STA303 Mini-portfolio

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

Smriti Kannan

2022-02-03

Contents

List of Figures

Introduction

This mini-portfolio highlights some skills across the three main sections of the portfolio: the statistical skills samples, the writing sample and the reflection. The statistical skills samples display various types of data analysis and data visualization knowledge. Some of the skills displayed include wrangling and exploring a data set, creating appropriate visualizations for data sets, creating confidence intervals and performing appropriate tests. Specifically, the final skills sample, investigating whether there is an association between cGPA and answering a question correctly highlights many of the skills listed above. This sample involves wrangling the data and performing appropriate analysis in the form of a statistical test and building a linear regression model.

The writing skills sample portrays some of my other qualifications in the form of addressing a job prompt to be a data scientist. The soft skills and analytical skills for such a job are addressed along with my own qualifications and skills I have for this type of job. It also highlights how my current studies are linked to such a job and highlights the skills I have learned during my university education.

The reflection highlights the strengths of this portfolio, how the skills displayed in this portfolio can be used in future work and study and what I would do differently in this portfolio next time.

Statistical skills sample

Setting up libraries

```r
# Loading the relevant libraries
library(tidyverse)
library(readxl)
```

Visualizing the variance of a Binomial random variable for varying proportions

```r
# Setting up the necessary objects
n1 <- 7
n2 <- 56
props <- seq(0, 1, 0.01)

# Creating a data frame for the values
for_plot <- tibble(props,
                   n1_var = n1 * props * (1-props),
                   n2_var = n2 * props * (1-props))

# Creating a variance plot for n1
for_plot %>%
  ggplot(aes(x=props, y=n1_var))+
  geom_line()+
  theme_minimal()+
  labs(x="Probability of success", y="Variance", caption = "Created by Smriti Kannan
↪   in STA303, Winter 2022")
```

