Module 3: Assignment

Aaron Hum

Stephanie Moyerman, PhD

LSC 541: Statistics for Biological Data Science I

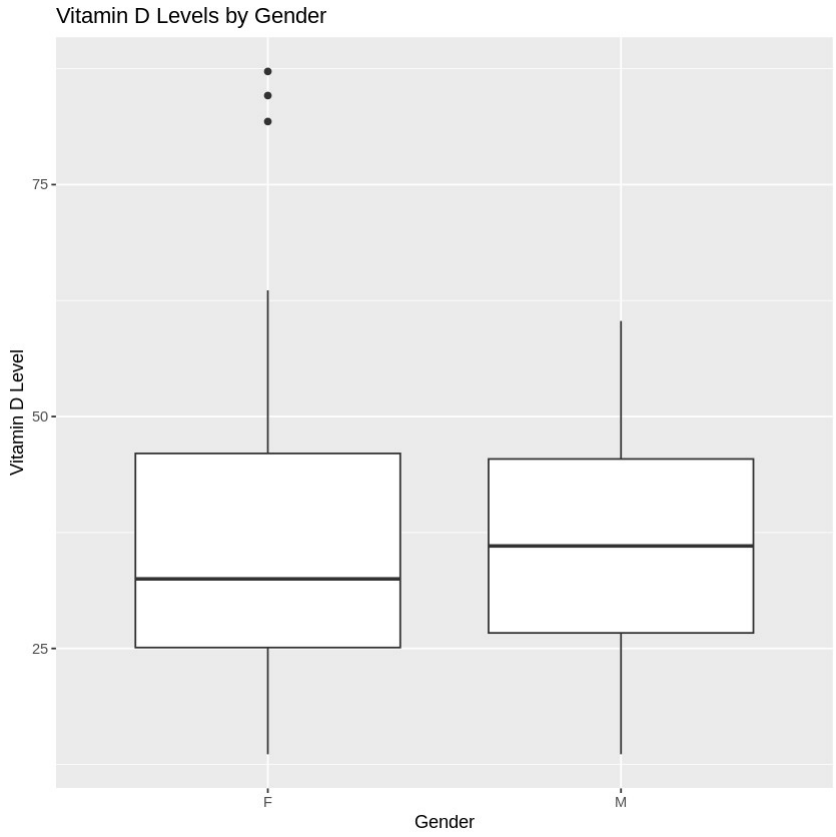
2024-07-16

```
# load library
library(ggplot2)

# read in data
data <- read.table('data1_LSC598.txt', header = T)

# omit null values
data <- na.omit(data)

data
```

```
   age_month gender vitD_level group
1  60        F      47.5       1
2  50        M      39.0       1
3  35        F      13.6       1
4  50        F      16.7       1
5  61        F      32.7       0
6  55        F      23.2       0
7  54        F      36.4       0
8  60        F      38.1       1
9  47        F      28.2       1
10 65        M      34.1       1
11 22        F      25.7       1
12 52        M      39.9       1
13 50        F      45.2       1
14 51        F      47.8       0
15 33        F      42.4       0
16 60        F      39.3       1
17 65        F      23.4       1
18 74        F      36.6       0
19 51        F      18.4       1
20 57        F      87.2       0
21 51        F      35.1       0
22 54        F      38.5       1
23 53        F      28.2       1
24 56        F      30.6       1
25 47        F      52.5       0
26 46        M      23.3       0
```

```
27 41         M       45.4        1
28 49         F       38.6        0
29 66         M       46.0        1
30 48         M       51.3        1
⋮  ⋮          ⋮       ⋮           ⋮
57 57         F       13.8        1
58 57         F       24.9        1
59 40         M       45.5        1
60 57         M       28.6        0
61 27         F       63.6        1
62 45         F       21.1        0
63 27         F       19.2        1
64 51         M       24.0        0
65 50         M       50.1        0
66 50         F       26.1        1
67 64         F       16.1        1
68 52         M       34.7        1
69 56         M       16.0        0
70 63         M       26.8        0
71 53         M       37.4        1
72 38         F       39.7        0
73 40         M       44.8        0
75 54         F       31.3        0
76 25         F       32.7        1
77 63         F       31.5        1
78 50         M       40.6        1
79 64         M       48.0        0
80 51         F       39.5        1
81 49         F       46.3        1
82 57         F       31.3        1
83 50         F       27.0        1
84 71         F       46.6        0
85 51         F       50.9        1
86 48         M       31.3        0
87 47         F       15.1        0
```

```
# 1. side-by-side boxplot for vitD_level by gender
ggplot(data, aes(x = gender, y = vitD_level)) +
  geom_boxplot() +
  labs(title = "Vitamin D Levels by Gender", x = "Gender", y =
"Vitamin D Level")
```



Interpretation

Median:

The median vitamin D levels for males and females are quite similar.

Variability:

Both genders exhibit similar variability in vitamin D levels as indicated by the similar IQR and whisker ranges.

Outliers:

The presence of outliers in the female group suggests that there are a few females with vitamin D levels significantly higher than the rest.

```
# 2b. calculate the correlation coefficient
correlation <- cor(data$age_month, data$vitD_level)
correlation
```

```
[1] 0.08259572
```

Interpretation

The correlation of ~0.0826 indicates a very weak positive linear relation ship betwen age_month and vitD_level. However, the value is extremely low, close to zero, meaning the increase is minimal and not practically significant.

Practical Significance:

Given that the correlation is so close to zero, age and vitamin D levels are essentially independent in this dataset and any observed relationship too weak to be meaningful.

```
# 3. perform t-test for vitD_level by gender
t_test_gender <- t.test(vitD_level ~ gender, data = data)
t_test_gender
```

```
     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
```

Interpretation

No Significant Difference:

The p-value (0.9194) is much greater than 0.05, indicating no statistically significant difference in vitamin D levels between females and males.

Confidence Interval Includes Zero:

The 95% confidence interval [-6.196153, 6.859056] includes zero, supporting the conclusion that the true difference in means could be zero.

Similar Means:

The mean vitamin D levels for females (36.05645) and males (35.72500) are very close, showing little practical difference.

```
# 4. perform t-test for vitD_level by group
t_test_group <- t.test(vitD_level ~ as.factor(group), data = data)
t_test_group
```

```
     Welch Two Sample t-test

data:  vitD_level by as.factor(group)
t = 0.89425, df = 64.73, p-value = 0.3745
alternative hypothesis: true difference in means between group 0 and
group 1 is not equal to 0
95 percent confidence interval:
 -3.896267 10.213690
sample estimates:
mean in group 0 mean in group 1
       37.83714        34.67843
```

Interpretation:

No Significant Difference:

The p-value (0.3745) is greater than 0.05, indicating no statistically significant difference in vitamin D levels between healthy and autism groups.

Confidence Interval Includes Zero:

The 95% confidence interval [-3.896267, 10.213690] includes zero, supporting the conclusion that the true difference in means could be zero.

Similar Means:

The mean vitamin D levels for healthy (37.83714) and autism (34.67843) groups are close, showing little practical difference.

Conclusion of t-tests

Overall, neither gender nor health status (healthy vs. autism) significantly affect vitamin D levels in this dataset.