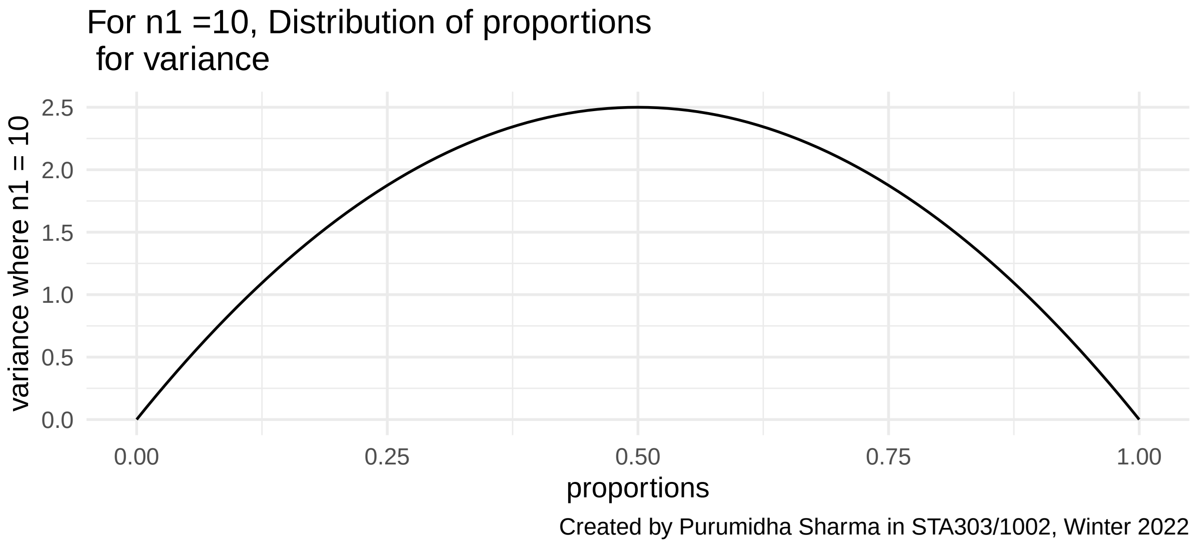


Figure 1: From our chosen n1 value = 10, this figure shows how the parabola formed maximizes at proportion = 0.5, since we want to find the largest variance when p = 0.5 for Binomial Random Variable. Here maximum variance is 2.5

```
# create a line plot using geom_line() (geom_line connects observations of data from
↪  left to right based on the variable on x-axis)

for_plot %>% ggplot(aes(x = props, y = n2_var)) + geom_line() + labs(title ="For n2
↪  =100, Distribution of proportions \n for variance", caption = "Created by
↪  Purumidha Sharma in STA303/1002, Winter 2022", x="proportions", y = "variance
↪  where n2 =100") + theme_minimal()
```

