STA303/1002 Mini-portfolio

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

Aliza Aziz Lakho

2022-02-03

Statistical skills sample

Setting up libraries

Visualizing the variance of a Binomial random variable for varying proportions

```r
#Setting the seed to ensure that we get the same results when we run this chunk
set.seed(448)

p <- 0.5
#Setting the two values for n to demonstrate that p = 0.5 will result in the largest
↪  variance for a Binomial random variable.
n1 <- 50
n2 <- 100

#Creating a vector of proportions
props <- seq(0, 1, by = 0.01)

#Creating a tiblle for plot
for_plot <- tibble(props, n1_var = n1*props*(1-props), n2_var = n2*props*(1-props))

#Creating a plot for varaince when n=50
g1 <- ggplot(data= for_plot, aes(x = props, y = n1_var)) +
  geom_line() +
  labs(caption = "Created by Aliza Aziz Lakho in STA303/1002, Winter 2022.", title =
↪  "Varaince of a Binomial RV when n=50")+
  xlab("Props")+
  ylab("Variance")+
  theme_minimal()

g1
```

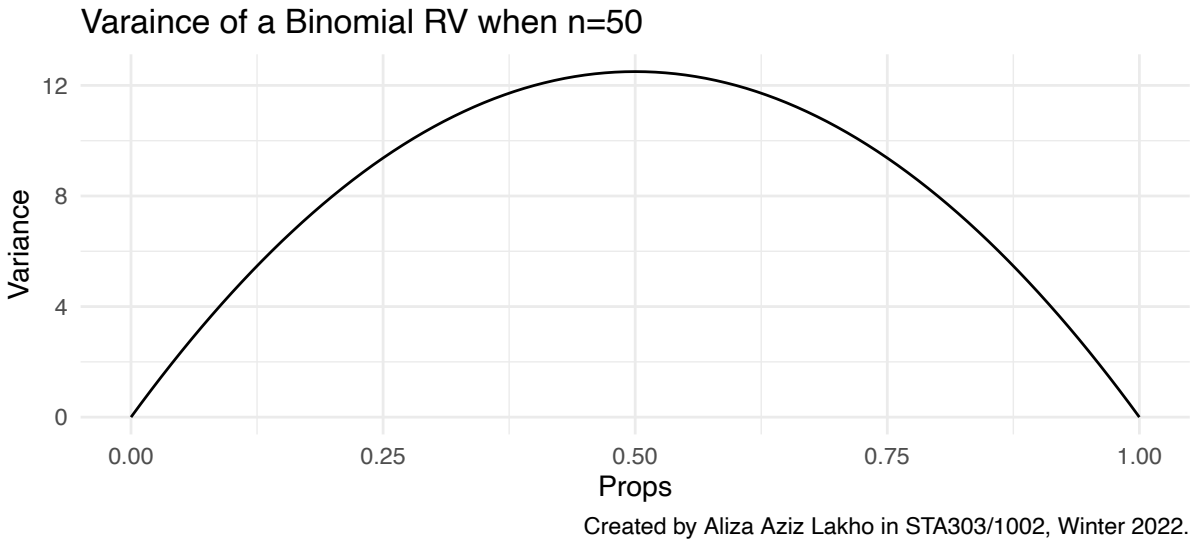


Figure 1: Bionomial RV

```
#Creating a plot for varaince when n=100
g2 <- ggplot(data= for_plot, aes(x = props, y = n2_var)) +
  geom_line() +
  labs(caption = "Created by Aliza Aziz Lakho in STA303/1002, Winter 2022.", title =
↪   "Varaince of a Binomial RV when n=100")+
  xlab("Props")+
  ylab("Variance")+
  theme_minimal()

g2
```

