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()
```

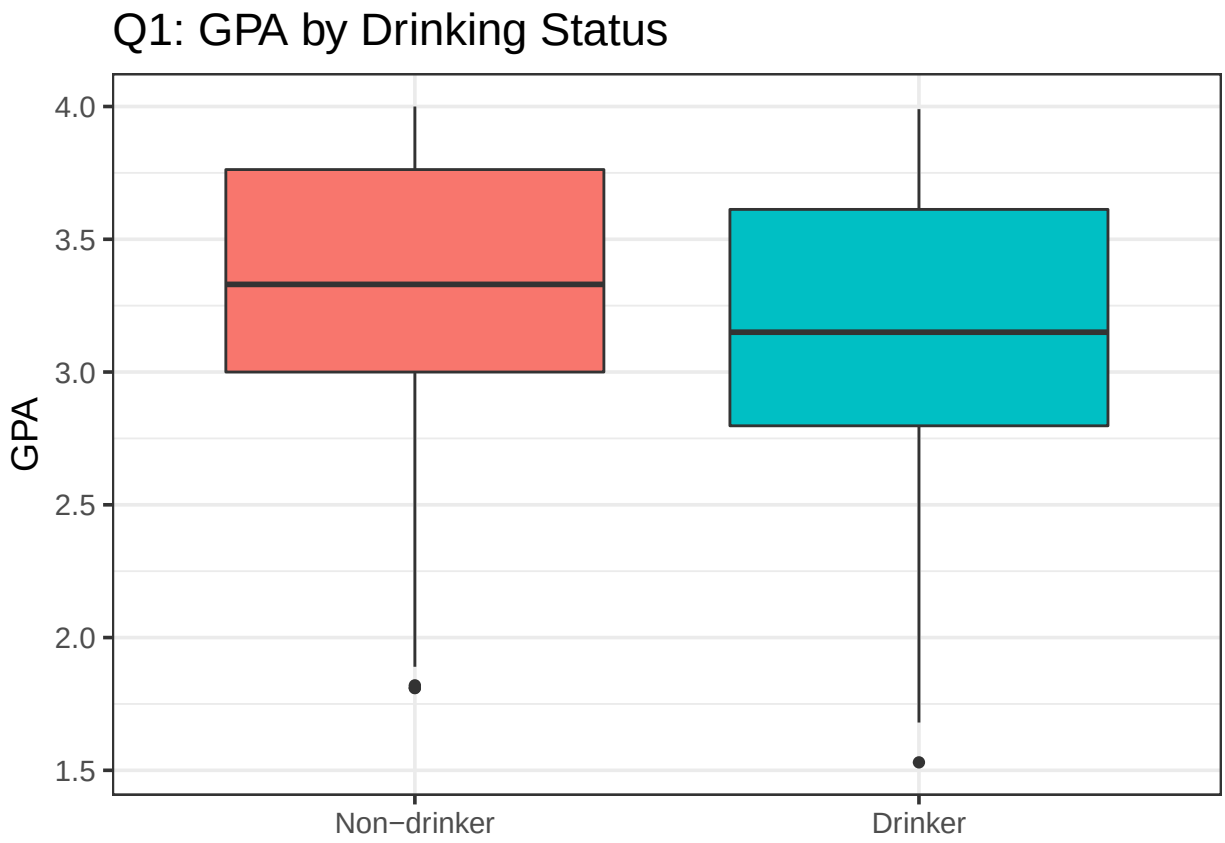
```
setwd("~/SpatialAnalysis/01-2022_PHStatsII/Assignments/A3/A3ptB")
casData <- read.csv("casData.csv")
# head(casData)
# Part A:  GPA by drinking
# class(casData$DRINK)
# table(casData$DRINK)
# make DRINK a factor with labels (helpful for visual displays, cuts down on code
# needed to label groups)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

Exercise 1:

