Final

Natalie Schmer

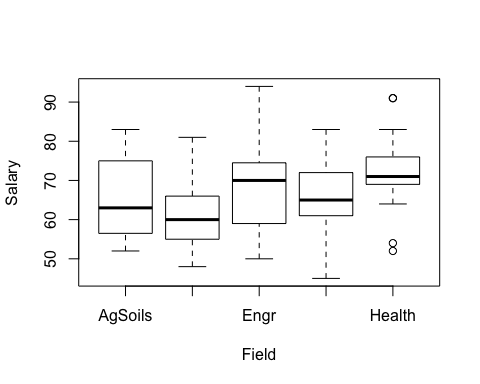
Signature for honor pledge: Natalie Schmer

# 1. Survey of Graduates

library(tidyverse)  
graduates <- read.csv("/Users/natalieschmer/Desktop/GitHub/stats\_511/data/career.csv")

### 1A. Summary of plot of Salar by Field

boxplot(Salary ~ Field, data= graduates)



*From this plot, I expect that health, out of all of the fields, health would be the one to be significanly different than the other fields. But, the boxplots all overlap a fair amount so I would not be surprised if there is not a significant difference.*

### 1B. Anova p-value

lm\_salaries <- lm(Salary ~ Field, data= graduates)  
anova(lm\_salaries)

## Analysis of Variance Table  
##   
## Response: Salary  
## Df Sum Sq Mean Sq F value Pr(>F)  
## Field 4 788.5 197.14 1.567 0.1943  
## Residuals 62 7799.9 125.81

*Since p > 0.05 at 0.19, we can conclude that variance among salaries are not significantly different between the different fields.*

### 1C. Contrast salary means for two particular fields.

*The two fields that seem to have a difference in salaries are health and wildlife ecology.*

em\_1 <- emmeans::emmeans(lm\_salaries, "Field")  
emmeans::contrast(em\_1, list(  
 a = c(0, 1, 0, 0, -1)  
))

## contrast estimate SE df t.ratio p.value  
## a -10.7 4.72 62 -2.268 0.0268

*p < 0.05 at 0.027, and so it is evident that the salaries between health and wildlife ecology are significantly different.*

### 1D. test for salaries by gender

t.test(Salary ~ Gender, data = graduates)

##   
## Welch Two Sample t-test  
##   
## data: Salary by Gender  
## t = 0.57701, df = 47.204, p-value = 0.5667  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -4.281558 7.726003  
## sample estimates:  
## mean in group F mean in group M   
## 68.22222 66.50000

*p > 0.05, so we can conclude there is not a significant difference in average salary by gender.*

### 1E Degree by Gender

graduates %>%   
 select(Gender, Degree) %>%   
 group\_by(Gender, Degree) %>%   
 summarise(n())

## # A tibble: 4 x 3  
## # Groups: Gender [2]  
## Gender Degree `n()`  
## <fct> <fct> <int>  
## 1 F MS 7  
## 2 F PhD 20  
## 3 M MS 30  
## 4 M PhD 10

gender\_degree <-matrix(c(7, 30, 20, 10), byrow = TRUE, nrow = 2)  
colnames(gender\_degree) <-c("Female", "Male")  
rownames(gender\_degree) <-c("MS", "PhD")  
  
gender\_degree

## Female Male  
## MS 7 30  
## PhD 20 10

chisq.test(gender\_degree)

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: gender\_degree  
## X-squared = 13.777, df = 1, p-value = 0.0002058

*p is <<< 0.05 at 0.0002, so we can conclude that there is a significant relationship between gender and degree level*

