PHStatsII-HW3b

Marty Ross

2022-10-02

knitr::opts\_chunk$set(echo = TRUE)  
library(tidyverse)

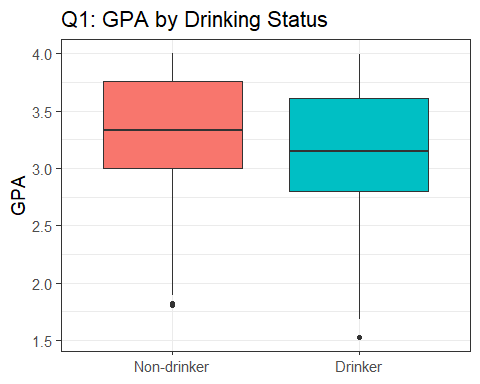
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --  
## v ggplot2 3.3.6 v purrr 0.3.4  
## v tibble 3.1.8 v dplyr 1.0.9  
## v tidyr 1.2.0 v stringr 1.4.1  
## v readr 2.1.2 v forcats 0.5.2  
## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

setwd("~/SpatialAnalysis/01-2022\_PHStatsII/Assignments/A3/A3ptB")  
casData <- read.csv("casData.csv")  
# head(casData)  
# Part A: GPA by drinking  
# class(casData$DRINK)  
# table(casData$DRINK)  
# make DRINK a factor with labels (helpful for visual displays, cuts down on code   
# needed to label groups)  
casData <- casData %>%  
 mutate(DRINK = factor(DRINK, levels = c(0, 1), labels = c("Non-drinker",  
 "Drinker")))

# Part A: What is the unadjusted relationship between GPA and consumption of alcohol?

Q1: Simple Boxplot of GPA by drinking status

ggplot(casData, aes(x = DRINK, y = GPA, fill = DRINK)) +  
 geom\_boxplot() +   
 theme\_bw(base\_size = 14) +  
 labs(title = 'Q1: GPA by Drinking Status') +  
 theme(axis.title.x = element\_blank(),  
 legend.position = 'none')



Q2-3. Regress by Drinking Status

model1 <- lm(GPA ~ DRINK, data = casData)  
# summary(model1)  
co1 <- model1$coefficients  
cat('Q2. Estimated mean difference in GPA for students who drink versus those',  
 'who do not: \n', round(co1[2], 2))

## Q2. Estimated mean difference in GPA for students who drink versus those who do not:   
## -0.18

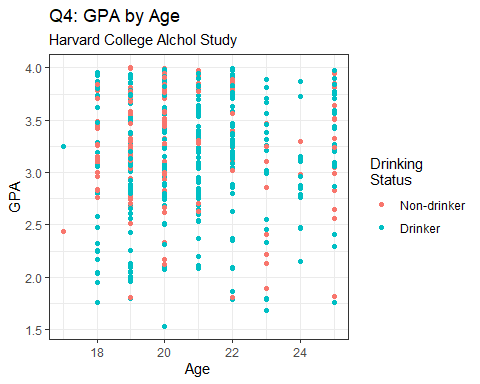
conf1 <- confint(model1)  
cat('Q3. 95% Confidence interval difference in GPA for students who drink versus those',  
 'who do not: \n', round(conf1[2,], 2))

## Q3. 95% Confidence interval difference in GPA for students who drink versus those who do not:   
## -0.28 -0.08

# Part B: What is the relationship between GPA and consumption of alcohol, when adjusted for a student’s age?

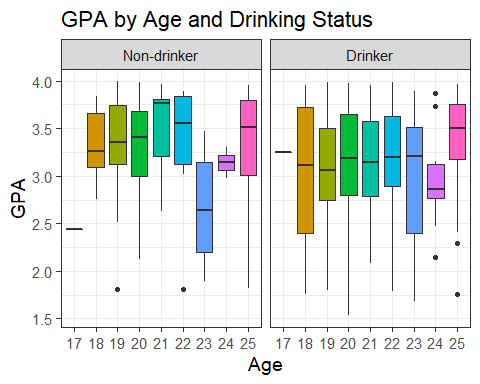
Q4: A scatterplot showing the relationship between GPA and a student’s age

ggplot(casData, aes(x = age, y = GPA, color = DRINK)) +  
 geom\_point() +  
 theme\_bw() +  
 labs(x = 'Age', color = 'Drinking\nStatus',  
 title = 'Q4: GPA by Age',  
 subtitle = 'Harvard College Alchol Study')