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(paste0('Q2. Estimated mean difference in GPA for students who drink versus those\n',
          '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
```

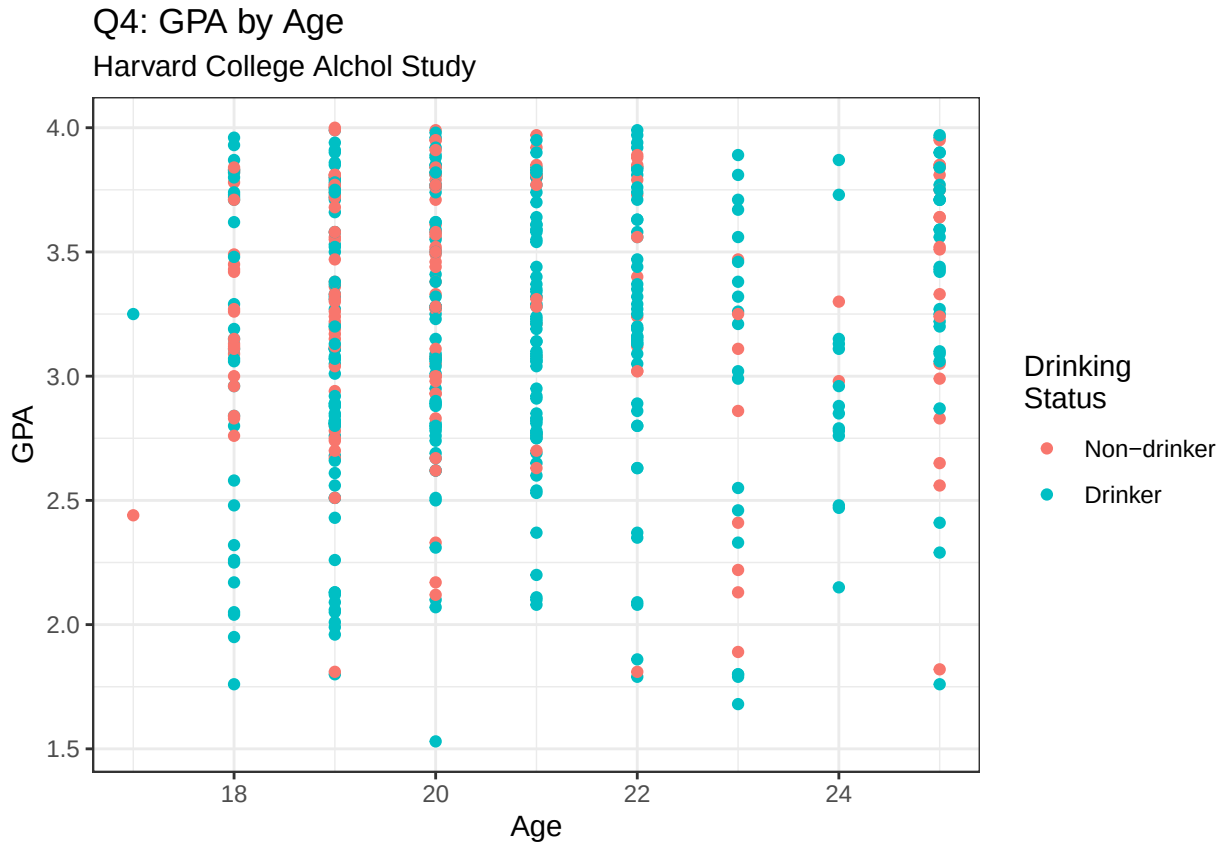
```
conf1 <- confint(model1)
cat(paste0('Q3. 95% Confidence interval difference in GPA for students who drink\n',
           'versus those who do not:\n  (', paste(round(conf1[2,], 2), collapse = ', '), ')'))
```

