ST517-HW1

Ben Tankus

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.

```
sportHeights <- read.csv(file = 'sport_heights.csv')
```

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.

$$H_0 : \mu_{baseball} = \mu_{basketball} = \mu_{soccer}$$

$$H_A : \mu_{baseball} \neq \mu_{basketball} \neq \mu_{soccer}$$

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

```
   # This is an R comment. Add your R code below.
#avgHeights
sportMeans <- tapply(sportHeights$height, sportHeights$sport, mean)
totalMean <- mean(sportMeans)

sportHeights$overall_mean <- with(sportHeights, mean(height))
sportHeights$group_mean <- with(sportHeights, ave(height,sport))


head(sportHeights)
```

```
##   player      sport   height overall_mean group_mean
## 1      1 basketball 70.20227     73.60851    73.2537
## 2      2 basketball 77.43825     73.60851    73.2537
## 3      3 basketball 76.68239     73.60851    73.2537
## 4      4 basketball 72.99433     73.60851    73.2537
## 5      5 basketball 71.44031     73.60851    73.2537
## 6      6 basketball 72.96603     73.60851    73.2537
```

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

```r
   # Put R code below.

betweenSS <- with(sportHeights, sum((group_mean - overall_mean)^2))

withinSS <- with(sportHeights, sum((height - group_mean)^2))

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

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

```r
   # Put R code below.

f = (betweenSS / (I-1)) / (withinSS / (N-I))
pval <- 1 - pf(f, I-1, N-1 )
print(paste("F-stat:", f))
```

```
## [1] "F-stat: 1.00451635949283"
```

```r
print(paste("P-Value:", pval))
```

```
## [1] "P-Value: 0.368688023746846"
```

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

```r
   # Put R code below.
test <- oneway.test(height~sport, data = sportHeights, var.equal = T)

test
```

```
##
##  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?

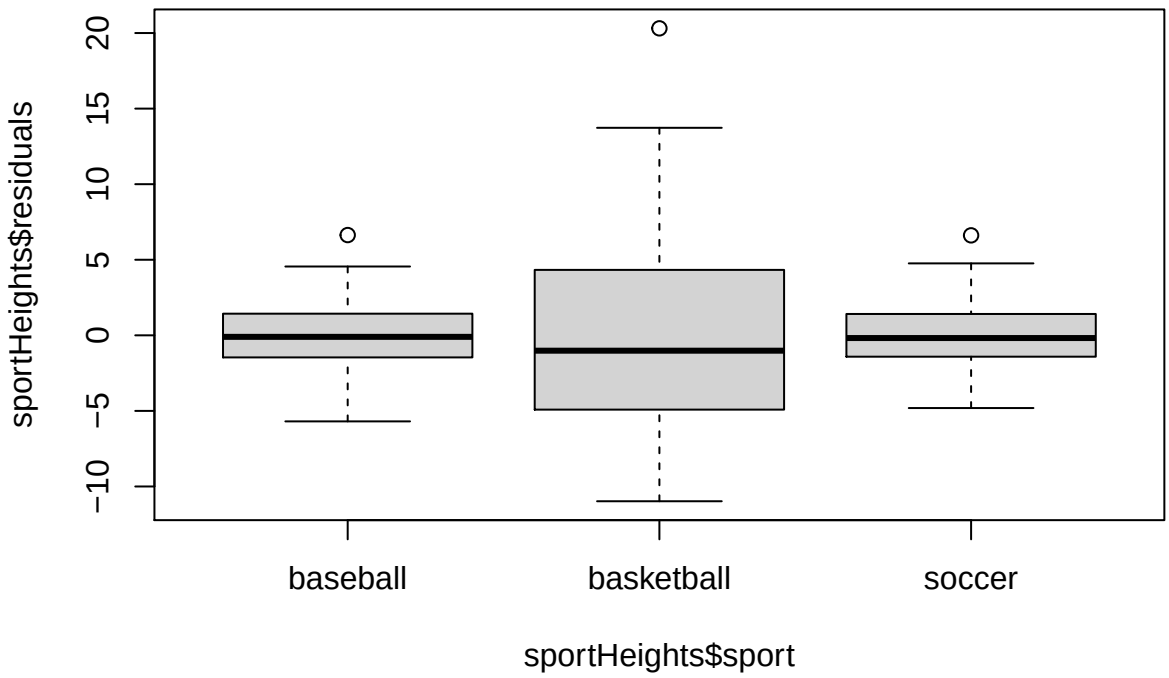
We do not have enough evidence to reject the null hypothesis. This means that the sports have players of similar heights.

(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?

```r
   # Put R code below.

sportHeights['residuals'] <- with(sportHeights, height - group_mean)

boxplot(sportHeights$residuals~sportHeights$sport)
```



The equal variance assumption is likely violated as the basketball variance is much greater than baseball and soccer.