### 1F. Field by Gender

gender\_field <- graduates %>%   
 select(Gender, Field) %>%   
 group\_by(Gender, Field) %>%   
 summarise(n()) %>%   
 pivot\_wider(names\_from = "Gender",  
 values\_from = "n()") %>%   
 as.matrix()  
  
gender\_field <-matrix(c(4, 3, 3, 7, 6, 13, 5, 13, 9, 4), byrow = TRUE, nrow = 5)  
colnames(gender\_field) <-c("Female", "Male")  
rownames(gender\_field) <-c("AgSoil", "Ecol", "Engr", "Envi", "Health")  
  
gender\_field

## Female Male  
## AgSoil 4 3  
## Ecol 3 7  
## Engr 6 13  
## Envi 5 13  
## Health 9 4

chisq.test(gender\_field)

## Warning in chisq.test(gender\_field): Chi-squared approximation may be incorrect

##   
## Pearson's Chi-squared test  
##   
## data: gender\_field  
## X-squared = 7.5628, df = 4, p-value = 0.109

chisq.test(gender\_field)$expected

## Warning in chisq.test(gender\_field): Chi-squared approximation may be incorrect

## Female Male  
## AgSoil 2.820896 4.179104  
## Ecol 4.029851 5.970149  
## Engr 7.656716 11.343284  
## Envi 7.253731 10.746269  
## Health 5.238806 7.761194

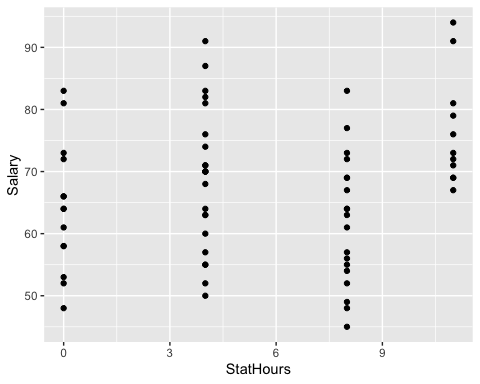
fisher.test(gender\_field)

##   
## Fisher's Exact Test for Count Data  
##   
## data: gender\_field  
## p-value = 0.1177  
## alternative hypothesis: two.sided

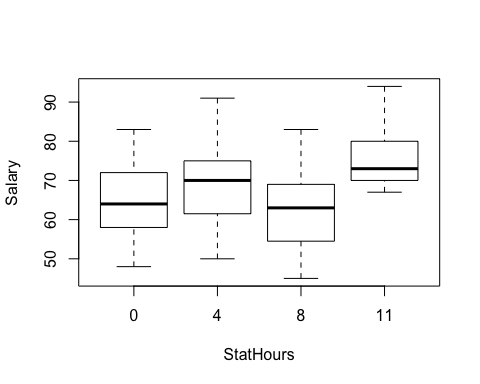
*Since the p > 0.05, we conclude that there is not a significant relationship between gender and field.*

### 1G. Statistics coursework with salary…

#Visualize   
ggplot(data = graduates, aes(x= StatHours, y = Salary))+  
 geom\_point()



boxplot(Salary ~ StatHours, data = graduates)



#anova because categorical  
graduates <- graduates %>%   
 mutate(StatHours = as\_factor(StatHours),  
 StatHours = fct\_relevel(StatHours, "0"))  
  
lm\_stat\_salaries <- lm(Salary ~ StatHours, data = graduates)  
summary(lm\_stat\_salaries)

##   
## Call:  
## lm(formula = Salary ~ StatHours, data = graduates)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -18.8261 -7.2727 -0.2143 7.0000 22.1739   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 64.214 2.803 22.911 < 2e-16 \*\*\*  
## StatHours4 4.612 3.555 1.297 0.19925   
## StatHours8 -2.214 3.694 -0.599 0.55101   
## StatHours11 12.331 4.225 2.918 0.00487 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 10.49 on 63 degrees of freedom  
## Multiple R-squared: 0.1933, Adjusted R-squared: 0.1549   
## F-statistic: 5.032 on 3 and 63 DF, p-value: 0.003444

anova(lm\_stat\_salaries)

## Analysis of Variance Table  
##   
## Response: Salary  
## Df Sum Sq Mean Sq F value Pr(>F)   
## StatHours 3 1660.1 553.36 5.0317 0.003444 \*\*  
## Residuals 63 6928.4 109.97   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