```
## 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

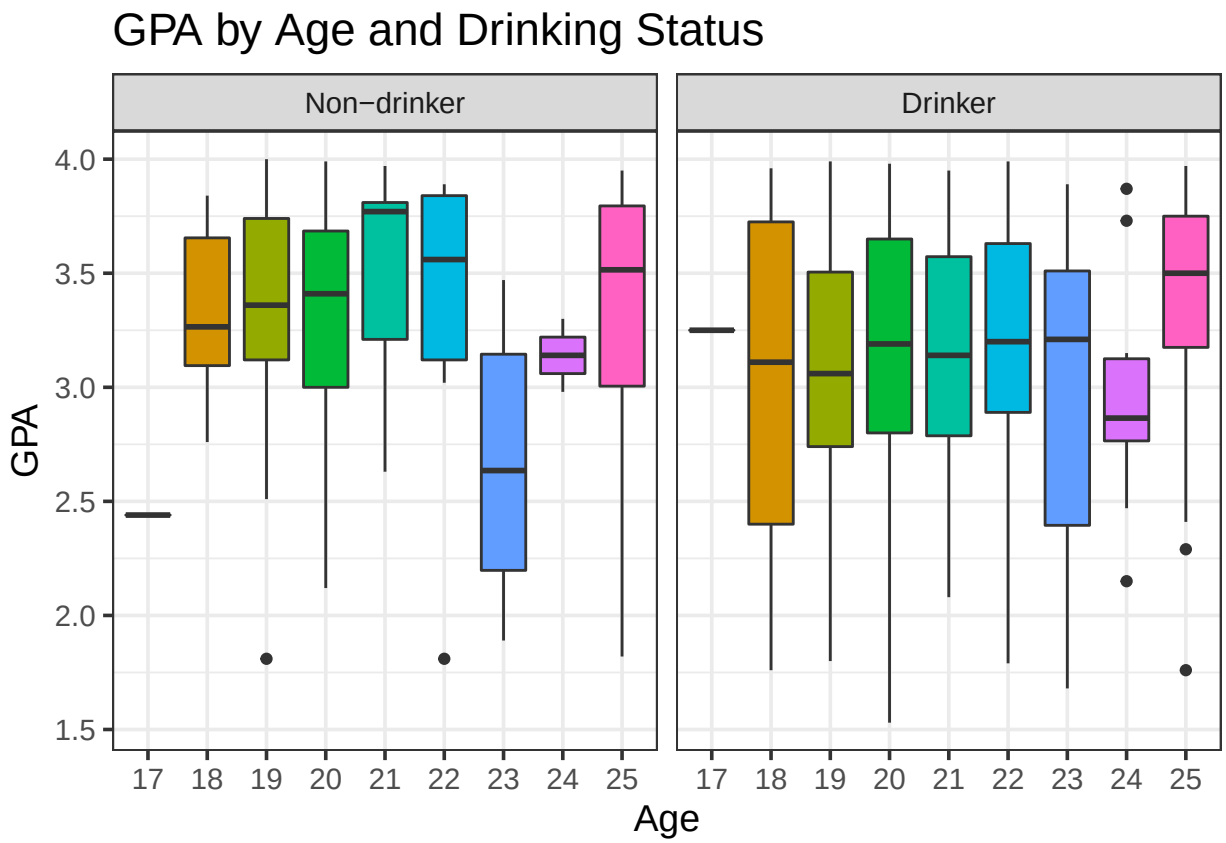
```
ggplot(casData, aes(x = age, y = GPA, color = DRINK)) +
  geom_point() +
  theme_bw() +
  labs(x = 'Age', color = 'Drinking\nStatus',
       title = 'Q4: GPA by Age',
       subtitle = 'Harvard College Alchol Study')
```



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)
conf2 <- confint(model2)
```

```
# The estimated mean GPA for 20-year-old students who drink
cat(paste0('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(paste0('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(paste0('Q8. The estimated mean difference in GPA for students who drink, compared\n',
    '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
```

```
cat(paste0('Q9. A 95% confidence interval for this estimated difference:\n  (',
    paste(round(conf2[2,],2), collapse = ', '), ')'))
```