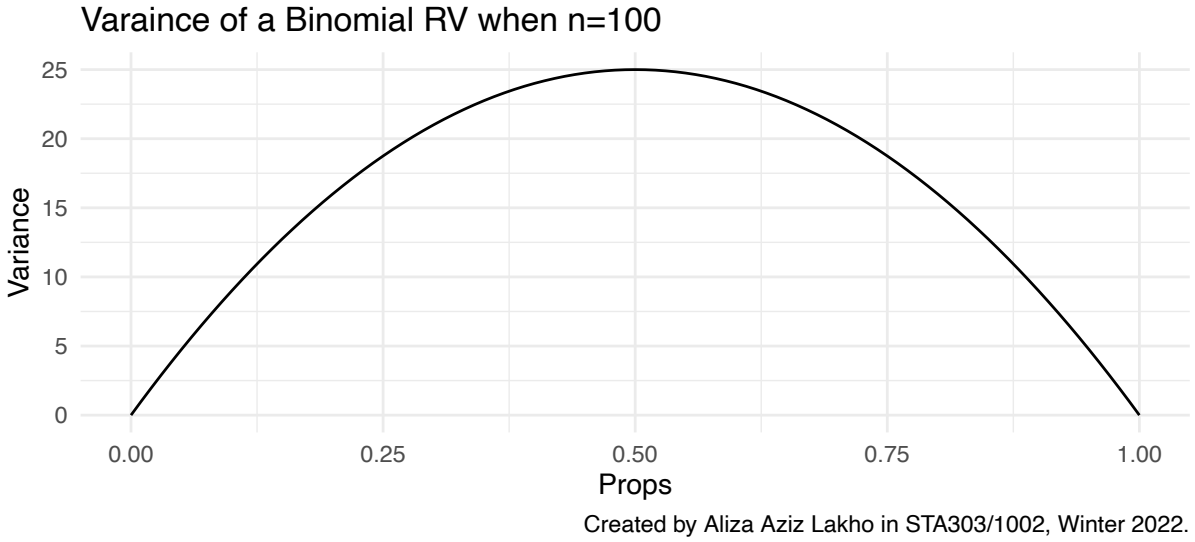
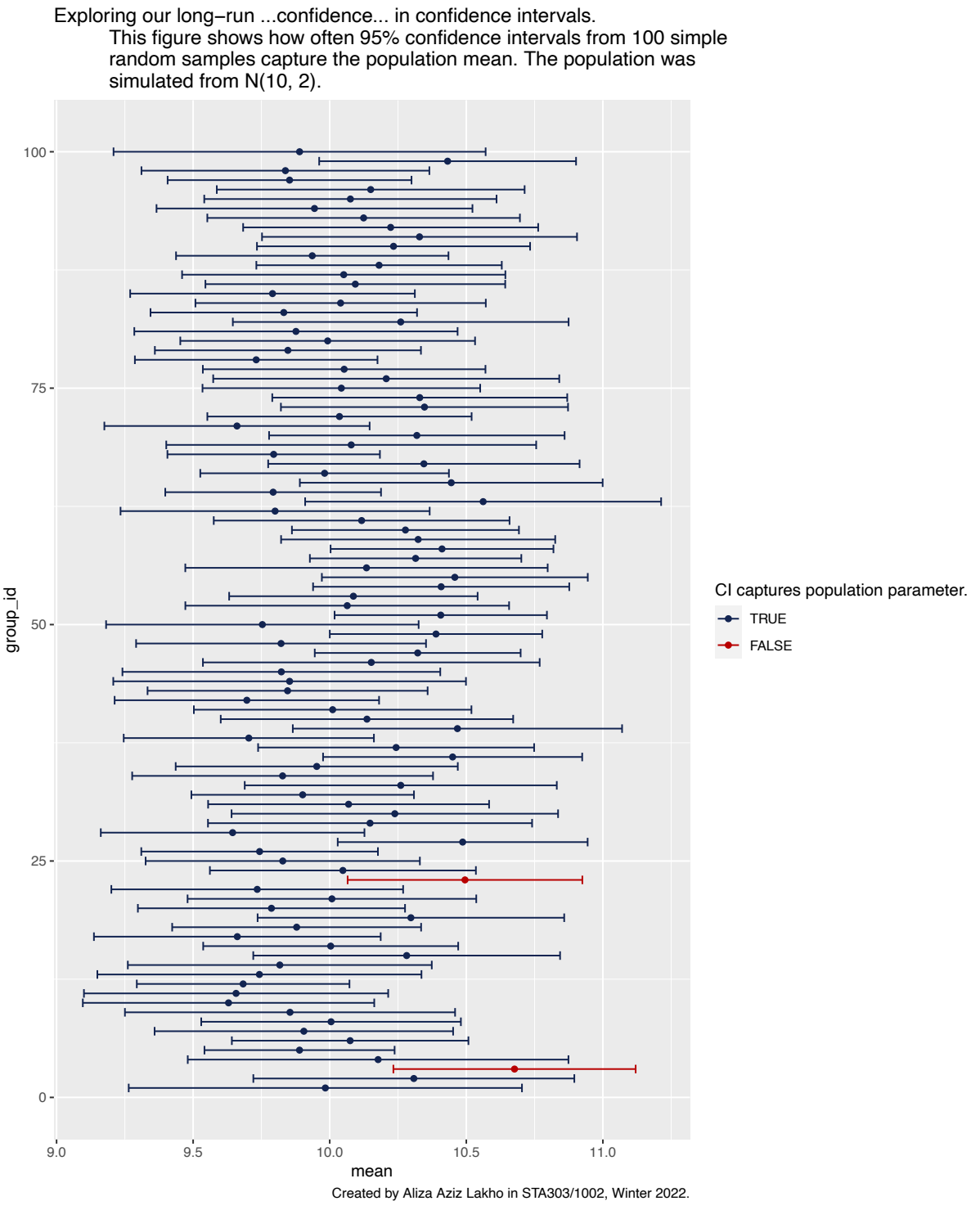