*I chose to do an anova for this question because upon visualizing the data, the predictor of stat hours seem to be a categoriacal variable rather than continuous, and an anova is more appropriate for categorical predictor data with continuous response, and salary is a continuous response.*

### 1H. What’s up with this gender as only predictor for salary from question 1D?

*Question D did not take into account the degree, which is important to consider because degree is a major infulence on salary. Additionally, each field pays differently which was not accounted for, and this also matters with degree because the salary for a given dgree in a given field may not be the same for that same degree in another field. This question could have had some sort of interaction or been a multiple regression for predicting salary while taking all variables into account.*

# 2. Jaundice

### 2A comparing proportions before and after.

{  
#2011  
totalbirths\_2011 <- 1098  
jaundice\_2011 <- 192  
  
#2013  
totalbirths\_2013 <- 1303  
jaundice\_2013 <- 88  
}  
  
#2011 proportion   
jaundice\_2011/totalbirths\_2011

## [1] 0.1748634

#2013 proportion   
jaundice\_2013/totalbirths\_2013

## [1] 0.06753645

#2: Test  
prop.test(c(192, 88), c(1098, 1303), correct = T)

##   
## 2-sample test for equality of proportions with continuity correction  
##   
## data: c(192, 88) out of c(1098, 1303)  
## X-squared = 65.59, df = 1, p-value = 5.551e-16  
## alternative hypothesis: two.sided  
## 95 percent confidence interval:  
## 0.08021121 0.13444265  
## sample estimates:  
## prop 1 prop 2   
## 0.17486339 0.06753645

*SInce p < 0.05, there is a significant difference in the proportions of cases with jaundice between 2011 and 2013.*

### 2B. 2011 Data only

births2011 <- matrix(c(101, 228,  
 50, 455,  
 41, 223),  
 byrow = T, nrow = 3)  
colnames(births2011) <- c("Jaundice", "Not Jaundice")  
rownames(births2011) <- c("Exclusive Breast Milk", "Mixed Feeding", "Exclusive Formula")  
  
births2011

## Jaundice Not Jaundice  
## Exclusive Breast Milk 101 228  
## Mixed Feeding 50 455  
## Exclusive Formula 41 223

chisq.test(births2011)

##   
## Pearson's Chi-squared test  
##   
## data: births2011  
## X-squared = 60.645, df = 2, p-value = 6.778e-14

chisq.test(births2011)$expected

## Jaundice Not Jaundice  
## Exclusive Breast Milk 57.53005 271.4699  
## Mixed Feeding 88.30601 416.6940  
## Exclusive Formula 46.16393 217.8361

####1. *Since p <<< 0.05, there is a significant association between feeding type and jaundice before the program was out in place.*

####2. *Based from the expected vs actual counts, it appears that the exclusive breastmilk feeding deviates the most from all the types of feeding.*

### 2C. Both before and after BF.

before\_after <- matrix(c(150, 1104,  
 69, 618,  
 61, 399),  
 byrow = T, nrow = 3)  
colnames(before\_after) <- c("Jaundice", "Not Jaundice")  
rownames(before\_after) <- c("Exclusive Breast Milk", "Mixed Feeding", "Exclusive Formula")  
  
before\_after

## Jaundice Not Jaundice  
## Exclusive Breast Milk 150 1104  
## Mixed Feeding 69 618  
## Exclusive Formula 61 399

#Odds ratio   
{  
ebm <- 150/1104  
mf <- 69/618  
ef <- 61/399  
}  
  
epitools::oddsratio(before\_after, method = "wald")$p.value

## NA  
## two-sided midp.exact fisher.exact chi.square  
## Exclusive Breast Milk NA NA NA  
## Mixed Feeding 0.2014258 0.2299437 0.2015282  
## Exclusive Formula 0.4669866 0.4567056 0.4682137