```
## 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, levels = c(1,2,3,4,5),
                            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')
```

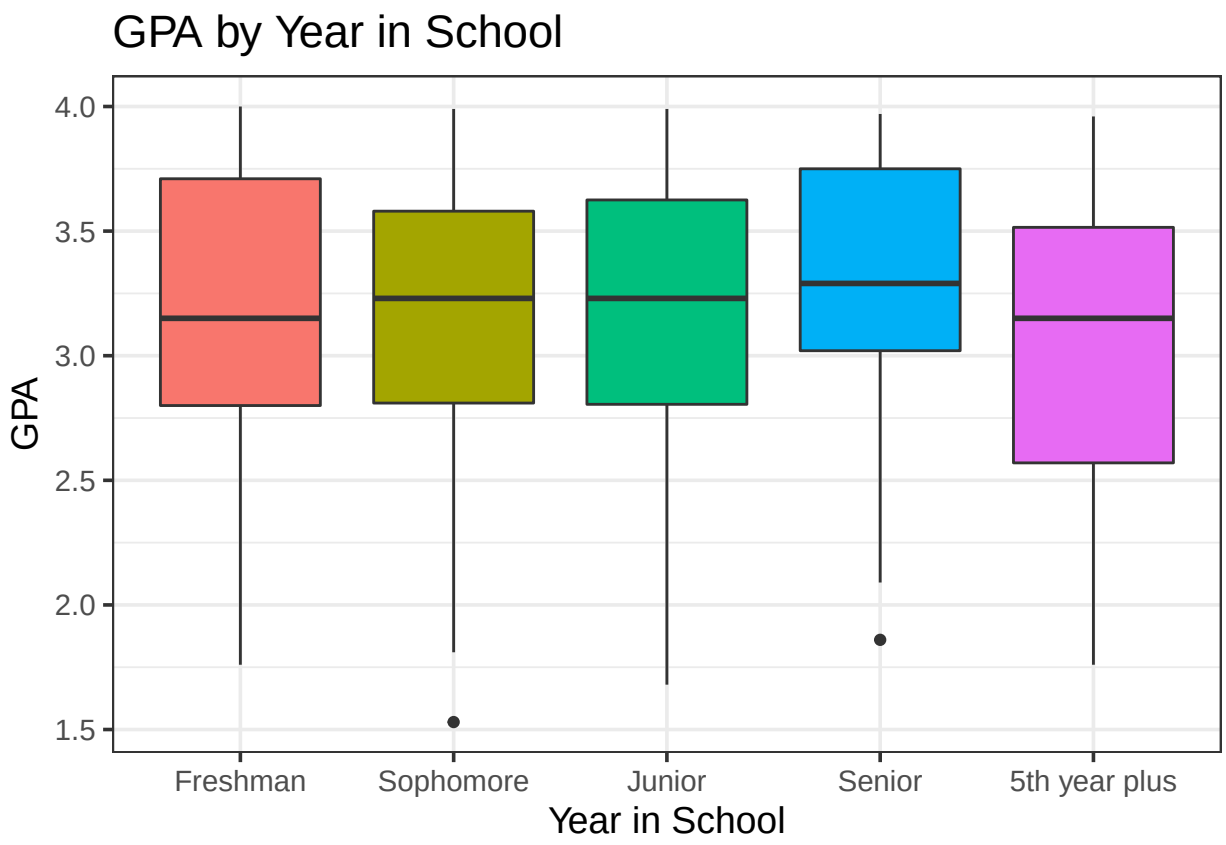


Table 1: Q11. Proportions of Drinkers by School Year

School Year	# Non-drinker	# Drinker	% Non-drinker	% Drinker
Freshman	48	72	40%	60%
Sophomore	42	72	36.8%	63.2%
Junior	31	84	27%	73%
Senior	34	83	29.1%	70.9%
5th year plus	9	25	26.5%	73.5%

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

```
yrdrnk_df <- casData %>%
  count(schoolyrf, DRINK) %>%
  group_by(schoolyrf) %>%
  spread(DRINK, n) %>%
  rename_with(~ paste0("# ", .x))
yrdrnk_df2 <- casData %>%
  count(schoolyrf, DRINK) %>%
  group_by(schoolyrf) %>%
  mutate(prop = paste0(round(n/sum(n)*100,1),"%")) %>%
  select(-n) %>%
  spread(DRINK, prop) %>%
  rename_with(~ paste0("% ", .x))
yrdrnk_df <- cbind(yrdrnk_df, yrdrnk_df2[2:3])
colnames(yrdrnk_df)[1] <- 'School Year'
knitr::kable(yrdrnk_df, format = 'latex', align = 'c',
             caption = 'Q11. Proportions of Drinkers by School Year')
```

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

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

