Week 02 Questions

Scott Graham 10131323

September 22, 2017

2.1

a.

$$\mathrm{P}(-|C) = \frac{1}{4}, \mathrm{P}(+|\bar{C}) = \frac{2}{3}$$

b.

$$\mathrm{P}(+|C) = 1 - \mathrm{P}(-|C) = 1 - \frac{1}{4} = \frac{3}{4}$$

c.

```
probs <- matrix(c(0.01*3/4, 0.99*2/3, 0.01*1/4, 0.99*1/3), nrow = 2, ncol = 2)
rownames(probs) <- c("C", "CBar")
colnames(probs) <- c("+", "-")
probs
```

```
##           +      -
## C    0.0075 0.0025
## CBar 0.6600 0.3300
```

```
sum_rows <- c()
for (i in 1:2){
  sum_rows[i] <- sum(probs[i, ])
}
sum_cols <- c()
for (j in 1:2){
  sum_cols[j] <- sum(probs[, j])
}
sum_rows
```

```
## [1] 0.01 0.99
```

```
sum_cols
```

```
## [1] 0.6675 0.3325
```

d.

From the "sum_cols" variable:

$$\mathrm{P}(+) = 0.6675, \mathrm{P}(-) = 0.3325$$

$$\text{Test Result} \sim \text{Bernoulli}(p = 0.6675)$$

e.

$$\mathrm{P}(C|+) = \frac{\mathrm{P}(+|C)\,\mathrm{P}(C)}{\mathrm{P}(+)} = \frac{\frac{3}{4}\times 0.1}{0.6675} =$$

```
3/4*sum_rows[1]/sum_cols[1]
```

```
## [1] 0.01123596
```

2.5

a.

Since we are dealing with likelihood, the 1.7 is the relative risk.

b.

Let C be having cancer, and D taking the drug:

$$\mathrm{P}(C|D) = 0.55\,\mathrm{P}(C|D^c) \implies RR = \frac{0.55}{1} = 0.55$$

Similarily

$$RR = \frac{1}{0.55} = 1.\overline{81}$$

2.7

a.

With an odds ratio, we aren't directly measuring the probability of an event, merely its odds. The correct iterpretation would be the odds of a female surviving is 11.4 times that of a male.

b.

$$\mathrm{P}(S|F) = 2.9\,\mathrm{P}(D|F) \implies 1 = 2.9\,\mathrm{P}(D|F) + \mathrm{P}(D|F) \implies \mathrm{P}(D|F) = \frac{10}{39} \implies \mathrm{P}(S|F) = \frac{29}{39}$$

$$\frac{\mathrm{P}(S|M)}{\mathrm{P}(D|M)} = \frac{2.9}{11.4} = \frac{29}{114} \implies 1 = \frac{29}{114}\mathrm{P}(D|M) + \mathrm{P}(D|M) \implies \mathrm{P}(D|M) = \frac{114}{143} \implies$$

$$\mathrm{P}(S|M) = \frac{29}{143}$$

c.

$$RR = \frac{\mathrm{P}(S|F)}{\mathrm{P}(S|M)} = \frac{\frac{29}{39}}{\frac{29}{143}} = \frac{143}{39}$$

The probability of survival for females was $3.\overline{66}$ times that for males.

2.11

a.

The difference in proportions for lung cancer is:

$$\hat{p}_{LC|S} - \hat{p}_{LC|NS} = 0.00140 - 0.00010 = 0.00130$$

So the probability of dieing from lung cancer increases by 0.0013 per year if one smokes.

The difference in proportions for heart disease is:

$$\hat{p}_{HD|S} - \hat{p}_{HD|NS} = 0.00669 - 0.00413 = 0.00256$$

So the probability of dieing from heart disease increases by 0.00256 per year if one smokes.

$$RR_{LC} = \frac{\hat{p}_{LC|S}}{\hat{p}_{LC|NS}} = \frac{0.00140}{0.00010} = 14$$

So the probability of dieing from lung cancer is 14 times higher per year for smokers vs. non-smokers.

$$RR_{HD} = \frac{\hat{p}_{HD|S}}{\hat{p}_{HD|NS}} = \frac{0.00669}{0.00413} = 1.619855$$

So the probability of dieing from heart disease is 1.619855 times higher per year for smokers vs. non-smokers.

$$OR_{LC} = \frac{\frac{\hat{p}_{LC|S}}{1-\hat{p}_{LC|S}}}{\frac{\hat{p}_{LC|NS}}{1-\hat{p}_{LC|NS}}} = \frac{\frac{0.00140}{1-0.00140}}{\frac{0.00010}{1-0.00010}} = 14.01823$$

So the odds of dieing from lung cancer is 14.01823 times higher per year for smokers vs. non-smokers.

$$OR_{HD} = \frac{\frac{\hat{p}_{HD|S}}{1-\hat{p}_{HD|S}}}{\frac{\hat{p}_{HD|NS}}{1-\hat{p}_{HD|NS}}} = \frac{\frac{0.00669}{1-0.00669}}{\frac{0.00413}{1-0.00413}} = 1.624029$$

So the odds of dieing from heart disease is 1.624029 times higher per year for smokers vs. non-smokers.

b.