(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?

```
   # Put R code below.
oneway.test(height~sport, data = sportHeights, 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 var.equal = False test still doesn't give convincing evidence to reject the null hypothesis, but it is very close.

(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 different conclusions show that the one-way ANOVA test is not robust 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!)

To evaluate the robustness we could replicate this same procedure many times for a known value, and then compare it to a simulated value.

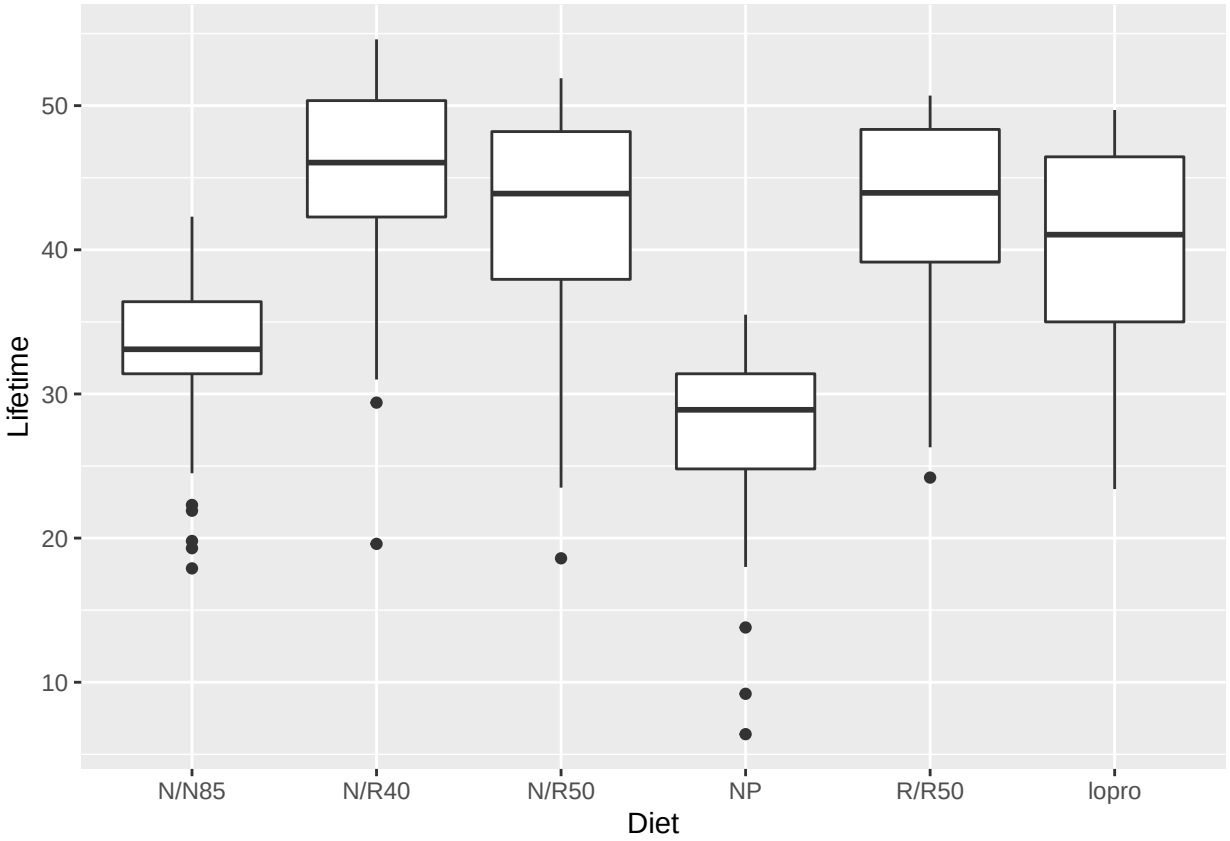
2. Using the data from the lab (|case0501| in the |Sleuth3| package), answer the question "Are there differences between the diets in their effect on lifetime?"

$$H_0 : \mu_{diet1} = \mu_{diet2} = \mu_{diet3} = \mu_{diet4} = \mu_{diet5} = \mu_{diet6}$$
$$H_A : \mu_{diet1} \neq \mu_{diet2} \neq \mu_{diet3} \neq \mu_{diet4} \neq \mu_{diet5} \neq \mu_{diet6}$$

```
# due to low pvalue var.equal can be either true or false and yield the same result (as seen in lab)
oneway.test(Lifetime ~ Diet, data = case0501, var.equal = T)
```

```
##
##  One-way analysis of means
##
## data:  Lifetime and Diet
## F = 57.104, num df = 5, denom df = 343, p-value < 2.2e-16
```

```
qplot(Diet, Lifetime, data = case0501, geom = "boxplot")
```



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

with a p-value of 2.2e-16 we have strong evidence to reject the null hypothesis that the means are equal. Looking at the boxplot it looks like most distributions are close enough to normal to satisfy normallity assumption. Reading the study introduction also leads me to belive that the data are independant, and measurements are representative of the study.

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.

Trial 2 -> Trial 1 -> Trial 3

Sum of Squares is based off variability, either within or between samples. The simplified formula to calculate the f-statistic is a simple fraction of BetweenSS / WithinSS, meaning as variability between samples increases, the f-stat will go up, while the opposite is true with variability within samples increases.

Trial 2 has large variability within samples and small varability between samples, so that has the lowest f-stat. Trial 3 has high varability between samples, and low variability within samples, so that is the highest f-stat, and trial 1 is somewhere in-between.

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?

One cannot simply average the group-level variances because pooled variance is the *degrees of freedom weighted* variance. The DF weight is lost when doing a simple average.