```
cat(paste0('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(paste0('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:
##   3.23
```

```
cat(paste0('Q14. The estimated mean difference in GPA for students who drink, compared\n',
    '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
```

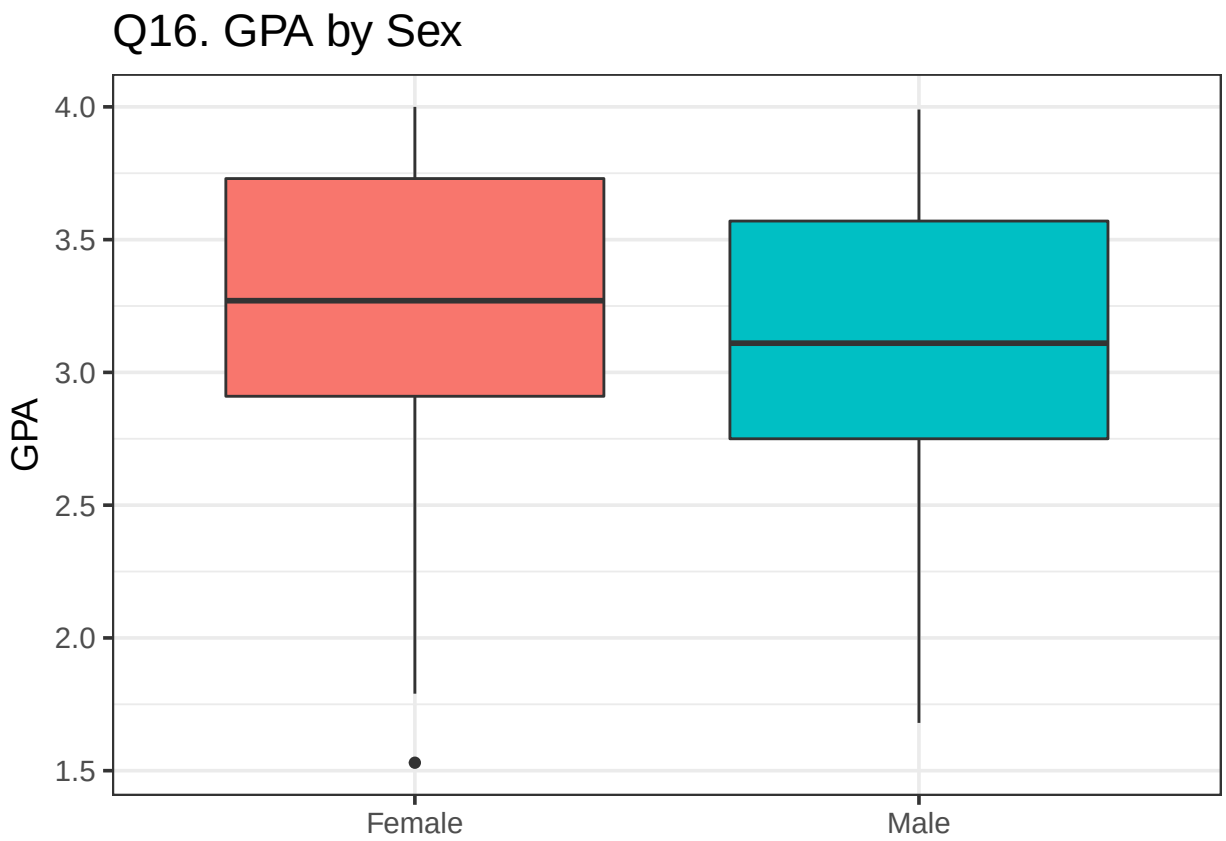
```
cat(paste0('Q15. A 95% confidence interval for this estimated difference:\n  (',
    paste(round(conf3[2,], 2), collapse = ', '),  ')'))
```