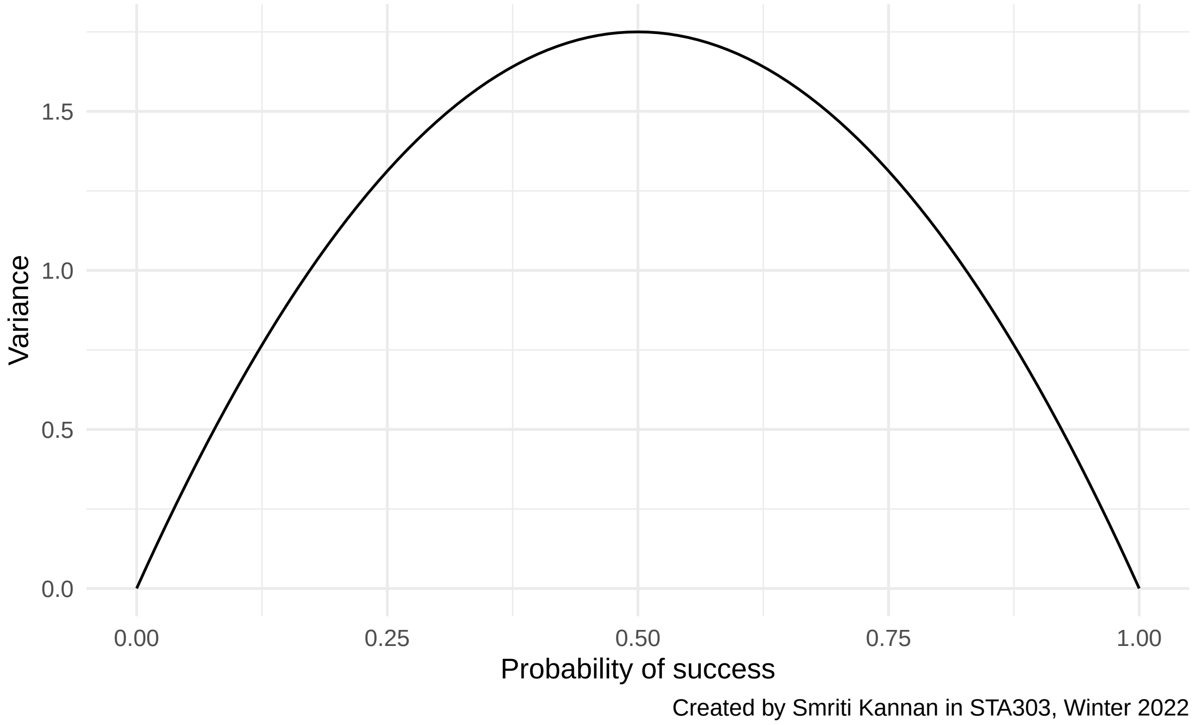
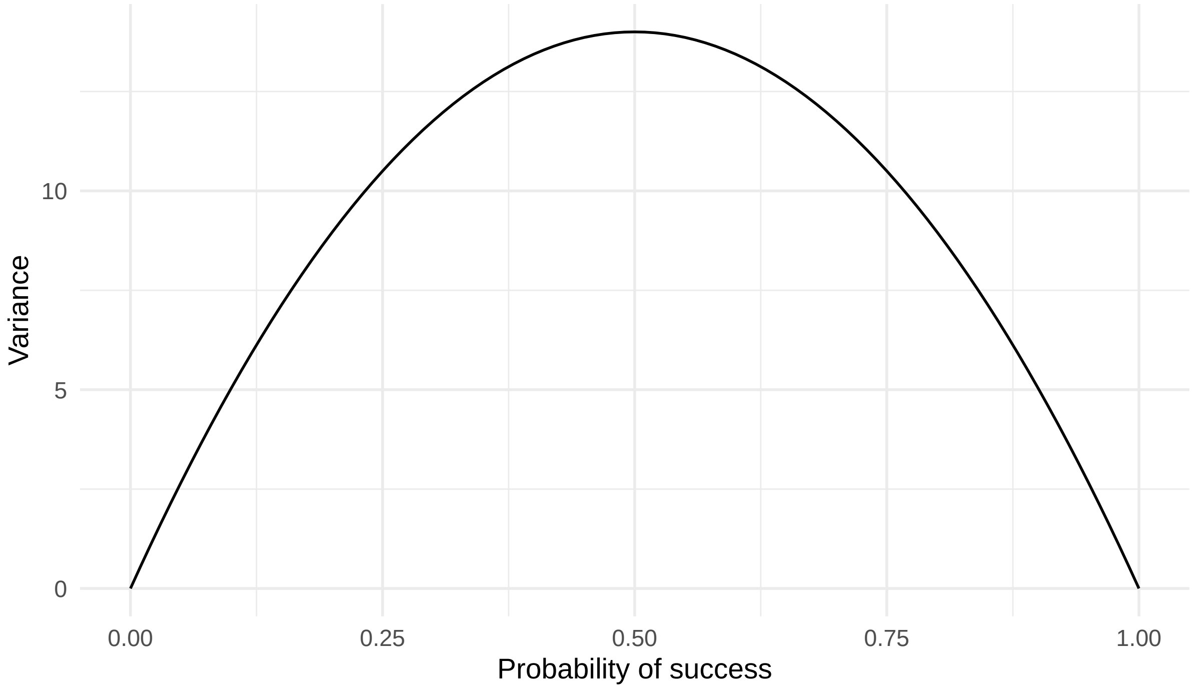


Figure 1: Variance of a Binomial random variable for varying proportions (sample size = 7)

```r
# Creating a variance plot for n2
for_plot %>%
  ggplot(aes(x=props, y=n2_var))+
  geom_line()+
  theme_minimal()+
  labs(x="Probability of success", y="Variance", caption = "Created by Smriti Kannan
↪   in STA303, Winter 2022")
```



Created by Smriti Kannan in STA303, Winter 2022

Figure 2: Variance of a Binomial random variable for varying proportions (sample size = 56)

