STAT 426 Assignment 3

Due Tuesday, September 14, 11:59 pm.

Submit through Moodle.

Name: Brianna Diaz

Netid: bdiaz22

Submit your work both as an R markdown (*.Rmd) document and as a pdf, along with any files needed to run the code. Embed your answers to each problem in the document below after the question statement. If you have hand-written work, please scan or take pictures of it and include in a pdf file, ideally combined with your pdf output file from R Markdown. Be sure to show your work.

Problem 1 (6 pts)

The relationship between coffee drinking and myocardial infarction (MI) was investigated in a retrospective study of women aged 30-49. The study included cases hospitalized for the occurrence of MI, and controls hospitalized for some other acute condition. Data for coffee consumption versus MI/control status, stratified by smoking status, are given below:

	Nonsmokers	
Coffee consumption	MI	Controls
≥ 5 cups/day	14	49
< 5 cups/day	75	381

	Smokers	
Coffee consumption	MI	Controls
≥ 5 cups/day	138	134
< 5 cups/day	260	416

(a) Compute the estimated conditional odds ratios for nonsmokers and for smokers.

Answer:

```r
coffee <- data.frame(Drink=c("5 cups or more", "5 cups or more","less then 5 cups", "les
Status=c("Nonsmoking", "Smoking", "Nonsmoking", "Smoking"),
MI=c(14,138,75,260),
Controls=c(49,134,381,416)
)
#coffee

c.tab <- xtabs(MI ~ Drink + Status, data=coffee)
#c.tab

#Odds Ratio
(OR.est <- c.tab[1,1]*c.tab[2,2]/(c.tab[1,2]*c.tab[2,1]))
```

```
## [1] 0.3516908
```

(b) Compute the estimated marginal odds ratio for the 2×2 table obtained by summing over smoking status.

Answer:

```r
more.coffee <- 14+138
less.coffee <- 75+260
controlm <- 49+134
controll <- 381+416

odds.ratio <- more.coffee*controll/(less.coffee*controlm)
odds.ratio
```

```
## [1] 1.976087
```

(c) Based on the results in (a) and (b), comment on whether stratifying on smoking status was important in this study.

Answer: Yes, stratifying on smoking status was important in this study.

Problem 2 (8 pts)

This problem refers to the data in Problem 1.

(a) Let $\theta_{CM(1)}$ denote the (conditional) odds ratio for coffee consumption (C) versus MI status (M) among nonsmokers only. Compute a 95% confidence interval for $\theta_{CM(1)}$.

Answer:

```r
(OR.est.non <- 14*381/(49*75))
```

```
## [1] 1.451429
```

```r
logOR.se = sqrt(sum(1/14,1/49,1/75,1/381))
logOR.CI <- log(OR.est.non) + c(-1,1) * qnorm(1-0.05/2) * logOR.se
exp(logOR.CI)
```

```
## [1] 0.7626552 2.7622507
```

(b) Interpret the result in (a), explaining whether or not there is a significant association between coffee consumption and MI for nonsmokers, at a significance level $\alpha = 0.05$.

Answer: Based off the results there is an association at the significance level at 0.05.

(c) Let $\theta_{CM(2)}$ denote the (conditional) odds ratio for coffee consumption versus MI status among smokers only. Compute a 95% confidence interval for $\theta_{CM(2)}$.

Answer:

```r
(OR.est.smoke <- 138*416/(134*260))
```

```
## [1] 1.647761
```

```r
logOR.se.smoke = sqrt(sum(1/138,1/416,1/134,1/260))
logOR.CI.smoke <- log(OR.est.smoke) + c(-1,1) * qnorm(1-0.05/2) * logOR.se.smoke
exp(logOR.CI.smoke)
```

```
## [1] 1.240691 2.188390
```

(d) Interpret the result in (c), explaining whether or not there is a significant association between coffee consumption and MI for smokers, at a significance level $\alpha = 0.05$.

Answer: Based off the results there is no association at the significance.

Problem 3 (6 pts)

This problem again refers to the data in Problem 1 and uses the notation of Problem 2.

(a) Using only the **nonsmoker** data, compute Pearson's X^2, the likelihood ratio G^2, and their p-values for testing $H_{01} : \theta_{CM(1)} = 1$ versus $H_{A1} : \theta_{CM(1)} \neq 1$. In other words test for independence between coffee consumption and MI status conditional on being in the nonsmoker group. What do you conclude? Is there significant association at level .05?

Answer:

```r
nonsmoke <- data.frame(nonsmoke=c(14,75), Ctrln=c(49,381))
cn.X2 <- chisq.test(nonsmoke, correct = FALSE)
cn.X2
```

```
##
##  Pearson's Chi-squared test
##
## data:  nonsmoke
## X-squared = 1.2993, df = 1, p-value = 0.2543
```

```r
cn.G2 <- 2*sum(cn.X2$observed * log(cn.X2$observed/cn.X2$expected))
data.frame(G2=cn.G2, df=1, pvalue=1-pchisq(cn.G2,1))
```

```
##         G2 df    pvalue
## 1 1.223595  1 0.2686563
```

```r
# There is no significant association at the level 0.05.
```

(b) Using only the **smoker** data, compute Pearson's X^2, the likelihood ratio G^2, and their p-values for testing $H_{02} : \theta_{CM(2)} = 1$ versus $H_{A2} : \theta_{CM(2)} \neq 1$. In other words test for independence between coffee consumption and MI status conditional on being in the smoker group. What do you conclude? Is there significant association at level .05?

Answer:

```r
smoke <- data.frame(smoke=c(138,260), Ctrls=c(134,416))
cs.X2 <- chisq.test(smoke, correct = FALSE)
cs.X2
```

```
##
##  Pearson's Chi-squared test
##
## data:  smoke
## X-squared = 11.996, df = 1, p-value = 0.0005332
```

```r
cs.G2 <- 2*sum(cs.X2$observed * log(cs.X2$observed/cs.X2$expected))
data.frame(G2=cs.G2, df=1, pvalue=1-pchisq(cs.G2,1))
```

```
##         G2 df      pvalue
## 1 11.90854  1 0.000558772
```

```
# There is a significant association at the 0.05.
```

(c) Finally, let's test the null hypothesis of conditional independence of coffee consumption and MI status given smoking status based on all of the data. This is a test of

$$H_0 : \theta_{CM(1)} = 1 \text{ and } \theta_{CM(2)} = 1$$

versus

$$H_A : \theta_{CM(1)} \neq 1 \text{ or } \theta_{CM(2)} \neq 1.$$

Let $G^2_{tot} = G^2_1 + G^2_2$ where G^2_1 is the likelihood ratio G^2 statistic from (a) and G^2_2 is the likelihood ratio G^2 statistic from (b). What is the asymptotic distribution of G^2_{tot} under H_0 (including degrees of freedom)? Compute G^2_{tot} and its p-value.

Answer:

G^2_{tot} is 13.132113 with a p-value of 0.001407321 with a degrees of freedom of 2.

```
g2.sum <- cs.G2 + cn.G2
g2.sum
```

```
## [1] 13.13213
```

```
no = (nrow(nonsmoke)-1)*(ncol(nonsmoke)-1)
yes = (nrow(smoke)-1)*(ncol(smoke)-1)

df = no + yes
df
```

```
## [1] 2
```

```
1 - pchisq(g2.sum, df = 2)
```

```
## [1] 0.001407321
```