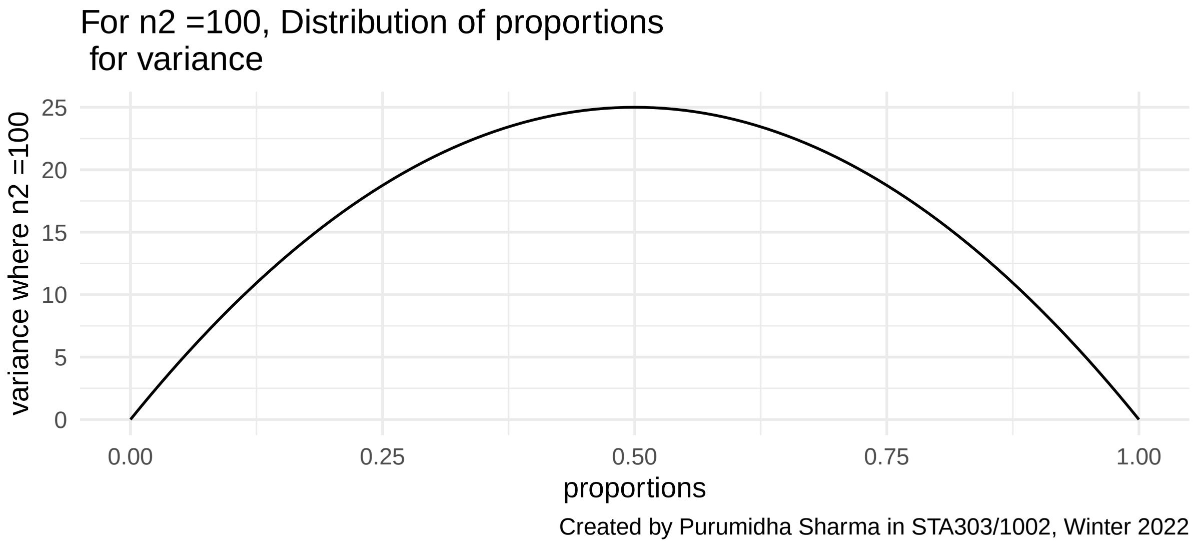


Figure 2: From our chosen n2 value = 100, this figure shows how the parabola formed maximizes at proportion = 0.5, since we want to find the largest variance when p = 0.5 for Binomial Random Variable. Here maximum variance is 25.

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

```r
set.seed(198)
number_of_samples <- 100
sim_mean <- 10
sim_sd <- sqrt(2) # 2 is variance in N(10,2)
sample_size <-  30
tmult <- qt(p = (1-0.95)/2, df = sample_size - 1, lower.tail = FALSE) # constructing a
↪  95% confidence interval (two side tail)


# Create a vector called population (simulated population) that uses simulated mean
↪  and simulated standard deviation where our number of observations (n) here is
↪  1000. Also, rnorm is the function to generate random numbers from a normal
↪  distribution.
population <- rnorm(n = 1000, mean = sim_mean, sd = sim_sd)

# the true (actual) mean for our population
pop_param <- mean(population)


# extract 100 samples of size 30 from our population
sample_set <- unlist(lapply(1:number_of_samples,
    function (x) sample(population, size = sample_size)))

# labeling the values from the 100 different samples above. By using rep() we
↪  replicate the values from 1 to the number of samples (shown below) 30 times (30 =
↪  sample_size)
group_id <- rep(1:number_of_samples, rep(sample_size,number_of_samples))

# create a dataframe of two columns
my_sim <- tibble(group_id, sample_set)

# dataframe to note confidence interval values
ci_vals <- tibble(group_by(my_sim, group_id) %>%
summarise(mean = mean(sample_set), sd = sd(sample_set))) %>%

# mutate() creates a new variable
mutate(lower = mean - tmult*sd/sqrt(sample_size), upper = mean +
↪  tmult*sd/sqrt(sample_size), capture = ifelse(pop_param >= lower & pop_param <=
↪  upper, TRUE, FALSE))
```

```
# store the proportion of intervals to capture population parameter
proportion_capture <- sum(ci_vals$capture) / number_of_samples


 # creating an error bar graph using confidence interval values to see the precision
 ↪  of mean using confidence interval values (lower and upper bound)
ggplot(data = ci_vals,mapping = aes(colour = capture)) + geom_errorbar(data=ci_vals,
↪  mapping=aes(ymin = lower, ymax = upper, y = mean, x = group_id)) +
↪  scale_color_manual("CI captures population parameter", values =
↪  c("#B80000","#122451"))  +
  geom_hline(yintercept = pop_param, linetype = "dashed") +
  geom_point(mapping = aes(y = mean, x = group_id), size = 1.5) + labs(caption =
↪  "Created by Purumidha Sharma in STA303/1002, Winter 2022",x = NULL, y = NULL) +
↪  theme(legend.position = "bottom")  + coord_flip()
```