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

```r
# Setting seed
set.seed(272)

# Setting up the necessary objects
sim_mean <- 10
sim_sd <- sqrt(2)
sample_size <- 30
number_of_samples <- 100

# Calculating the appropriate t-multiplier
tmult <- qt(0.975, 29)

# Simulating the population values
population <- rnorm(1000, mean = sim_mean, sd = sim_sd)

# Calculating the true population mean
pop_param <- mean(population)

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

# Creating a data frame for the sample data from the simulated population
group_id <- rep(c(1:100), each=30)
my_sim <- tibble(group_id,sample_set)

# Creating another data frame with confidence interval values for the samples
ci_vials <- my_sim %>%
  group_by(group_id) %>%
  summarise(
    mean = mean(sample_set),
    sd = sd(sample_set),
    lower = mean - tmult*(sd/sqrt(30)),
    upper = mean + tmult*(sd/sqrt(30)),
    capture = isTRUE(lower <= pop_param && pop_param <= upper)
  )

# Calculating the proportion of intervals that capture the true population parameter
```

```r
proportion_capture = sum(ci_vials$capture == TRUE)/number_of_samples

# Plotting the confidence intervals
ci_vials %>%
  ggplot(aes(x=group_id, y=mean, group=capture, color=capture)) +
  geom_point() +
  geom_errorbar(aes(ymin = lower, ymax = upper)) +
  coord_flip() +
  theme_minimal() +
  scale_color_manual(values = c("#B80000", "#122451"), name = "CI captures population
↪  parameter") +
  geom_hline(yintercept = pop_param, linetype="dotted") +
  labs(x=NULL, y=NULL, caption = "Created by Smriti Kannan in STA303, Winter 2022") +
  theme(legend.position = "bottom",
        axis.text.y  = element_blank())
```