Lung cancer is more strongly related to one's smoking habits compared to Heart Disease. While its difference in probability of death is smaller than heart disease, the likelihood and odds are both much greater than their heart disease counterparts.

2.23

Let $\alpha = 0.05$.

$$H_0 : \text{Highest Degree and Religious Beliefs are independent}$$

$$H_1 : \text{Highest Degree and Religious Beliefs are dependent}$$

P-Value:

```r
edu_rel_tbl <- matrix(c(178, 570, 138, 138, 648, 252, 108, 442, 252), nrow = 3, ncol = 3)
colnames(edu_rel_tbl) <- c("Fundamentalist", "Moderate", "Liberal")
rownames(edu_rel_tbl) <- c("< High School", "High School or Junior College", "Bachelor or Graduate")
chisq.test(edu_rel_tbl, correct = FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  edu_rel_tbl
## X-squared = 69.157, df = 4, p-value = 3.42e-14
```

Because our p-value is $< \alpha$, we reject our null hypothesis of independence, and assume some sort of dependency based on our sample.

```r
chisq.test(edu_rel_tbl, correct = FALSE)$stdres
```

```
##                                Fundamentalist   Moderate   Liberal
## < High School                        4.534918 -2.5521511 -1.941705
## High School or Junior College        2.553268  1.2859224 -3.994697
## Bachelor or Graduate                -6.809750  0.7009713  6.252547
```

The large standarized residuals for the Fundamentalist and Liberal categories shows that their may exist some relation between education and those categories, hence their may be dependency.

2.27

a.

Let $\alpha = 0.05$.

$$H_0 : \text{Family Income and Aspirations are independent}$$

$$H_1 : \text{Family Income and Aspirations are dependent}$$

```r
aspirations_tbl <- matrix(c(9, 44, 13, 10, 11, 52, 23, 22, 9, 41, 12, 27), nrow = 4, ncol = 3)
colnames(aspirations_tbl) <- c("L", "M", "H")
rownames(aspirations_tbl) <- c("Some HS", "Graduate HS", "Some College", "Graduate College")
aspirations_tbl
```

```
##                   L  M  H
## Some HS           9 11  9
## Graduate HS      44 52 41
## Some College     13 23 12
## Graduate College 10 22 27
```

```r
chisq.test(aspirations_tbl, correct = FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  aspirations_tbl
## X-squared = 8.8709, df = 6, p-value = 0.181
```

```r
chisq.test(aspirations_tbl, correct = FALSE)$expected
```

```
##                         L        M         H
## Some HS           8.07326 11.47253  9.454212
## Graduate HS      38.13919 54.19780 44.663004
## Some College     13.36264 18.98901 15.648352
## Graduate College 16.42491 23.34066 19.234432
```

At that level, we fail to reject our null hypothesis of independence, based on our sample. However, the family income levels are ordinal instead of purely categorical, our tests may not be accurate.

b.

```r
chisq.test(aspirations_tbl, correct = FALSE)$stdres
```

```
##                            L          M          H
## Some HS           0.4061328 -0.1898118 -0.1903291
## Graduate HS       1.5828205 -0.5440627 -0.9459053
## Some College     -0.1286367  1.3041565 -1.2374420
## Graduate College -2.1078423 -0.4031584  2.4360173
```

For low income families, we see a fair bit less number of students who aspire to graduate from college than what we would expect if they were independent. The opposite is true for high income families, where a fair bit more number of students expect to graduate from college, than we'd expect under the null hypothesis. For medium income families, our standardized residuals are all fairly small, so there may not be dependency. Note that all our standardized residuals are between $[-2.5, 2.5]$, so most likely no strong dependencies exists based on this sample.

c.

```r
library(coin)
```

```
## Loading required package: survival
```

```r
aspirations_tbl <- as.table(aspirations_tbl)
lbl_test(aspirations_tbl)
```

```
##
##  Asymptotic Linear-by-Linear Association Test
##
## data:  Var2 (ordered) by
##   Var1 (Some HS < Graduate HS < Some College < Graduate College)
## Z = 2.1792, p-value = 0.02932
## alternative hypothesis: two.sided
```

```r
pchisq(statistic(lbl_test(aspirations_tbl))^2, 1, lower.tail = FALSE)
```

```
##
## 0.02931658
```

Since our p-value is $< \alpha$, we reject our null hypothesis, and assume some dependency based on the sample.