*Since p-values > 0.05, and 1 is included in the confidence intervals, we conclude that there is not a relationship between the bf program and jaundice.*

####2. *The results may not be consistant between the A, B, and C with the proportion test, chi-square test, and odds ratio becaue of things like table size and the number of comparisons, in that proportions can only compare 2 proportions but chi square tests and odds ratios can have more than 2 groups. Additionally, differences in sample sizes could play a role. This was seen with the Bird example, where sample sizes made it so the chi square test was not appropriate and gave a different p-value. For this example, two other variables that might help in this explaination include how long the mothers and children were allowed to spend together per day since 24 hours per day is ideal, and if pacifiers were used, since it is advised that they are not used.*

# 3. Bike data

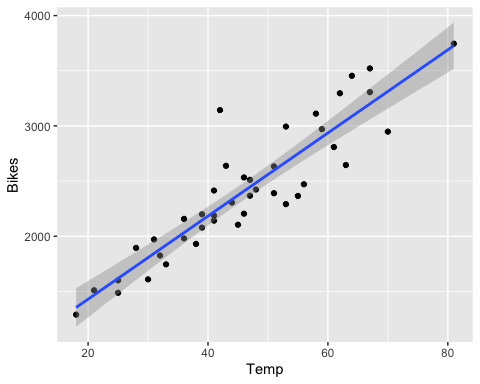
bikedata <- read.csv("/Users/natalieschmer/Desktop/GitHub/stats\_511/data/bikes.csv")  
head(bikedata)

## Bikes Temp Precip Time  
## 1 2204 46 no morning  
## 2 1971 31 no afternoon  
## 3 3521 67 no afternoon  
## 4 2188 41 yes morning  
## 5 1600 25 yes afternoon  
## 6 2948 70 no afternoon

### 3A. plot orginal data

ggplot(bikedata, aes(x = Temp, y = Bikes))+  
 geom\_point()+  
 geom\_smooth(method = "lm")

## `geom\_smooth()` using formula 'y ~ x'



### 3B. Comments on diagnostics for original data

*Concerns: There are many points on the residuals vs fitted plot that fall far from 0, and there are some points that are not falling along the 1:1 line (mostly on the extremes, bottom left and upper right) on the Q-Q plot, so this data may not be normal.*

### 3C. Log Transform diagnostics comments (compared to orginal)

*Log-transforming the Bikes variable did not help the model in this case. The residuals were even farther from 0 and points were more so off of the 1:1 Q-Q plot line.*

### 3D. Large residual values (raw, standardized, and rstudent values)

#ID outlier, row 31  
#temp 42, 3143 bikes   
  
#outlier test, Rstudent residual  
car::outlierTest(lm\_3a)

## rstudent unadjusted p-value Bonferroni p  
## 31 4.244234 0.00013104 0.0055038

rstudent(lm\_3a)[31]

## 31   
## 4.244234

#Rstudentized residual  
rstandard(lm\_3a)[31]

## 31   
## 3.555007

#Raw residual  
resid(lm\_3a)[31]

## 31   
## 885.1651

lm\_3a$residuals[31]

## 31   
## 885.1651

*The outlier here is row 31, with a temperature of 42 f, and 3143 bikes*

### 3E Temp and Bikes relationship observations.

summary(lm\_3a)

##   
## Call:  
## lm(formula = Bikes ~ Temp, data = bikedata)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -403.74 -171.76 -30.17 124.98 885.17   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 676.02 132.86 5.088 8.94e-06 \*\*\*  
## Temp 37.66 2.76 13.645 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 252.3 on 40 degrees of freedom  
## Multiple R-squared: 0.8232, Adjusted R-squared: 0.8187   
## F-statistic: 186.2 on 1 and 40 DF, p-value: < 2.2e-16

*Based on the model, it does appear that there is a clear and significant positive correlation between temperature and number of bikes on campus, in that more students bike to campus when it is warmer outside. The relationship is pretty strong, signified an R-squared of 0.82.*