Figure 2: Bionomial RV

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

```
#Setting the seed to ensure that we get the same results when we run this chunk
set.seed(448)

#Using  N(10,2), and taking 100 independent, random samples of size 30 observations
↪  each from it.
sim_mean <- 10
sim_sd <- sqrt(2)
sample_size <- 30
number_of_samples <- 100

#Calculating the appropriate t-multiplier
tmult <- qt(0.975, sample_size - 1)

#Creating a simulated population using sim_mean and sim_sd with 1000 values
population <- rnorm(1000, mean = sim_mean, sd = sim_sd)

#Actual true mean
pop_param <- mean(population)

#Getting 100 samples of size 30 from 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
group_id <- rep(1:number_of_samples, each = sample_size)

#Contains our simulation
my_sim <- tibble(group_id, sample_set)

#Finding confidencce interval values
ci_vals <- my_sim %>%
  group_by(group_id) %>%
  summarise(mean= mean(sample_set), sd= sd(sample_set))%>%
  mutate(lower= (mean - tmult*sd/sqrt(sample_size)),
         upper= (mean + tmult*sd/sqrt(sample_size)))%>%
  mutate(capture = ifelse(pop_param >= lower & pop_param <= upper, TRUE, FALSE))

#Proportion of intervals created that capture the population parameter
```

```
proportion_capture <- sum(ci_vals$capture == TRUE)/number_of_samples

g3 <- ggplot(ci_vals, aes(x= group_id, y= mean, color= capture))+
  geom_point()+
  geom_errorbar(aes(ymin= lower, ymax=upper))+
  coord_flip()+
  scale_color_manual(values = c("TRUE" = "#122451", "FALSE" = "#B80000"))+
  ggtitle("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).")+
  labs(caption = "Created by Aliza Aziz Lakho in STA303/1002, Winter 2022.",
       colour = "CI captures population parameter.")
g3
```



98

The reason that we can include the population parameter in this plot is that because our population capture is almost 95% (it is 94%) and so it will contain the population parameter.

