ST517-HW1

Put your name here

```
sport_heights <- read.csv("sport_heights.csv")
```

1. Are basketball, baseball, and soccer players the same height on average? Suppose we take a random sample of 50 players from each of the three sports. Load the sample data `sport_heights.csv` provided with the homework file.

   Now you will perform an Analysis of Variance F-test step by step. The code in the second lecture this week will be helpful.

   (a) State the null and alternative hypothesis, in statistical notation, for testing whether players from the three sports have the same mean height.

   $$H_0 : \mu_1 = \mu_2 = \mu_3$$

   $$H_A : \textit{The } \mu_i \textit{ vary across some of the populations}$$

   (b) Add columns to `heights` for the overall average height and the average height for the sport of each player.

```
Height <- sport_heights$height
Sport <- sport_heights$sport

sport_heights$overall_mean <- with(sport_heights, mean(Height))

sport_heights$group_mean <- with(sport_heights, ave(Height, Sport))
```

   (c) Calculate the between group sum of squares and within group sum of squares and their corresponding degrees of freedom.

```
between_group_SS <- with(sport_heights, sum((sport_heights$group_mean - sport_heights$overall_mean)^2))

within_group_SS <- with(sport_heights, sum((Height - sport_heights$group_mean)^2))
```

   (d) Calculate the F-statistic, and give a p-value.

```
I <- length(unique(sport_heights$sport))
N <- nrow(sport_heights)

F_stat <- (between_group_SS/(I-1))/(within_group_SS/(N-I))
F_stat
```

```
## [1] 1.004516
```

(e) Now use `oneway.test()` to verify your answer. Use `var.equal = TRUE` and you should get the same answer.

```
oneway.test(Height ~ Sport, data = sport_heights, var.equal = T)
```

```
##
##  One-way analysis of means
##
## data:  Height and Sport
## F = 1.0045, num df = 2, denom df = 147, p-value = 0.3687
```

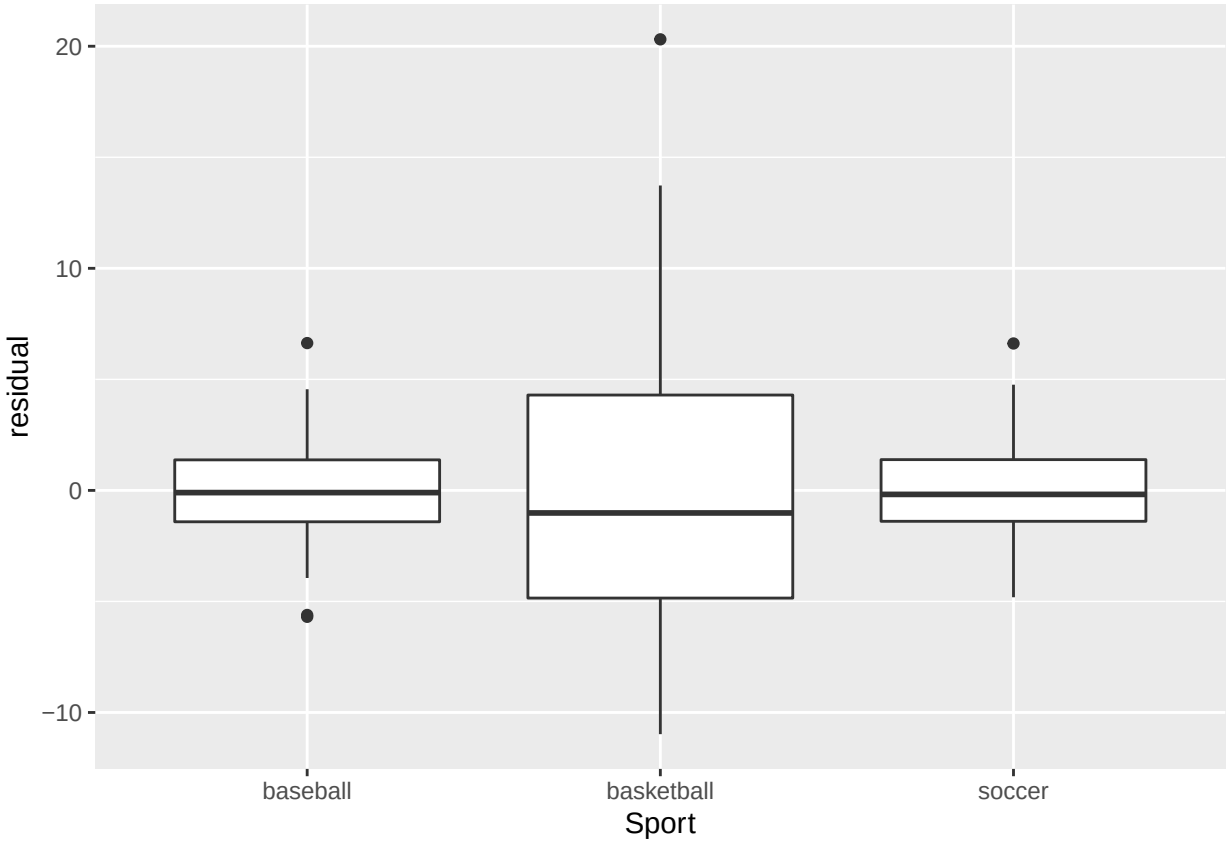
(f) In two sentences or less, what do you conclude?