```
## Q15. A 95% confidence interval for this estimated difference:
##   (-0.29, -0.09)
```

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, levels = c(0, 1), labels = c('Female', 'Male')))
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

Table 2: Q17. Proportions of Drinkers by Sex

Drinking Status	# Non-drinker	# Drinker	% Non-drinker	% Drinker
Female	108	209	34.1%	65.9%
Male	56	127	30.6%	69.4%

```
sxdrnk_df <- casData %>%
  count(Sex, DRINK) %>%
  group_by(Sex) %>%
  spread(DRINK, n) %>%
  rename_with(~ paste0("# ", .x))
sxdrnk_df2 <- casData %>%
  count(Sex, DRINK) %>%
  group_by(Sex) %>%
  mutate(prop = paste0(round(n/sum(n)*100,1),"%")) %>%
  select(-n) %>%
  spread(DRINK, prop) %>%
  rename_with(~ paste0("% ", .x))
sxdrnk_df <- cbind(sxdrnk_df, sxdrnk_df2[2:3])
colnames(sxdrnk_df)[1] <- 'Drinking Status'
knitr::kable(sxdrnk_df, 'latex', align = 'c',
             caption = 'Q17. Proportions of Drinkers by Sex')
```

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

```
cat(paste0('Q18. The estimated mean difference in GPA for students who drink, compared to\n',
    '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(paste0('Q19. A 95% confidence interval for this estimated difference:\n  (',
    paste(round(conf4[2,], 2), collapse = ', '), ')\n'))
```

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

```
cat(paste0('Q20. The estimated mean difference in GPA for male students, compared to female\n',
    '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
```

Table 3: Q24. Counts, Proportions of HiExp by Smoking status

Health Care Costs	Have not smoked	Have smoked	Have not smoked	Have smoked
Not high expenditures	1998	1879	95.9%	94.2%
High expenditures	86	115	4.1%	5.8%

```
cat(paste0('Q21. A 95% confidence interval for this estimated difference:\n  (',
    paste(round(conf4[3,], 2), collapse = ', '), ')'))
```

```
## Q21. 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

Exercise 2

Part A: What is the unadjusted relationship between high expenditures and smoking?

```
nmesData <- read.csv("nmesData.csv")
# head(nmesData)
nmesData$eversmk <- factor(nmesData$eversmk, levels = c(0, 1),
                           labels = c('Have not smoked','Have smoked'))
nmesData$highexp <- factor(nmesData$highexp, levels = c(0, 1),
                           labels = c('Not high expenditures','High expenditures'))
# table of high expenditures indicator by smoking status
smkhiexp_df <- nmesData %>%
  count(eversmk, highexp) %>%
  group_by(eversmk) %>%
  spread(eversmk, n)
smkhiexp_df2 <- nmesData %>%
  count(eversmk, highexp) %>%
  group_by(eversmk) %>%
  mutate(prop = paste0(round(n / sum(n)*100,1),"%")) %>%
  select(-n) %>%
  spread(eversmk, prop)
smkhiexp_df <- cbind(smkhiexp_df, smkhiexp_df2[,2:3])
colnames(smkhiexp_df)[1] <- 'Health Care Costs'
knitr::kable(smkhiexp_df, format = 'latex', align = 'c',
             caption = 'Q24. Counts, Proportions of HiExp by Smoking status')
```

```
cat(paste0('Q24a. Number and proportion of smokers with high health expenditures\n  ',
           smkhiexp_df[2,3], ', ', smkhiexp_df[2,5]), '\n')
```

```
## Q24a. Number and proportion of smokers with high health expenditures
##   115, 5.8%
```

```
cat(paste0('Q24b. Number and proportion of nonsmokers with high health expenditures\n  ',
           smkhiexp_df[2,2], ', ', smkhiexp_df[2,4]))
```

```
## Q24b. Number and proportion of nonsmokers with high health expenditures
##   86, 4.1%
```

```
model5 <- glm(highexp ~ eversmk, family=binomial(link=logit), data=nmesData)
#summary(model5)

co5 <- coefficients(model5)
conf5 <- confint(model5)
```

```
## Waiting for profiling to be done...
```

```
od_smk <- exp(co5[1] + co5[2]*1)
cat(paste0('Q25. The estimated probability of a high expenditure for a smoker\n  ',
           round(od_smk / (1 + od_smk), 3)), '\n')
```

```
## Q25. The estimated probability of a high expenditure for a smoker
##   0.058
```

```
cat(paste0('Q26. The estimated probability of a high expenditure for a non-smoker\n  ',
           round(exp(co5[1] / (1 + exp(co5[1]))), 3)), '\n')
```

```
## Q26. The estimated probability of a high expenditure for a non-smoker
##   0.049
```

```
cat(paste0('Q27. The estimated odds ratio of a high expenditure, comparing smokers to ',
           'nonsmokers\n  ', round(exp(co5[2]), 3)), '\n')
```

```
## Q27. The estimated odds ratio of a high expenditure, comparing smokers to nonsmokers
##   1.422
```

```
cat(paste0('Q28. A 95% confidence interval for this true odds ratio of high expenditure, ',
           '\ncomparing smokers to nonsmokers\n  (',
           paste(round(exp(conf5[2,]),3), collapse = ', '), ')'))
```

```
## Q28. A 95% confidence interval for this true odds ratio of high expenditure,
## comparing smokers to nonsmokers
##   (1.069, 1.898)
```

