Assignment 10

Natalie Schmer

1. Vaccine for tuberculosis

```
TB <-array(c( 619, 2537, 10, 8,
            87892, 87886, 499, 505,
            7232, 7470, 45, 29),
        dim=c(2,2,3),
        dimnames=list( Trt=c("Ctrl","Trt"),
                    Response=c("TBneg","TBpos"),
                    Study=c("1","2","3")))

#A Odds Ratios by Study

#by hand (will be shown in solutions)

#Study 1
{
  ctrl_1 <- (10/629)/(1-(10/629))
  trt_1 <- (8/2545)/(1-(8/2545))
}

ctrl_1/trt_1

## [1] 5.123183

#Study 2
{
  ctrl_2 <- (499/88391)/(1-(499/88391))
  trt_2 <- (505/88931)/(1-(505/88931))
}

ctrl_2/trt_2

## [1] 0.9941223

#Study 3
{
  ctrl_3 <- (45/7277)/(1-(45/7277))
  trt_3 <- (29/7499)/(1-(29/7499))
}

ctrl_3/trt_3

## [1] 1.60279
```

```
#B Breslow Day Test
library(metafor)
tb_bd <- metafor::rma.mh(ai = TB[1,1,],
                         bi = TB[1,2,],
                         ci = TB[2,1,],
                         di = TB[2,2,])

tb_bd$BD

## [1] 17.66826

tb_bd$BDp

## [1] 0.0001456754
```

The p-value from this test is p = 0.0001456754, so much less than 0.05, and we colnclude that the odds ratios are not equal across the three studies and that the information across studies should not be combined.

2. Mock Final Exam Question with Lab grades/attendence

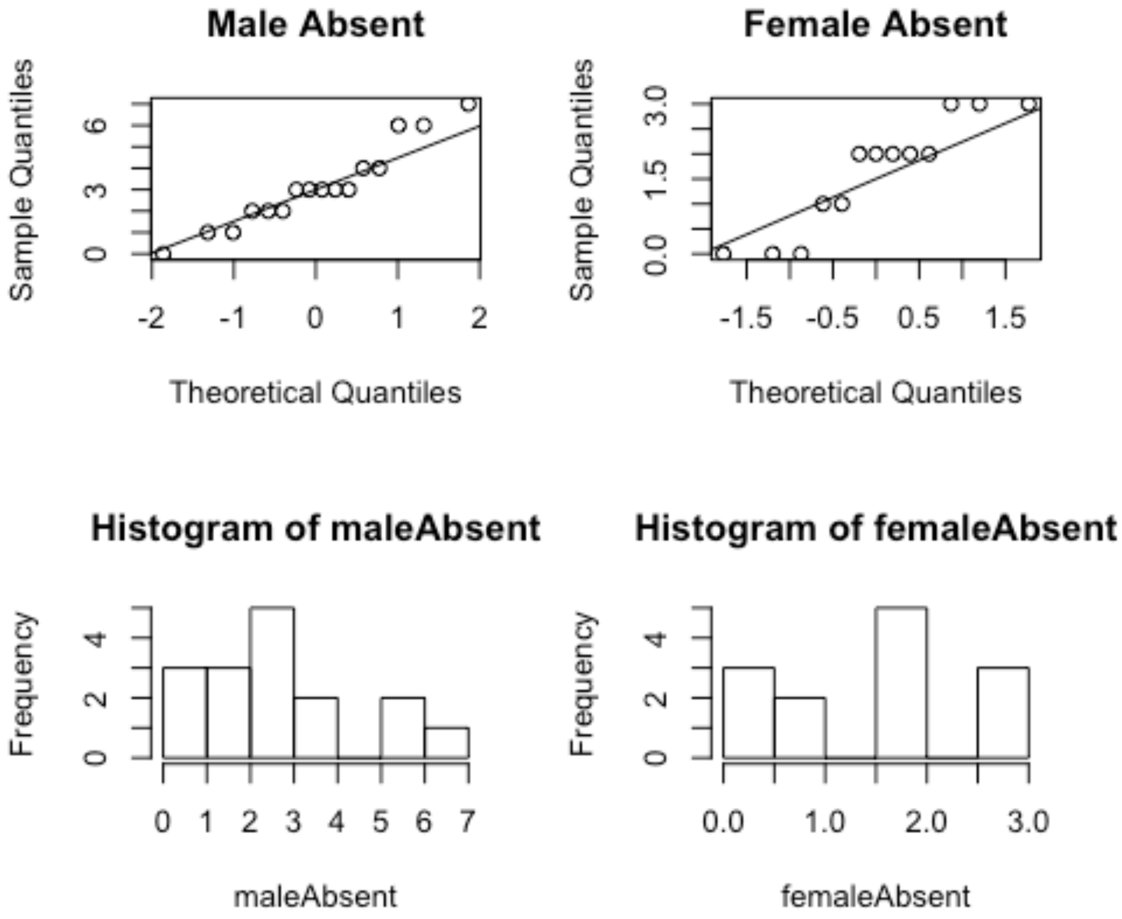
```
LabGrades <-
read.csv("/Users/natalieschmer/Desktop/GitHub/stats_511/data/LabGrades.csv"
, header = TRUE)

str(LabGrades)

## 'data.frame':    29 obs. of  6 variables:
##  $ Student   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Major     : Factor w/ 2 levels "Engineering",..: 2 2 2 2 2 2 2 2 2 2
...
##  $ DaysAbsent: int  2 3 4 7 6 3 0 1 2 0 ...
##  $ Test1     : int  92 83 67 89 91 90 75 89 84 93 ...
##  $ Test2     : int  92 85 70 91 88 91 79 90 85 97 ...
##  $ Gender    : Factor w/ 2 levels "F","M": 1 1 2 2 2 1 1 1 2 1 ...
```