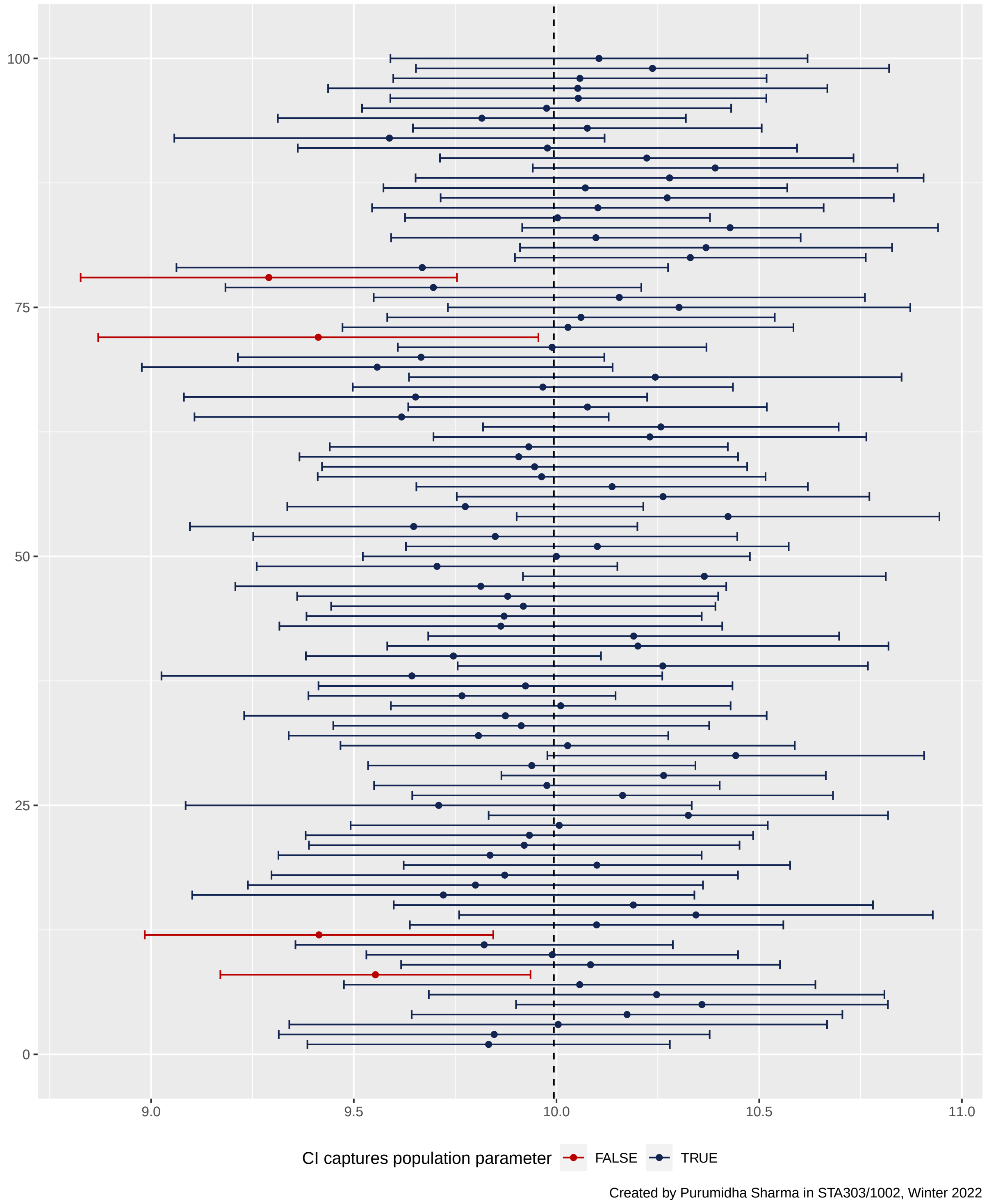


Figure 3: Exploring our long-run "confidence" in confidence intervals. This figure shows how often 95% confidence intervals from 100 simple random samples capture the population mean. The population was simulated from N(10, 2)

96 % of my intervals capture the population parameter.

The reason we can include the population parameter in this plot is because we are just seeing what random variation from the mean looks like, and in our case, we are given a population parameter, and we are generating samples from it, which is not the case generally.