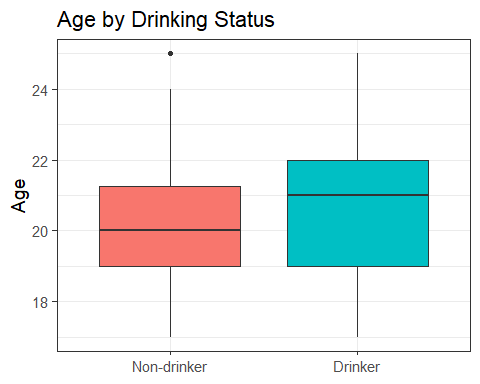


Q5. Side-by-side boxplots showing the difference in age between students who drink and students who do not drink. Note: I wasn’t certain I was fulfilling this correctly based on the starter code and how the question was phrased, so I’m giving a couple options.

casData <- casData %>% mutate(agef = factor(age))  
  
ggplot(casData, aes(x = agef, y = GPA, fill = agef)) +  
 geom\_boxplot() +  
 facet\_wrap(facets = casData$DRINK) +  
 theme\_bw(base\_size = 14) +  
 theme(legend.position = 'none') +  
 labs(x = 'Age', title = 'GPA by Age and Drinking Status')



ggplot(casData, aes(x = DRINK, y = age, fill = DRINK)) +  
 geom\_boxplot() +  
 theme\_bw(base\_size = 14) +  
 theme(legend.position = 'none',  
 axis.title.x = element\_blank()) +  
 labs(y = 'Age',  
 title = 'Age by Drinking Status')



Q6-9 Multiple Regression on GPA, Drinking + age

model2 <- lm(GPA ~ DRINK + age, data = casData)  
# summary(model2)  
co2 <- coef(model2)  
# The estimated mean GPA for 20-year-old students who drink  
cat('Q6. The estimated mean GPA for 20-year-old students who drink:\n ',   
 round(co2[1] + co2[2]\*1 + co2[3]\*20,2))

## Q6. The estimated mean GPA for 20-year-old students who drink:  
## 3.11

cat('Q7. The estimated mean GPA for 20-year-old students who do not drink:\n ',   
 round(co2[1] + co2[2]\*0 + co2[3]\*20, 2))

## Q7. The estimated mean GPA for 20-year-old students who do not drink:  
## 3.3

cat('Q8. The estimated mean difference in GPA for students who drink, compared to students of the same age who do not:\n ', round(co2[2], 2))

## Q8. The estimated mean difference in GPA for students who drink, compared to students of the same age who do not:  
## -0.19

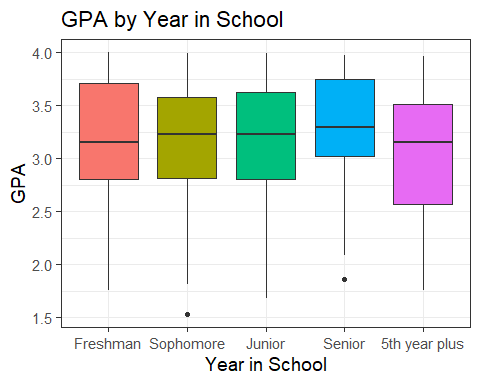
conf2 <- confint(model2)  
cat(paste0('Q9. A 95% confidence interval for this estimated difference:\n (',  
 round(conf2[2,1],2), ',', round(conf2[2,2],2),')'))

## Q9. A 95% confidence interval for this estimated difference:  
## (-0.29,-0.08)

## Part C: What is the relationship between GPA and consumption of alcohol, when adjusted for a student’s year in school?

Q10. Side-by-side boxplots showing the relationship between GPA and a student’s year in school. [Hint: To include school year as a categorical variable, you must use as.factor(schoolyr) instead of just schoolyr in your R code.]

# created factor version of year in school  
casData <- casData %>%  
 mutate(schoolyrf = factor(schoolyr,   
 labels = c("Freshman", "Sophomore", "Junior", "Senior",   
 "5th year plus")))  
ggplot(casData, aes(x = schoolyrf, y = GPA, fill = schoolyrf)) +  
 geom\_boxplot() +  
 theme\_bw(base\_size = 14) +  
 theme(legend.position = 'none') +  
 labs(x = 'Year in School',  
 title = 'GPA by Year in School')



Q11. A table showing the number and proportions of students in each school year who drink/don’t drink

yrdrnk\_df <- casData %>%  
 group\_by(schoolyrf, DRINK) %>%  
 summarise(n = n(), .groups = 'drop\_last') %>%  
 mutate(prop = paste0(round(n/sum(n)\*100,1),"%")) %>%  
 select(-n) %>%  
 spread(DRINK, prop)  
colnames(yrdrnk\_df)[1] <- 'School Year'  
knitr::kable(yrdrnk\_df, 'html', align = 'c', caption = 'Q11. Proportions of Drinkers by School Year', table.attr = "style='width:40%;'")