The reason why we cannot usually compare the population parameter to our confidence interval in practice is because we are only using one random sample whose confidence interval can or cannot perhaps have 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

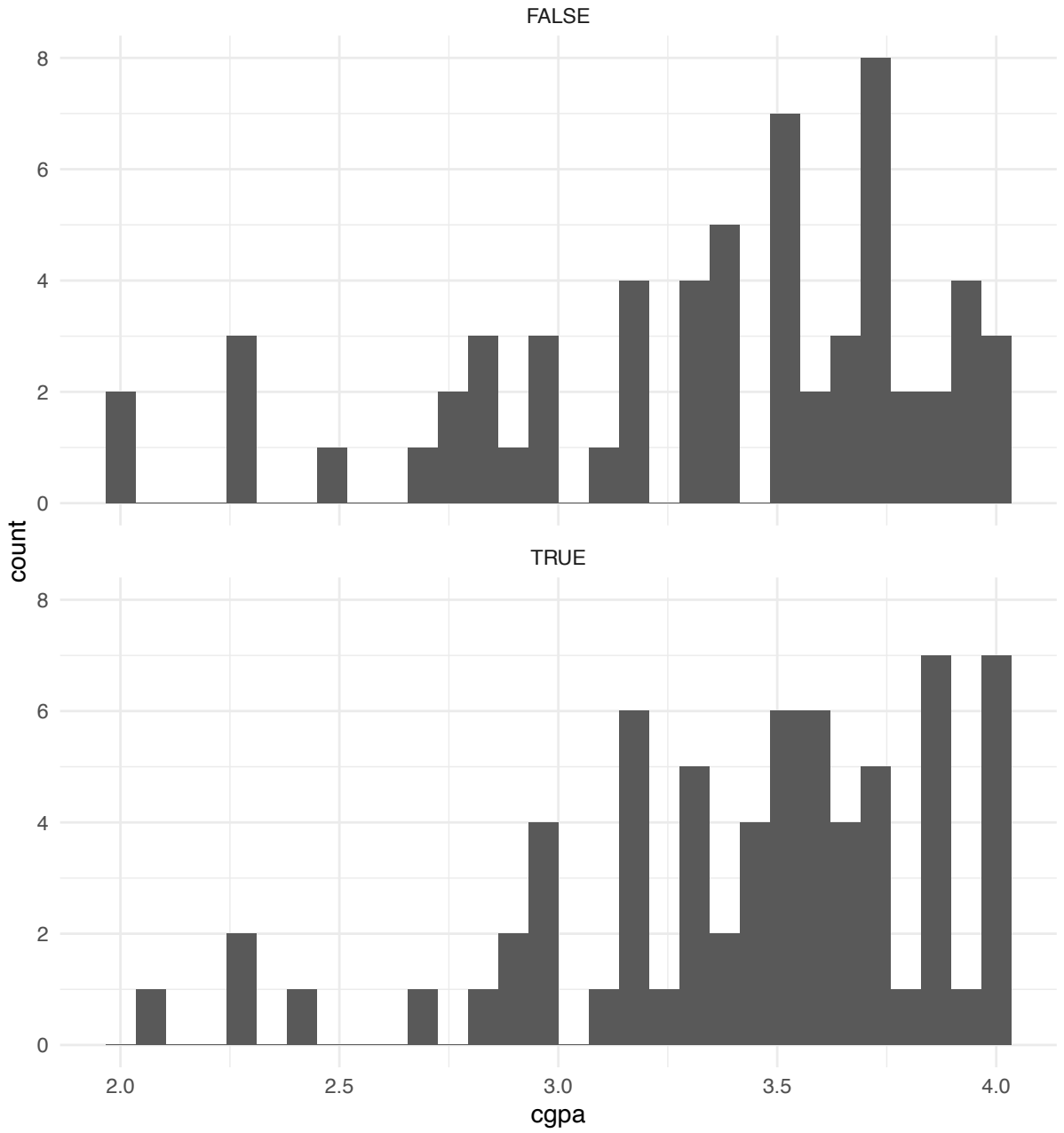
The goal is the study is to find whether if there is some sort of relationship between cGPA and STA303/1002 students correctly answering a question on global poverty rates. The survey asked the students two questions: Q1.What their current cumulative grade point average (CGPA) was at U of T. Q2)Whether the proportion of people living below the global poverty line had halved, doubled or stayed about the same in the last 20 years. To answer this question, we will be using hypothesis testing on a sample of 200 responses to this survey.