Generally speaking, we cannot usually compare the population parameter to our confidence interval in practice (when working with data that has not been simulated) because true values we interested in (also called population statistics) is hidden from us. In real world setting, all you have access to is a dataset which will not give mean or variance or distribution that generated it. Hence, the job of confidence interval is to give "the likely range" of values which the mean and variance can take. Instead, we center the confidence interval around the sample mean and sample variance, which are statistics and therefore functions of the dataset and not parameters. The whole point of classical parametric statistics is that you're trying to guess what the parameters are using statistics.

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

Goal

Firstly, we are interested in finding whether the factor "correct" which maps TRUE/FALSE values to those students who answered the question correctly or wrong. (The question is: Whether the proportion of people living below the global poverty line had halved, doubled or stayed about the same in the last 20 years.)

The goal of this task is to test whether whether the factor "correct" impacts the cGPA between students who correctly answer the question or not (as mentioned above). This will be done by choosing which hypothesis test is the best to use to investigate if either of the two groups respond in a similar way with the effect of cGPA of STA303 students.

Wrangling the data

```
# Wrangle!

# first loading the excel file to be able to read it!
cgpa_data <-
↪  read_excel("~/sta303-w22-mini-portfolio/data/sta303-mini-portfolio-poverty.xlsx",
↪  sheet = 1) %>%
# clean the data using the janitor package, in which clean_names() function makes all
↪   the names consistent (making all words lowercase, creating spaces between
↪   characters/words replaced with underscore, or special characters being removed or
↪   simplifying it)
janitor::clean_names() %>%

#renames the column variable (to the left value)
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)
↪   %>% filter(!is.na(cgpa)) %>% filter(cgpa > 0.0 & cgpa <= 4.0) %>%   # filter
↪   removes values specific to your condition inside the parameter of filter()


  # creates a new column name with the following condition:
 mutate(correct = case_when(
str_detect(global_poverty_ans, "Halved") ~ TRUE, (str_detect(global_poverty_ans,
↪   "Doubled") ~ FALSE), (str_detect(global_poverty_ans, "Stayed about the same") ~
↪   FALSE)))
```