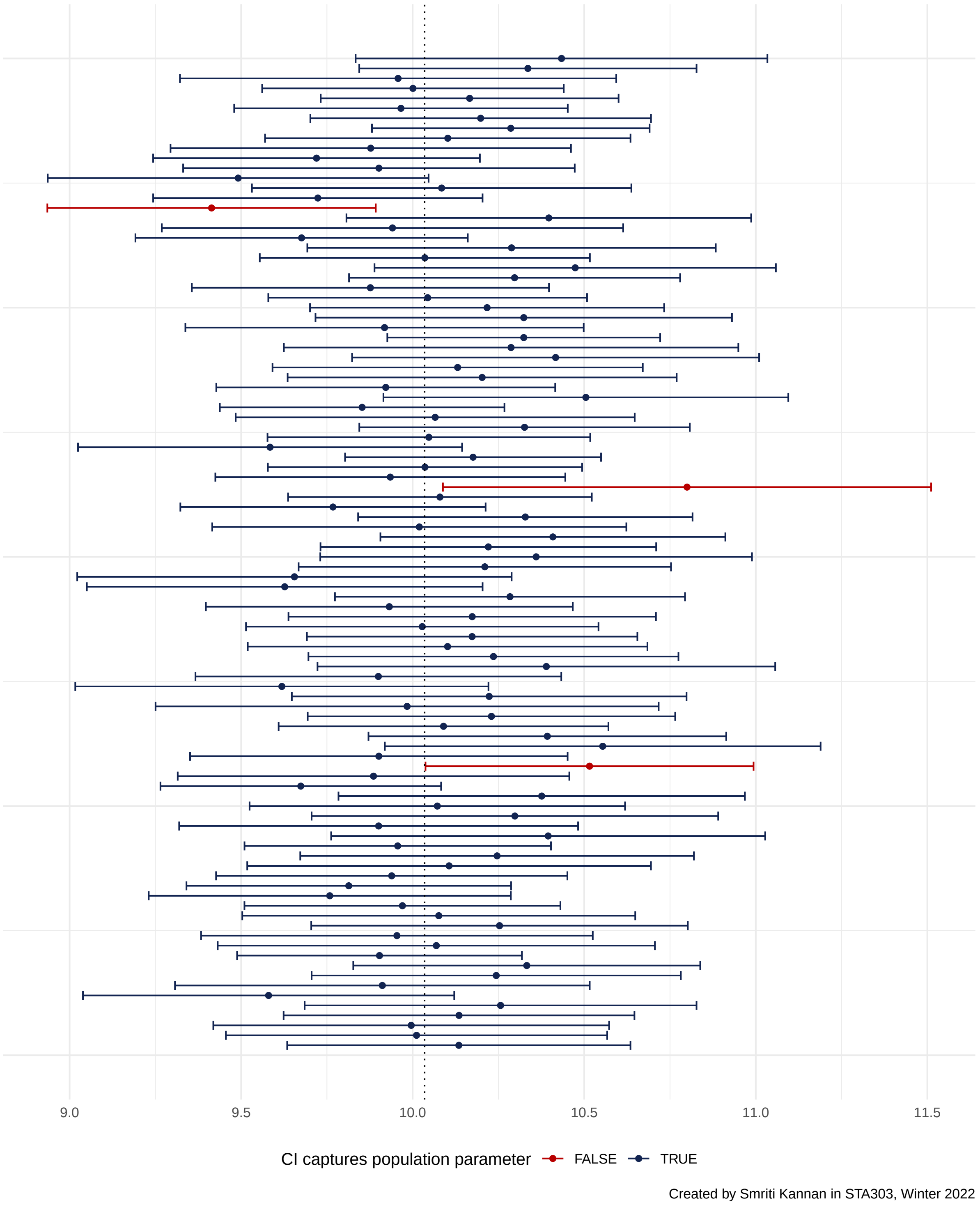


Figure 3: Exploring our long-run "confidence" in confidence intervals. This figure shows how often 95% confidence intervals from 100 simple random samples capture the population mean. The population was simulated from N(10, 2)

97 % of my intervals capture the population parameter.

In Figure 3, the graph has a dotted line which represents the true population parameter. We can do so because in this case, the population was simulated and therefore it was possible to calculate the true population parameter with ease. However, usually with real life data this is not possible since it is highly unlikely we would have the entire population data set. We would only be given a sample of the target population data set. Therefore, confidence intervals are highly useful as they can predict a range of values the true population parameter might take (with a certain degree of confidence).

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 analysis is to understand if there is a difference in the cGPA among students who answered a question on global poverty correctly and students who answered it incorrectly. The question was in the last 20 years did the proportion of people living below the poverty line decrease by half, double or stay the same. The data set used for this analysis has 200 observations that contain information on student's cGPA and their answer to the poverty question. In order to conduct this analysis, a new variable is added to the data set which shows if the student answered correctly or not and then the data is visualized and a hypothesis test is conducted.