Q11. Proportions of Drinkers by School Year

School Year

Non-drinker

Drinker

Freshman

40%

60%

Sophomore

36.8%

63.2%

Junior

27%

73%

Senior

29.1%

70.9%

5th year plus

26.5%

73.5%

Q12-15. Model GPA by drinking status and school year

# multiple regression of GPA on drinking and school year  
model3 <- lm(GPA ~ DRINK + schoolyr, data = casData)  
# summary(model3)  
co3 <- coef(model3)  
conf3 <- confint(model3)

cat('Q12. The estimated mean GPA for freshman students who drink:\n ',   
 round(co3[1] + co3[2]\*1, 2), '\n')

## Q12. The estimated mean GPA for freshman students who drink:  
## 3.06

cat('Q13. The estimated mean GPA for senior students who drink:\n ',   
 round(co3[1] + co3[2]\*1 + co3[5]\*1, 2), '\n')

## Q13. The estimated mean GPA for senior students who drink:  
## NA

cat('Q14. The estimated mean difference in GPA for students who drink, compared to students of the same school year who do not:\n ', round(co3[2],2), '\n')

## Q14. The estimated mean difference in GPA for students who drink, compared to students of the same school year who do not:  
## -0.19

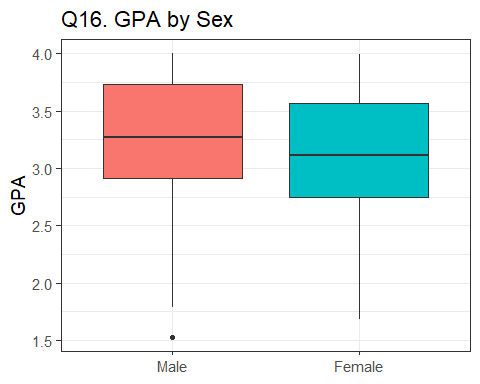
cat('Q15. A 95% confidence interval for this estimated difference:\n ',   
 paste0('(', round(conf3[2,1], 2), ', ', round(conf3[2,2], 2), ')'))

## Q15. A 95% confidence interval for this estimated difference:  
## (-0.29, -0.08)

## Part D: What is the relationship between GPA and consumption of alcohol, when adjusted for a student’s sex?

Q16. Side-by-side boxplots showing the difference in GPA between male and female students

casData <- casData %>%  
 mutate(Sex = factor(male, labels = c('Male','Female')))  
ggplot(casData, aes(x = Sex, y = GPA, fill = Sex)) +  
 geom\_boxplot() +  
 theme\_bw(base\_size = 14) +  
 theme(legend.position = 'none',  
 axis.title.x = element\_blank()) +  
 labs(title = 'Q16. GPA by Sex')



Q17. A table showing the number and proportions of students of each sex who drink/don’t drink

sxdrnk\_df <- casData %>%  
 group\_by(Sex, DRINK) %>%  
 summarise(n = n(), .groups = 'drop\_last') %>%  
 mutate(prop = paste0(round(n/sum(n)\*100,1),"%")) %>%  
 select(-n) %>%  
 spread(DRINK, prop)  
colnames(sxdrnk\_df)[1] <- 'Sex'  
knitr::kable(sxdrnk\_df, 'html', align = 'c', caption = 'Q17. Proportions of Drinkers by Sex', table.attr = "style='width:30%;'")

Q17. Proportions of Drinkers by Sex

Sex

Non-drinker

Drinker

Male

34.1%

65.9%

Female

30.6%

69.4%

model4 <- lm(GPA ~ DRINK + male, data=casData)  
#summary(model4)  
co4 <- coef(model4)  
conf4 <- confint(model4)

cat('Q18. The estimated mean difference in GPA for students who drink, compared to students of the same sex who do not:\n ', round(co4[2], 2), '\n')

## Q18. The estimated mean difference in GPA for students who drink, compared to students of the same sex who do not:  
## -0.18

cat('Q19. A 95% confidence interval for this estimated difference:\n ',   
 paste0('(', round(conf4[2,1], 2), ', ', round(conf4[2,2], 2), ')\n'))

## Q19. A 95% confidence interval for this estimated difference:  
## (-0.28, -0.07)

cat('Q20. The estimated mean difference in GPA for male students, compared to female students \_of the same drinking status\_:\n ', round(co4[3],2), '\n')

## Q20. The estimated mean difference in GPA for male students, compared to female students \_of the same drinking status\_:  
## -0.16

cat('Q15. A 95% confidence interval for this estimated difference:\n ',   
 paste0('(', round(conf4[3,1], 2), ', ', round(conf4[3,2], 2), ')'))

## Q15. A 95% confidence interval for this estimated difference:  
## (-0.26, -0.06)

Q22. According to the College Alcohol Study from the Harvard School of Public Health, it was found that students who drank, defined as having drank alcohol in the past 30 days, were observed as having a 0.19 lower mean GPA (95% CI: -0.29,-0.09) than students who did not drink alcohol when controlling for year in school, a significant finding at the 0.05 confidence level.

Q23. See above