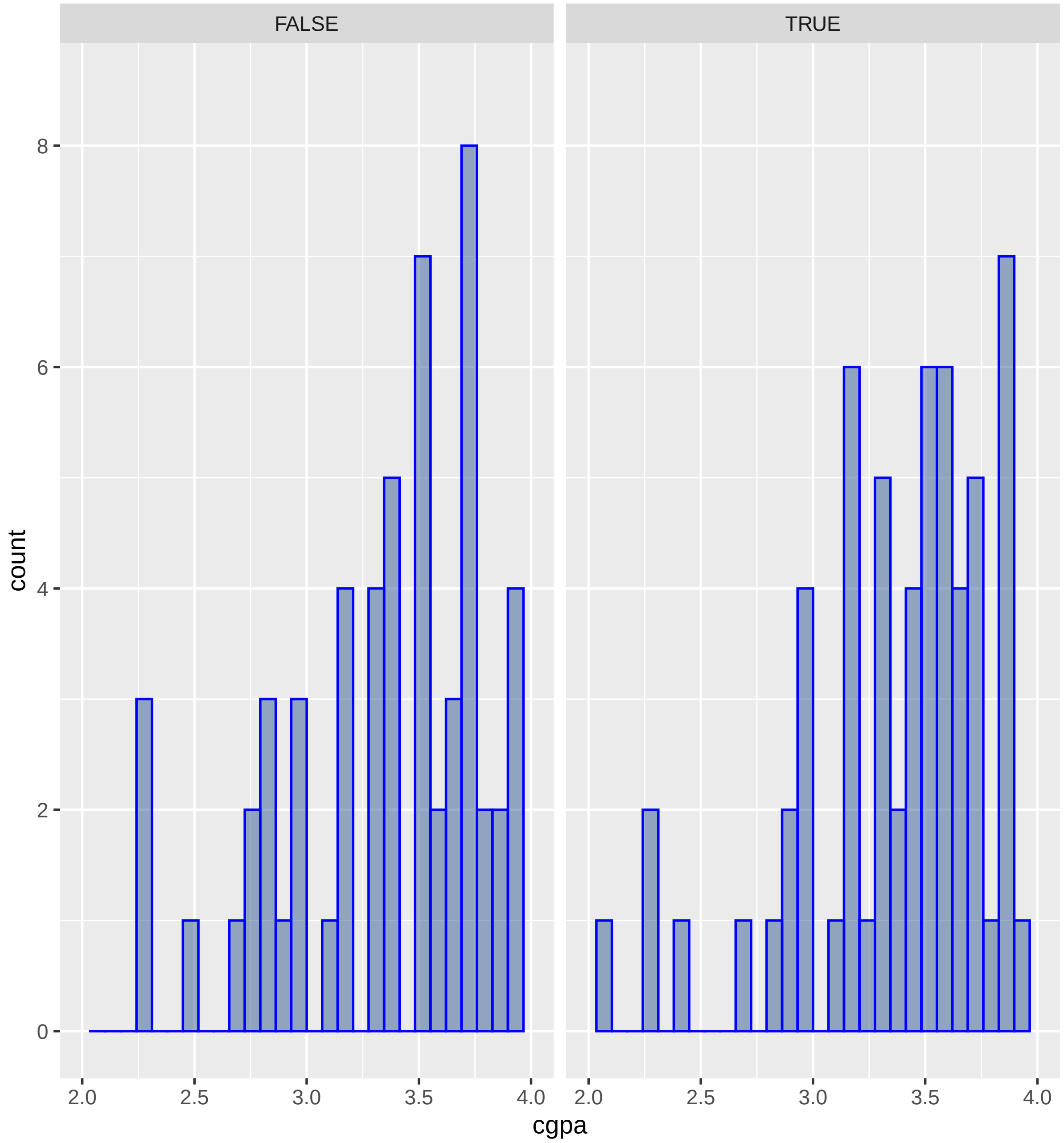
Please note: The reason we remove observations with 0 cGPA is because, during the student survey in STA303 class, 0 signifies those students who did not want to share their cGPA. These observations will not help us in any analysis since we are investigating the difference between cGPA students who correctly answer the global poverty line question and those who do not. Thereby, 0 will not help us, and thereby we exclude the students who put a 0 cGPA.

Visualizing the data

```
# Visualize!

# Create a set of histograms, in one figure, positioned on top of each other, that
↪   will allow you examine the data in a useful way.
cgpa_data %>%
ggplot(aes(x = cgpa)) +
geom_histogram(alpha = 0.5, fill="#375E97", col="blue") +
scale_x_continuous(limits = c(min(cgpa_data$cgpa), max(cgpa_data$cgpa))) +
↪   facet_wrap(~correct) + coord_cartesian(ylim = c(0, 8.5))  # using facet_wrap
↪   splits graphs by the groups for the variable, in our case, (true and false) for
↪   the "correct" variable
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 4 rows containing missing values (geom_bar).
```