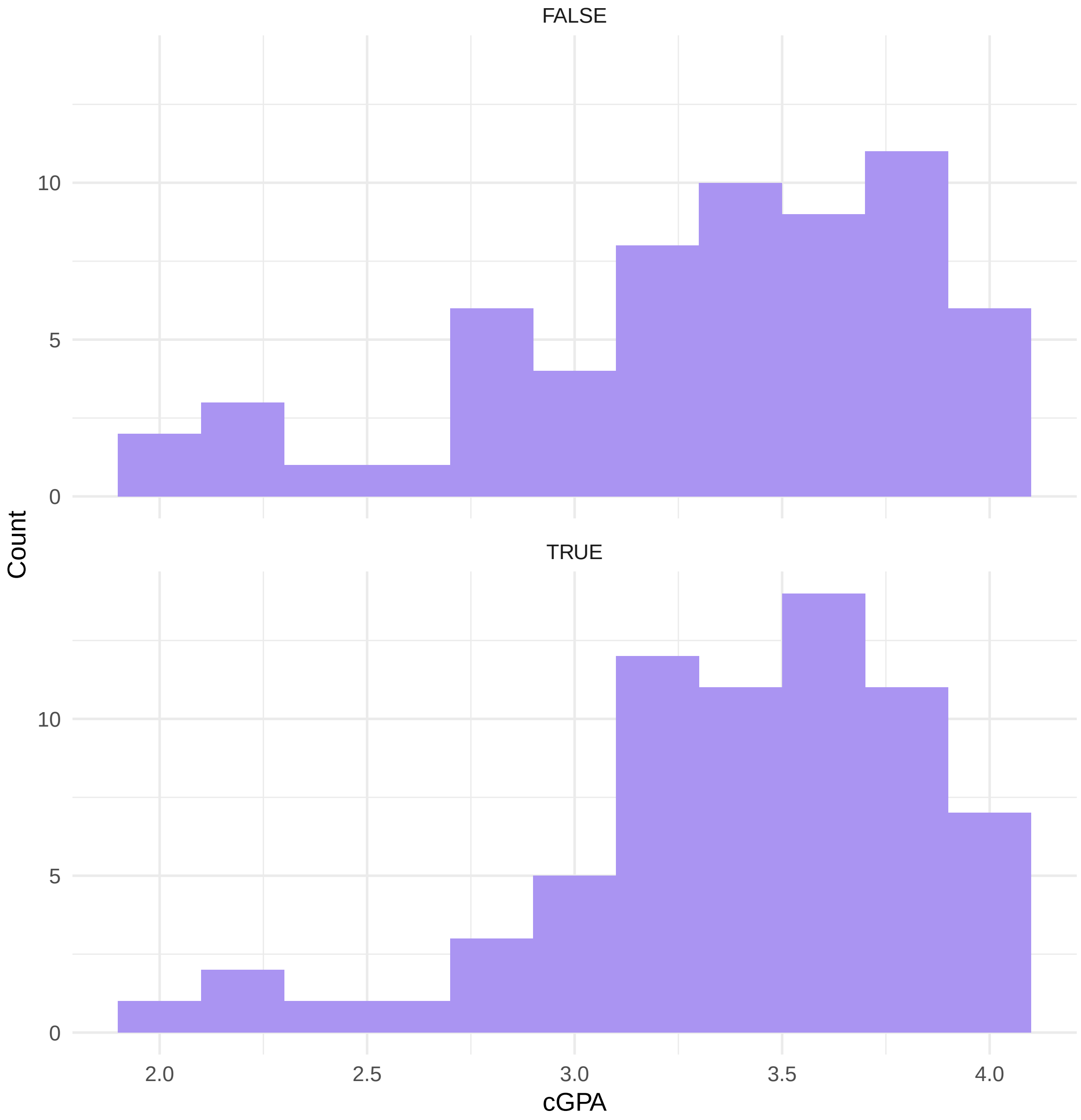
Wrangling the data

```
# Loading the data
cgpa_data <-
↪  read_excel("~/sta303-w22-mini-portfolio/data/sta303-mini-portfolio-poverty.xlsx")

# Cleaning the data and adding a variable
library(janitor)
cgpa_data <- clean_names(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_ha
         ↪  %>%
  filter(0 < cgpa) %>%
  filter(cgpa <= 4) %>%
  mutate(correct = case_when(global_poverty_ans == 'Halved' ~ TRUE, global_poverty_ans
↪   == 'Doubled' ~ FALSE,
                             global_poverty_ans == 'Stayed about the same' ~ FALSE))
```

Visualizing the data

```
# Creating two histograms which display the distribution of cGPAS
cgpa_data %>%
  ggplot() +
  geom_histogram(aes(x=cgpa), fill = "#AA94F2", binwidth = 0.2) +
  facet_wrap(~ correct, nrow = 2) +
  theme_minimal() +
  labs(x="cGPA", y="Count", caption = "Created by Smriti Kannan in STA303, Winter
↪   2022")
```



Created by Smriti Kannan in STA303, Winter 2022

Figure 4: Distribution of CGPAs for those who answered the global poverty question correctly and incorrectly

