STA-290 HW 8

```r
library(ggplot2)
library(dplyr)

# Prettier graphs
theme_set(theme_bw())
```

Question 1

In professional basketball games during the 2009-2010 season, when Kobe Bryant of the Los Angeles Lakers shot a pair of free throws, 8 times he missed both, 152 times he made both, 33 times he made only the first shot, and 37 times he made only the second. Is it possible that the successive free throws are independent, or is there evidence to suggest a "hot streak" effect? The data are tabulated in the `freethrow` data frame below:

```r
# Create freethrow data (copy and paste this into your own R session)
freethrow <- matrix(c(152,33,37,8), nrow = 2, byrow = TRUE)
rownames(freethrow) <- c("Make 1st", "Miss 1st")
colnames(freethrow) <- c("Make 2nd", "Miss 2nd")
print(freethrow)
```

```
##          Make 2nd Miss 2nd
## Make 1st      152       33
## Miss 1st       37        8
```

1. What is the null hypothesis of this experiment?
2. Using the table provided, find a table of *expected values* for each cell
3. Using your table of observed and expected values, find the χ^2 statistic associated with this table along with the degrees of freedom
4. Using your critical value sheet, if we were to test this hypothesis at level $\alpha = 0.05$, what conclusion would we come to regarding the independence of the first and second free throw?

Null: Independence

```r
## Expected values
chisq.test(freethrow)[["expected"]]
```

```
##          Make 2nd Miss 2nd
## Make 1st  152.022  32.9783
## Miss 1st   36.978   8.0217
```

```r
## chisq stat
chisq.test(freethrow)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  freethrow
## X-squared = 0.0000000000000000000000000000393, df = 1, p-value = 1
```

Based on this, we would reject.

Question 2

Reconsider the anorexia data that we investigated in Homework 7:

```
anorexia <- read.csv("https://collinn.github.io/data/anorexia.txt")
```

- **Part A:** Use the `mutate` function to again create a variable called `Diff` that records the difference in pre and post weights
- **Part B:** State the null hypothesis for testing the difference and pre and post weights for each of the groups considered in the dataset
- **Part C:** Perform an ANOVA for the hypothesis stated in Part B. What do you conclude?
- **Part D:** Use *post-hoc* testing to determine if there are any pairwise differences between these groups. How do your findings here compare with the conclusions you had in Homework 7?

```
## Part 1
anorexia <- mutate(anorexia, diff = Postwt - Prewt)
```

Null: no difference in treatment between all groups

```
## Based on this, there appears to be different in treat
aov(diff ~ Treat, anorexia) %>% summary()
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Treat        2    612   306.1     5.4 0.0066 **
## Residuals   69   3911    56.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Looking at post-hoc testing, we identify difference in FT vs Control, just as we did after bonferonni correction in Homework 7

```
TukeyHSD(aov(diff~Treat, anorexia))
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = diff ~ Treat, data = anorexia)
##
## $Treat
##                 diff     lwr     upr   p adj
## Control-CBT -3.2069 -8.0773  1.6635 0.26231
## FT-CBT       4.5078 -1.0006 10.0162 0.12990
## FT-Control   7.7147  2.0901 13.3393 0.00451
```

Question 3

This question will again consider the `mtcars` dataset built into R

```
data(mtcars)
```

We will be investigating the relationship between the weight of a car (independent variable) and its miles per gallon (dependent variable). In addition to this, we will also be using the number of carburetors as a second independent variable.

- **Part A:** Create a linear model predicting `mpg` with the covariates `wt` and `carb`. Based on the results, does it appear that the number of carburetors has a relationship with fuel economy (mpg)?
- **Part B:** By default, `carb` is stored in the dataset as an *integer* value. Use the `mutate` function to create a new variable in the `mtcars` dataset called `carb_factor` that is equal to `carb_factor = fator(carb)`.

This will turn the new variable into a *categorical* one instead of an integer

- **Part C:** Create a new linear model, this time predicting `mpg` with `wt` and `carb_factor`. What has changed this time? Specifically, what do the covariates in the new model represent, and how is this different from what we saw in Part A? (Hint: how do the estimates for `factor_carb` change as the number of carburetors increases?)
- **Part D:** Based on your assessment in Part C, which of these two models do you think is more appropriate for predicting miles per gallon? In other words, does the number of carburetors appear to make more sense as a continuous variable or a categorical one?

Part A: Based on output below, we do see evidence of relationship between carb and mpg

```
lm(mpg ~ wt + carb, mtcars) %>% summary()
```

```
##
## Call:
## lm(formula = mpg ~ wt + carb, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.521 -2.122 -0.047  1.455  5.974
##
## Coefficients:
##             Estimate Std. Error t value             Pr(>|t|)
## (Intercept)   37.730      1.760   21.43 < 0.0000000000000002 ***
## wt            -4.765      0.576   -8.27         0.0000000041 ***
## carb          -0.822      0.349   -2.35                0.026 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.84 on 29 degrees of freedom
## Multiple R-squared:  0.792,  Adjusted R-squared:  0.778
## F-statistic: 55.4 on 2 and 29 DF,  p-value: 0.000000000125
```

Part B and C:

What we see here is a set of coefficient estimates that, ultimately, are testing to see if there is a difference between any individual group of vehicles with a specified number of carburetors against the group of vehicles that only have one. In other words, we have lost the sense of ordinality that comes with treating carburetors as a number that increases or decreases and have replaced them with disjoint groups. Interestingly, we do still see the estiamted size of the difference increase along with the number of carbs in each group

```
lm(mpg ~ wt + factor(carb), mtcars) %>% summary()
```

```
##
## Call:
## lm(formula = mpg ~ wt + factor(carb), data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -4.56  -1.79   0.00   1.41   5.72
##
## Coefficients:
##               Estimate Std. Error t value           Pr(>|t|)
## (Intercept)     36.834      2.100   17.54 0.0000000000000015 ***
## wt              -4.615      0.706   -6.54 0.0000007566610666 ***
```

```
## factor(carb)2   -1.222      1.521   -0.80               0.43
## factor(carb)3   -2.721      2.309   -1.18               0.25
## factor(carb)4   -3.058      1.797   -1.70               0.10
## factor(carb)6   -4.351      3.255   -1.34               0.19
## factor(carb)8   -5.359      3.337   -1.61               0.12
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.04 on 25 degrees of freedom
## Multiple R-squared:  0.795,  Adjusted R-squared:  0.746
## F-statistic: 16.2 on 6 and 25 DF,  p-value: 0.00000016
```

Part D:

Arguments can be made for each. In particular, one might note that the number of carburetors really does constitute a class of vehicle. Further, each vehicle can only take integer values for number of carburetors. Alternatively, and the argument I lean towards, leaving it as a numeric values allows me to estimate the rate (on average) for which miles per gallon decreases as the number of carburetors increases.