HW3

Greg Finkelberg

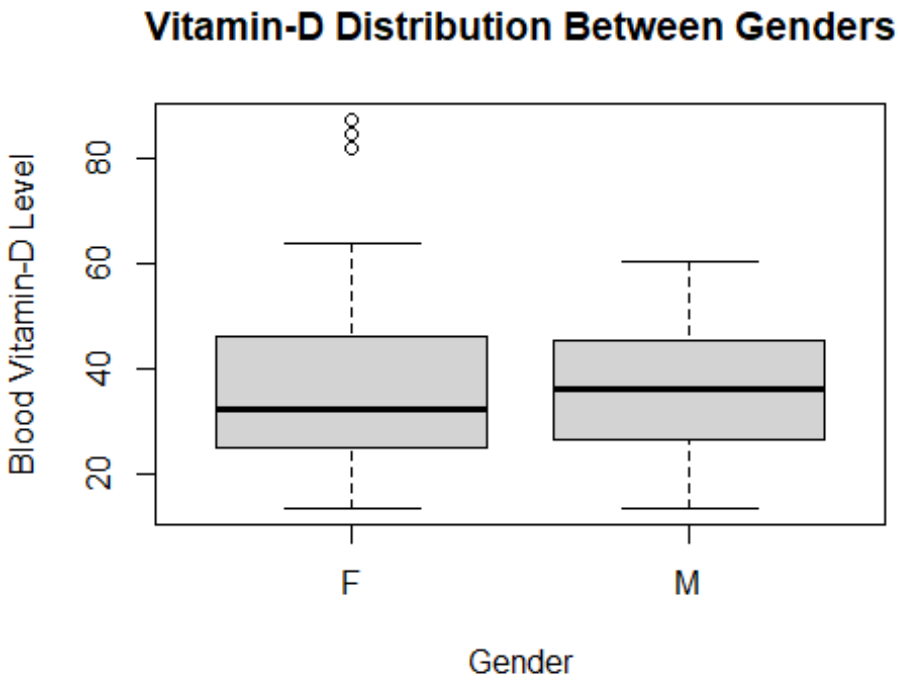
2023-10-21

R packages

```
library(tidyverse)
library(mosaic)
library(quarto)
```

1. Box plot comparison

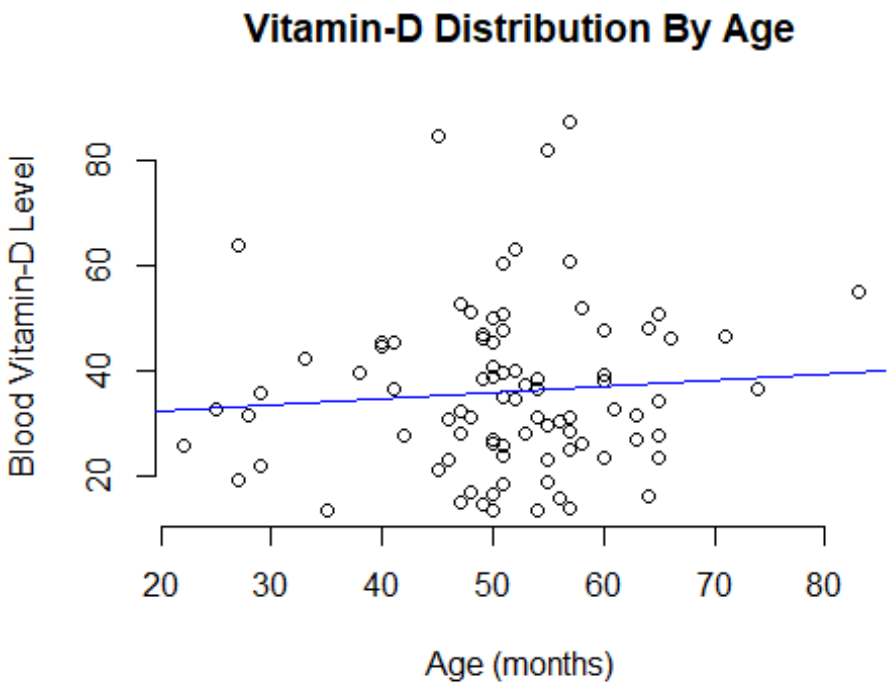
```
data1_LSC598 <- read.csv("C:/Users/Greg/Downloads/data1_LSC598.txt", sep="")

df <- data1_LSC598 %>% mutate(autism = case_when(group == '0' ~ 'no', group
== '1' ~ "yes"))

vitD <- df %>% drop_na(vitD_level)

boxplot(vitD_level ~ gender, data = vitD, main = "Vitamin-D Distribution
Between Genders", xlab = "Gender", ylab = "Blood Vitamin-D Level")
```



Visually, there does not appear to be much of a difference in the distribution of vitamin-D levels based on gender. The mean vitamin-d level for males is slightly lower than females, but if we were to remove the outliers from the analysis this gap would likely shrink.