Testing

In order to investigate if there is an association between a student's cGPA and correctly answering a question on global poverty rates a Mann-Whitney U test will be used. This is a non-parametric test. A two-sample t-test is similar to the Mann-Whitney U but will not be used as it requires the assumption that the data be normally distributed in each group. By looking at the histograms above we can see clearly that the distribution for both groups is not normal but skewed. Therefore, a Mann-Whitney U test is more appropriate as we do not need to assume normality. However, this test does require the observations to be independent and random which are, for the most part, satisfied by the data set.

```
# Conducting the test to check if there is an association between cGPA and if they
↪  answered the question correctly
# A non-parametric test is used: 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
```

```
# Conducting the same test by constructing an equivalent linear regression 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
```

The p-value from the test and the equivalent linear model are 0.35 and 0.351 respectively. Therefore, there is no statistically significant evidence that there is a difference in the mean cGPA of students who answered the global poverty question correctly and students who answered it incorrectly.

Writing sample

Introduction

Working as a data scientist involves various roles and requires a broad set of skills. In this piece, I will go over some of the important analytical and soft skills that are needed to work as a data scientist. I will also go over some of my own qualifications and how my current studies are linked to this field of work.

Soft skills

Some of the most important soft skills to perform this job well include strong teamwork skills and good time management. Good teamwork skills are highly important as working as a Data Scientist would involve collaboration with peers. Working towards deadlines is also important as one might be working on multiple projects for numerous clients. Taking part in the Young Enterprise Program in Scotland helped me develop these skills. The 1-year program involved working with a team to create, develop and sell a product.

Analytic skills

Being a data scientist also requires good analytical skills. Important analytical skills include research skills and good knowledge of software (for example R) in order to carry out data analysis. Completing STA304: Surveys, Sampling and Observational Data in Fall 2020 further allowed me to develop and strengthen these skills. The final project for this course involved an independent research and data analysis task. My research project looked into the different factors that can affect a person's mental health using 2017 GSS data. My work on this can be found on my GitHub: https://github.com/smritikannan/A-study-on-mental-health.git

Connection to studies

Studying to be a statistics major has helped me gain both analytical and programming knowledge. Some additional useful skills I hope to gain after completing STA303 include a deeper understanding of linear models, understanding ethical considerations in data analysis and being able to communicate statistical results accurately to a range of audiences. I am also currently studying multivariate analysis in another statistical course, which will increase my theoretical and programming knowledge for multivariable analysis.

Conclusion

Therefore, in order to be a good data scientist one needs both analytical and soft skills. Taking part in the Young Enterprise Program helped me develop my soft skills such as teamwork and working towards a deadline. Pursuing a statistics major has greatly increased my theoretical and programming knowledge in data analysis. Conducting my own research and analysis for some of my courses has greatly improved my analytical and programming skills.

Word count: 396 words

Reflection

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

I am quite proud of the data visualization in this mini portfolio. Specifically, Figure 3 which displays the confidence interval values and which intervals capture the true population mean. I believe it is a very useful and informative plot which can be used to explain the analysis done in that section for both audiences with a statistical background and those without. It makes understanding what was done easier and the use of colour clearly highlights which confidence intervals capture the true population parameter and which do not. Graphs and figures are highly useful in data analysis as it is an easy way to visualize the data and is also useful when presenting analysis to other people.

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

This mini-portfolio has allowed me to display and improve some of my data analysis and presentation skills. I've learned how to produce good and informative data figures and also how to conduct hypothesis tests in different ways. The first method would be to use the correct test function and conduct an appropriate hypothesis test. The other method would be to construct a linear regression model that is equivalent to conducting the test. These skills are useful as they can be used in future data analysis work. I can produce better data visualization figures and can conduct hypothesis tests in an alternative way, which can prove to be useful for future research tasks and data analysis.

What is something I'd do differently next time?

Something I would do differently next time is try a different type of test for the final statistical skill sample. The test conducted was a non-parametric test to test for a difference in the mean cGPAs among students who answered a question on global poverty correctly and those who answered incorrectly. Non-parametric tests can be useful when the data is not normally distributed; however, the Mann-Whitney U test is just rank, not signed rank. Parametric tests are more powerful and in this case the data could be reanalysed and a two sample t-test could be used provided that the Central Limit Theorem can be applied and all other assumptions are satisfied.