Due to the F-stat of 1.0045 we can see that there is very little variation between the different groups we are looking at. Based on the F-stat and the p-value I conclude that there is very little difference in player heights between the three sports.

(g) Create a column of the residuals in `heights` by subtracting the average height from the sport of each player (i.e. the second column you created in (b)) from the `height` column. Create side-by-side box-plots of these residuals. Do you think the equal variance assumption is violated?

```
sport_heights$residual <- Height - sport_heights$group_mean

qplot(Sport, residual, data = sport_heights, geom = "boxplot")
```



I think the equal variance assumption is violated because while baseball and soccer have about the same variance basketball has a much larger variance ruining the assumption that they will all have the same variance.

(h) Use `oneway.test()` again, this time with `var.equal = FALSE`. Give the F-statistic, p-value, and denominator degrees of freedom. Does the test indicate a different conclusion?

```
oneway.test(Height ~ Sport, data = sport_heights, var.equal = F)
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  Height and Sport
## F = 2.5609, num df = 2.000, denom df = 90.256, p-value = 0.08284
```

The F-stat is 2.56, the p-value is 0.08, and the degrees if freedom is 90.256.
Yes, I beleive that that this data comes to a different conclusion based on p-value and df.

(i) In actuality, the data is generated from distributions with slightly different means, and non-constant variances. In two sentences or less, comment on what the difference in conclusions from the two tests. Does this tell us anything about the robustness of the F-test to non-constant spreads?

The assumption of them having the same variation changes the data to show that there is almost no difference because we believe there is almost no difference. Yes, this tells us we cannot assume for the robustness and spreads.

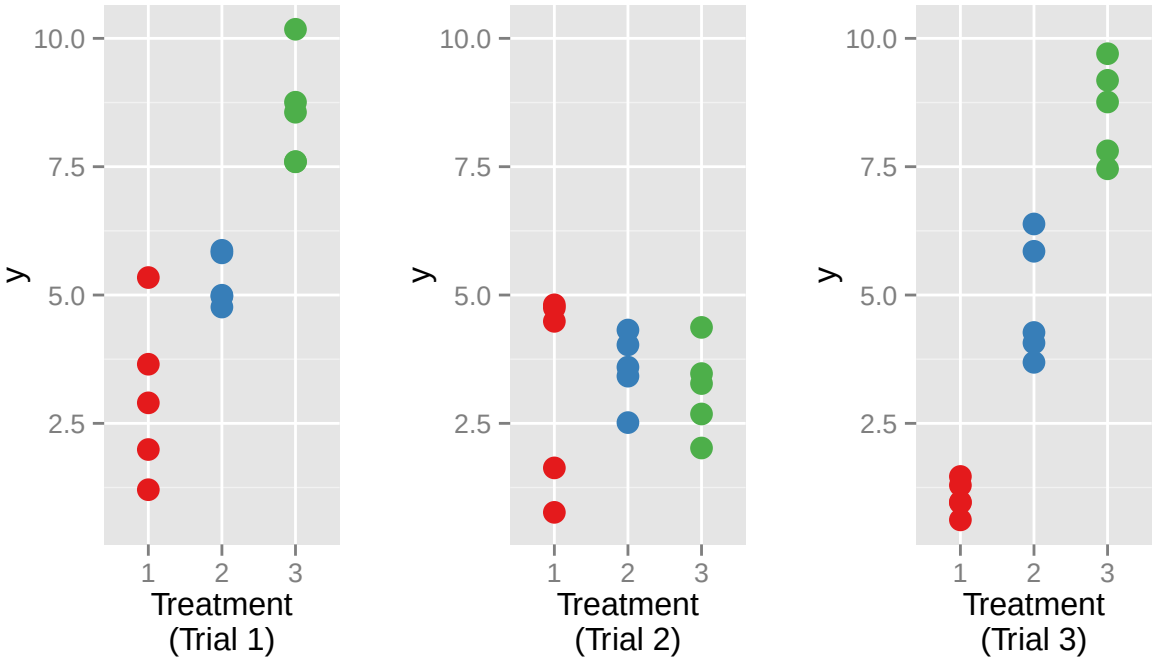
(j) Food for thought: how would you perform a simulation to evaluate the robustness of the F-test to violation of the equal variance assumption? (You do not need to do this for the homework!)

I would manually make sure that there are a variety of variances within the data to violate the equal variance assumption.

2. Using the data from the lab (`case0501` in the `Sleuth3` package), answer the question "Are there differences between the diets in their effect on lifetime?"

   From the data we hypothesize that there is no difference in lifespane based on diet. Due to the data and graphs we can see that there is an increase in average lifespan based on certain diets, therefore, we can summarize that our hypotheses is incorrect. There is no validity to the assumption that all lifespans would be the same no matter the diet because we can see that there is a large variation in lifespans.

3. The following plot represents three trials of an experiment in which there are three treatments and five responses in each treatment. Without doing any calculations, order the trials in increasing order of F-statistic. Explain your reasoning.



Treatment 2, treatment 1, and treatment 3. I chose this order because treatment 2 has the most overlap so it's F-stat should be small because there is not much variation. Treatment 1 is next because there is some overlap but some variation. Finally, treatment 3 would have the largest F-stat because there is so much variation in the data and no overlap.

4. When calculating the denominator of the F-statistic, a.k.a. the pooled variance, why do we not simply average the group-level sample variances?

   I believe we cannot do that because that would assume too much about the variances. If we were to simply average the group-level sample variances we would be seperating the variances out instead of viewing them all together.