2A. Days Absent by Gender

```
maleAbsent <- subset(LabGrades$DaysAbsent, LabGrades$Gender == "M")

femaleAbsent <- subset(LabGrades$DaysAbsent, LabGrades$Gender == "F")
```

** 2Ai. Diagnostics **

```
par(mfrow = c(2,2))
qqnorm(maleAbsent, main = "Male Absent")
qqline(maleAbsent)
qqnorm(femaleAbsent, main = "Female Absent")
qqline(femaleAbsent)
hist(maleAbsent)
hist(femaleAbsent)
```



```
shapiro.test(maleAbsent)

##
##  Shapiro-Wilk normality test
##
## data:  maleAbsent
## W = 0.93548, p-value = 0.2973

shapiro.test(femaleAbsent)

##
##  Shapiro-Wilk normality test
##
## data:  femaleAbsent
## W = 0.86385, p-value = 0.04332
```

Based on the qqplots and histograms, the data appears somewhat normal– the data roughly falls along the qqline, but the histograms look to be slighly non-normal, specifically on the outer ranges. The shapiro wilk test confirms that the male absent data is normal, but female absent is not normal, although it is very close.

** 2Aii 2-Sample t-test **

```
##
##  Welch Two Sample t-test
##
## data:  maleAbsent and femaleAbsent
## t = 2.6321, df = 24.728, p-value = 0.0144
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.3277449 2.6914859
## sample estimates:
## mean of x mean of y
##  3.125000  1.615385
```

Hypotheses: H0: $\mu_M = \mu_F$ vs HA: μ_M =/= μ_F

The p-value is p =0.0144, indicating that mean absences between male and female student are not equal to eachother. So, we reject the null hypothesis and conclude that one of the groups of students have more absent days than the other, which appears to be males.

2B. Differences in Exam1 and Exam2

```
LabGrades$TestDiff <- LabGrades$Test2 - LabGrades$Test1
# Bi.
shapiro.test(LabGrades$TestDiff)

##
##  Shapiro-Wilk normality test
##
## data:  LabGrades$TestDiff
## W = 0.95638, p-value = 0.267

#p = 0.267, data is normal

# Bii.
t.test(LabGrades$TestDiff, alternative = "greater")

##
##  One Sample t-test
##
## data:  LabGrades$TestDiff
## t = 3.2781, df = 28, p-value = 0.001396
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  0.6801281       Inf
## sample estimates:
## mean of x
##  1.413793

#p-value = 0.001396
```

p < 0.05, indicting that the the average difference for Exam 2 vs Exam 1 is more than 0.

2Ci Summary Table

```
LabGrades %>%
                filter(Gender == "F" & Major == "Engineering") %>%
          count()

## # A tibble: 1 x 1
##       n
##   <int>
## 1     4

{
  table_c2 <- matrix(c(12, 4, 4, 9), nrow = 2, byrow = T)
  colnames(table_c2) <- c("Engineering", "Not Engineering")
  rownames(table_c2) <- c("Male", "Female")
}

table_c2

##        Engineering Not Engineering
## Male            12               4
## Female           4               9
```

2Cii

```
chi_table_c2 <- chisq.test(table_c2)

chi_table_c2$expected

##        Engineering Not Engineering
## Male      8.827586        7.172414
## Female    7.172414        5.827586
```

2ciii

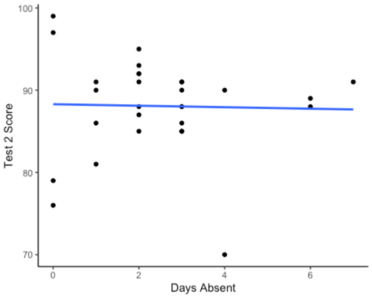
```
chi_table_c2$p.value

## [1] 0.04480381

#p = 0.04480381
```

Since p < 0.05, we can conclude there is an association between gender and being an engineering major.

2D.

```
ggplot(data= LabGrades, aes(x = DaysAbsent, y =Test2 ))+
geom_point()+
  geom_smooth(method = "lm", se = F) +
  labs(x = "Days Absent",
       y = "Test 2 Score") +
  theme_classic()
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Based on the linear regression line, there is not a significant linear relationship between attendance and exam score, becuase even with fewer days absent, there seems to be a wider range of test scores and does not establish a clear relationship.