Testing

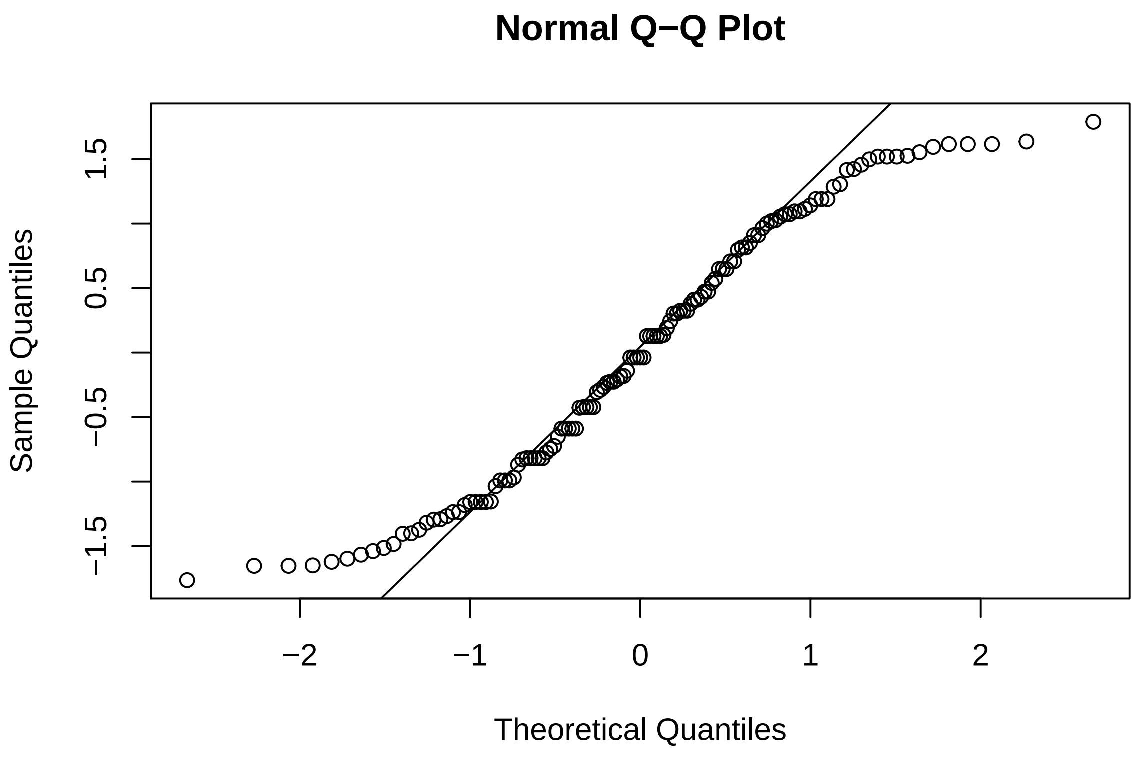
```
# Test!

# Choose an appropriate test to test whether there is an association between cGPA and
↪  if a student in STA303/1002 answered this question correctly

# This linear model is equivalent to Mann Whitney U-test
mod = lm(rank(cgpa)~correct, data = cgpa_data)
summary(mod)
```

```
##
## 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
```

```
# Mann-Whitney U
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
```

```
# this is the Normal Q-Q plot to see if the normality assumption of our model is met
↪  or not
r <- rstudent(mod)
qqnorm(r)
qqline(r)
```



To interpret our model above, we have applied an equivalent way of writing a linear model using the lm() regression method to get our p-value corresponding to F-statistic which is 0.3508. Also, we have performed Mann Whitney U-test which gives a very similar p-value of 0.35. The null hypothesis for Mann Whitney U-test roughly says,the two groups have the same population distribution. Based on this, we know our p-value is not significant (0.35) and there's no evidence against our null hypothesis that the two groups (true and false) have the same population distribution. Overall, this means that the cGPA of students who do answer questions correctly for global poverty line and those who don't, answer in the same way.

Now, to justify on why we used Mann Whitney U-test instead. If we think about t-tests, one of the assumptions is that the data is normally distributed, and the independence assumption is even more important. However, just to get an idea before hand, when we plot the standardized residuals Normal QQ plot, we see none of the residuals fall along the line, the normality assumption is not met because we see many extreme outliers, and many data points are not aligned to the line.

We cannot use Anova for testing for this task because we have 2 groups, and in anova, you need atleast 3 groups or more. For the similar reason,although Kruskal-Wallis Rank Sum test is a non parametric test that doesn't rely on normality assumptions (and is more laid back in assumptions) yet is similar to being an one-way ANOVA test but on the rank transformed (y). The similarity is that you need to deal with 3 or more groups, and since ours has 2 groups, our options drive us to either Mann Whitney U-test or Wilcoxon test. We now know, we need to use a non-parametric test since this data is not normally distributed.