Part B: What is the relationship between high expenditures and smoking when adjusted for an individual's age?

```r
model6 <- glm(highexp ~ eversmk + age, family=binomial(link=logit), data=nmesData)
# summary(model6)
co6 <- coefficients(model6)
conf6 <- confint(model6)
```

```
## Waiting for profiling to be done...
```

```r
od_smk50 <- exp(co6[1] + co6[2]*1 + co6[3]*50)
cat(paste0('Q29. The estimated probability of a high expenditure for a 50-year-old smoker\n  ',
           round(od_smk50 / (1 + od_smk50), 3), '\n'))
```

```
## Q29. The estimated probability of a high expenditure for a 50-year-old smoker
##   0.048
```

```r
od_nosmk60 <- exp(co6[1] + co6[3]*60)
cat(paste0('Q30. The estimated probability of a high expenditure for a 60-year-old ',
           'non-smoker\n  ', round(od_nosmk60 / (1 + od_nosmk60), 3), '\n'))
```

```
## Q30. The estimated probability of a high expenditure for a 60-year-old non-smoker
##   0.05
```

```r
cat(paste0('Q31. The estimated odds ratio of a high expenditure, comparing smokers to ',
           'nonsmokers\nof the same age\n  ', round(exp(co6[2]),3), '\n'))
```

```
## Q31. The estimated odds ratio of a high expenditure, comparing smokers to nonsmokers
## of the same age
##   1.54
```

```r
cat(paste0('Q32. A 95% confidence interval for this estimated odds ratio\n  (',
           paste(round(exp(conf6[2, ]),3), collapse = ', '),')\n'))
```

```
## Q32. A 95% confidence interval for this estimated odds ratio
##   (1.149, 2.073)
```

```r
cat(paste0('Q33. The estimated odds ratio of a high expenditure for two groups of the ',
           'same\nsmoking status but whose age differs by 10 years\n  ',
           round(exp(co6[3]*10),3), '\n'))
```

```
## Q33. The estimated odds ratio of a high expenditure for two groups of the same
## smoking status but whose age differs by 10 years
##   1.63
```

```r
cat(paste0('Q34. A 95% confidence interval for this estimated odds ratio\n  (',
           paste(round(exp(conf6[3,]*10), 3), collapse = ', '),')'))
```

```
## Q34. A 95% confidence interval for this estimated odds ratio
##   (1.498, 1.778)
```

```
paste0(round(exp(co6[2]),3), ' (', paste(round(exp(conf6[2,]),3), collapse = ', '), ')')
```

```
## [1] "1.54 (1.149, 2.073)"
```

Q35. According to data collected from 4078 individuals in the 1987 National Medical Expenditures Survey (NMES), the status of ever having smoked is associated with an estimated 54% greater odds of higher health care expenditures (95% CI: 14.9%, 107.3%) when adjusted for age. This is a significant finding at the 0.05 level.

Q36. See above