ST 517: Data Analytics I
Module 1 Homework

1. (3 points) 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 files (you will need to download it onto your computer first):

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

   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.

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

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

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

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

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

   (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?

   (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?

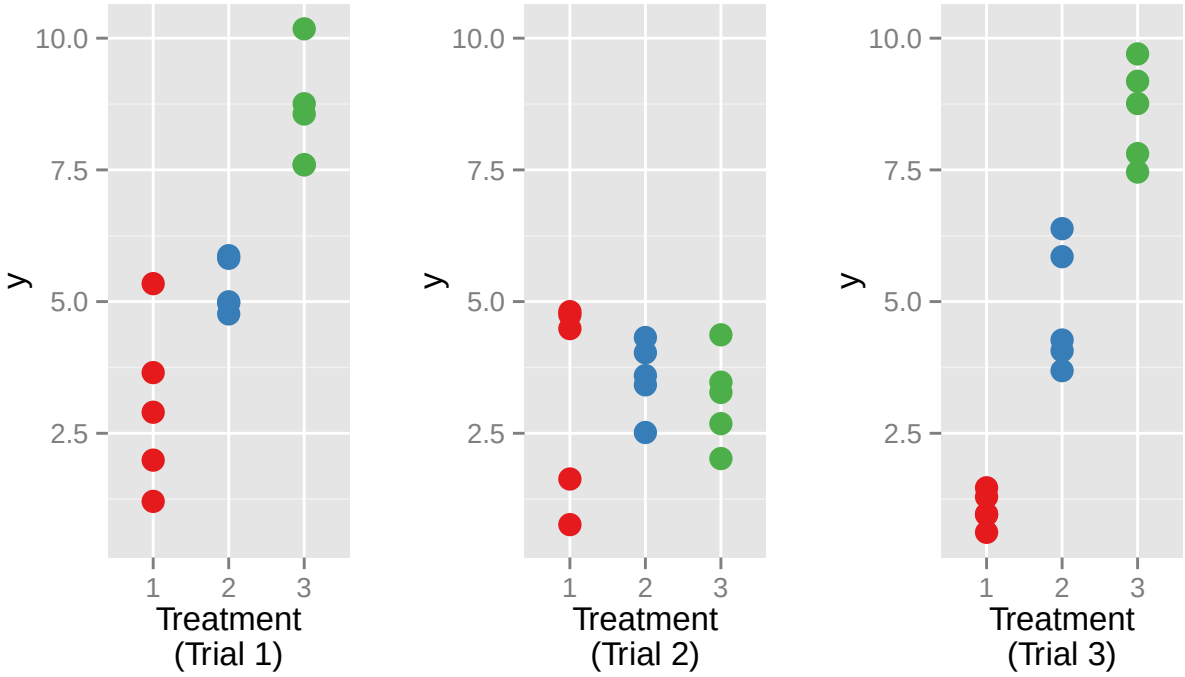
   (i) In actuality, the data is generated from distributions with slightly different means, and non-constant variances. In two sentences or less, comment on what the difference in conclusions from the two tests. Does this tell us anything about the robustness of the F-test to non-constant spreads?

   (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!)

2. (3 points) 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?"

   Make sure you clearly state your hypotheses, summarize your calculations, comment on the validity of the assumptions, and include a short summary of your results in non-technical language.

3. (2 points) 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.



4. (2 points) 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?