So, we don't use Wilcoxon test here because its mostly used for one group's centre or pairs, we aren't really dealing with that here. Rather, Mann Whitney U-test is a very useful method for our scenario since we are comparing 2 groups. Mann Whitney U-test is the best non-parametric version of the t-test, and is conceptually similar. Since, we originally thought of t-test but because normality assumption was being violated, the non parametric version of t-test is more appropriate. Thus, Mann Whitney U-test is the best test to go for this scenario.

Writing sample

Introduction

As a third-year Statistics major student at the University of Toronto, I love to learn and use various data science skills with an aim to identify business challenges in the real world and enable effective solution to decision-makers using technology to improve operational efficiency. My analytical skills, soft skills, and my inquisitive nature to learn more in the field of data analytics will be an asset to this company.

Soft skills

Introspecting myself, I possess strong verbal and written skills. For example, in one of my "Software Design" course, we learnt Agile and Scrum methodology. Through this, we learnt to summarize in a few sentences about the entire week's progress on the group project on which we were working. We also communicated our project's vision, core concepts and demonstrated technical skills through effective public speaking skills.

Another example I would like to share is about active listening. This skill requires on listening to key words being spoken and capturing that. For example, I am part of a Digital society club, where annual general meeting is held by presidents to speak about dates on organizing events, and tasks to complete for executives like myself. As an active listener, I focus on the dates they emphasize on, the criteria for organizing events and I make a note of it for future purposes. The purpose of active listening is that the intent through which people are trying to convey information is captured within us meaningfully.

Analytic skills

The analytic skills relating to software use and performing data analysis that I possess is SQL, R, and data exploratory tools such as ggplot, and tableau. For example, at University Of Toronto, I worked on a project to investigate if smoking affects systolic blood pressure and if any other factors also influence blood pressure. This project used statistical inference analysis, and relied only on R coding, and using exploratory tools such as ggplot. Lastly, I have learnt SQL through a database course taken at the University of Toronto that helps us learn about relations, and extracting relevant information from databases.

Connection to studies

As an inquisitive person open to learning new technologies/methods, a skill I would like to develop is on using python as a data analysis tool. I will achieve this by taking a crash course, getting a certification and focus on making my own data analytic project by using only Python. Another field that I would like to work on is, networking. The best way to expand my existing network is via Linkedln, and being part of professional groups such as, ToastMasters club. Lastly, being resourceful is important because this is a skill used to overcome problems that may arise, and using what's available to create an effective solution.

Conclusion

In conclusion, I am confident in my demonstration of technical skills, analytical skills, soft skills. I am willing to learn new, connect to deeper/newer technologies and methods to be up to date. This will showcase my true talent and provide a growth for me at a personal and professional level.

Word count: 500 words

Reflection

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

I strongly believe that this mini-portfolio has improved my data exploratory skills, such as, data visualization, data wrangling, choosing the right hypothesis test given the scenario, and lastly, writing skills. In short, I am proud of learning different aspects of data analysis through the medium of mini-portfolio. This will be very helpful for future employment purposes.

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

Majority of a data analyst's routine work goes into organizing data, extracting relevant information, cleaning any missing values, formatting the variables in the right manner, and making sure the data is tidy. These are necessary measures before stepping into the analysis and making decisions based on a well-formed data. The point of explaining this is to show how I have learned how to do those initial steps that marks importance before the analytical and decision step. This demonstrates my methodology adopted for future work and study after STA303/1002.

What is something I'd do differently next time?

We all know, there's always room for improvement, that's the only way we become a better version of ourselves whether personal or career wise. If given a next time for a similar project like this, I would want to explain in detail about when a particular hypothesis testing should be used, i would like to show regression assumptions plots in detail to show how the regression model assumptions are met for different hypothesis testing and why. In short, the "how" and "why" part is what i would like to expand in my work next time. Another thing, since I am newer to the non-parametric tests, I would want to expand my knowledge on when to use the different tests for non-parametric hypothesis testing and why. This will be more helpful for future non-normal data relevant analysis as well.

All filler text sourced from: Hipster Ipsum