2. Scatter plot

```
plot(x = vitD$age_month, y = vitD$vitD_level, main = "Vitamin-D Distribution
By Age", xlab = "Age (months)", ylab = "Blood Vitamin-D Level", frame =
FALSE)
abline(lm(vitD_level ~ age_month, data = vitD), col = "blue")
```




```
cor(x = vitD$age_month, y = vitD$vitD_level)

[1] 0.08259572
```

There seems to be almost no correlation between age and vitamin-D level. A correlation coefficient of 0.08 is very weak. The line on the scatter plot has a slightly positive slope, but it is close to being horizontal which indicates that there is not much of a trend.

3. Hypothesis test (Vitamin-D ~ Gender)

Null hypothesis: The true difference in mean vitamin-D level between genders equals 0.
Alternative hypothesis: The true difference in mean vitamin-D level between genders does not equal 0.

Based only on the box plot comparison from part 1, I would guess that mean vitamin-D level are not significantly different between genders (i.e. we would fail to reject the null

hypothesis). Since the samples are independent and the population standard deviation is not known, I ran a two-sample t-test at a 95% confidence interval (0.05 significance level).

```
t.test(vitD_level ~ gender, paired = FALSE, data = vitD)


    Welch Two Sample t-test

data:  vitD_level by gender
t = 0.10163, df = 58.166, p-value = 0.9194
alternative hypothesis: true difference in means between group F and group M
is not equal to 0
95 percent confidence interval:
 -6.196153  6.859056
sample estimates:
mean in group F mean in group M
       36.05645        35.72500
```

As I suspected, the p-value of 0.9194 is well above 0.05 meaning we would fail to reject the null hypothesis. The means do not appear to be significantly different.

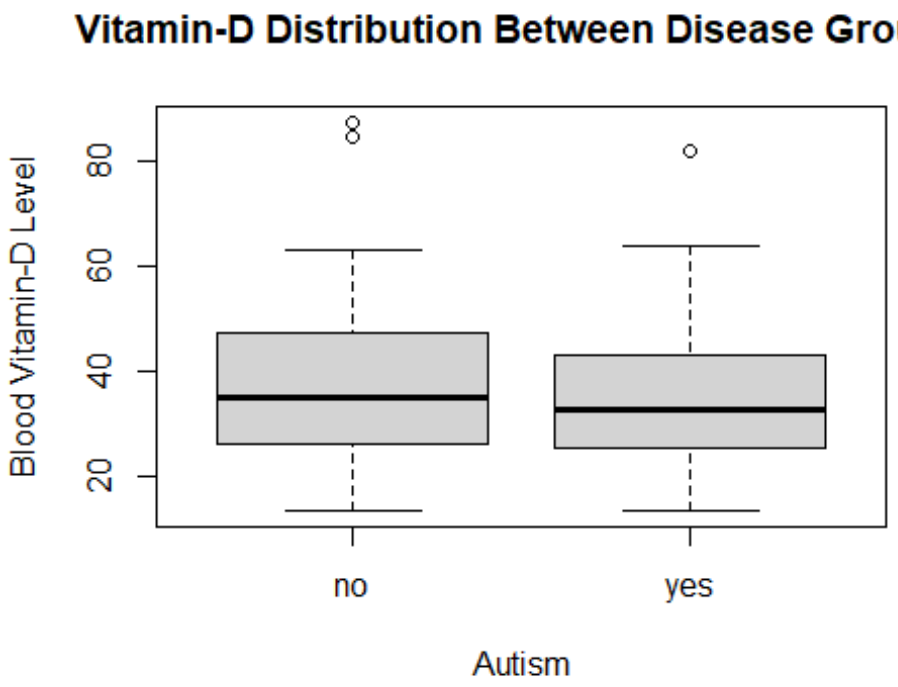
4. Hypothesis test (Vitamin-D ~ Autism)

Null hypothesis: The true difference in mean vitamin-D level between disease groups equals 0.
Alternative hypothesis: The true difference in mean vitamin-D level between disease groups does not equal 0.

Visually, the box plot comparison of vitamin-D levels between disease groups looks almost identical. Therefore I would expect again that we would fail to reject the null hypothesis as the true means will likely not be significantly different.

Since the samples are independent and the population standard deviation is not known, I ran a two-sample t-test at a 95% confidence interval (0.05 significance level).

```
boxplot(vitD_level ~ autism, data = vitD, main = "Vitamin-D Distribution
Between Disease Groups", xlab = "Autism", ylab = "Blood Vitamin-D Level")
```




```
t.test(vitD_level ~ autism, data = vitD)


    Welch Two Sample t-test

data:  vitD_level by autism
t = 0.89425, df = 64.73, p-value = 0.3745
alternative hypothesis: true difference in means between group no and group
yes is not equal to 0
95 percent confidence interval:
 -3.896267 10.213690
sample estimates:
 mean in group no mean in group yes
         37.83714          34.67843
```

The p-value of 0.3745 is lower than the t-test comparing gender and vitamin-D levels, but it is still too high to reject the null hypothesis. There does not appear to be a significant difference between mean vitamin-D levels of the disease groups.