Wrangling the data

```
#importing the dataset
cgpa_data <- read_xlsx("data/sta303-mini-portfolio-poverty.xlsx")

cgpa_data <- cgpa_data%>%
  clean_names()%>%
  rename(cgpa=
↪  what_is_your_c_gpa_at_u_of_t_if_you_dont_want_to_answer_you_can_put_a_0)%>%
  rename(global_poverty_ans=
↪  in_the_last_20_years_the_proportion_of_the_world_population_living_in_extreme_poverty_has)%>%
  na.omit(cgpa_data)%>%
  filter(cgpa > 0.0 & cgpa <= 4.0)%>%
  mutate(correct = ifelse(global_poverty_ans == "Halved", TRUE, FALSE))
```

Visualizing the data

```
ggplot(cgpa_data, aes(x= cgpa))+
  geom_histogram(bins=30)+
  theme_minimal()+
  facet_wrap(cgpa_data$correct, ncol=1)
```



Testing

For the purposes of testing, we will use t-test. As we can see in the above visualizations, since we have large datasets, we can say that the mean is normally distributed owing to the central limit theorem. For this case, we can assume equal variance since the ratio of cgpa variance is less than 3 (this is depicted below). We can also safely assume that the responses form a simple random

sample. Lastly, we have a large enough sample size and our data, cgpa, tends to be continous which allows us to use t-testing.

```
#Shwoing that the ratio is less than 3:
True_Data = filter(cgpa_data, correct == TRUE)
var_true = var(True_Data$cgpa)

False_Data = filter(cgpa_data, correct == FALSE)
var_false = var(False_Data$cgpa)

var_false/var_true
```

```
## [1] 1.358846
```

```
#performing t-test
t.test(cgpa ~ cgpa_data$correct, data = cgpa_data)
```

```
##
##  Welch Two Sample t-test
##
## data:  cgpa by cgpa_data$correct
## t = -1.205, df = 118.95, p-value = 0.2306
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.26951549  0.06558743
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            3.329507            3.431471
```

```
#performing linear regression between cGPA and the correctness of the survey question
lr <- lm(cgpa~correct, data = cgpa_data)
summary(lr)
```

```
##
## Call:
## lm(formula = cgpa ~ correct, data = cgpa_data)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.39147 -0.23147  0.07049  0.38049  0.67049
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.32951    0.06093  54.648   <2e-16 ***
## correctTRUE  0.10196    0.08392   1.215    0.227
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4758 on 127 degrees of freedom
## Multiple R-squared:  0.01149,    Adjusted R-squared:  0.003708
## F-statistic: 1.476 on 1 and 127 DF,  p-value: 0.2266
```

By looking at the results of the linear regression model, we can conclude that that since the estimare for correctTRUE is 0.1, students who have correctly answered on the survey would typically have a higher cGPA (by about 0.1) than those who did not. This model also shows that since the p-value is not significant of correctTRUE we can conclude that there is no evidence to suggest a difference in student's cGPA's.

Conclusion

To summarize, I believe that I have necessary skills that complement this job posting of Data Scientist at Yelp. I also believe in further improving one self, for which I will take up on tasks such as volunteering that make me a well rounded candidate for this position.

Word count: 467 words