536HW4

Coco_Luo

2022-11-01

Question 1

	Yes	No	Row Total
Placebo	28	656	684
Aspirin	18	658	676
Column Total	46	1314	1360

Question 2

The frechet bounds is $max\{0, n_{i+} + n_{+j} - n_{++}\} \le n_{ij} \le min\{n_{i+}, n_{+j}\}$ Thus, we have

$$max\{0, 46 + 684 - 1360\} = 0 \le n_{11} \le 46 = min\{46, 684\},$$
$$max\{0, 1314 + 684 - 1360\} = 638 \le n_{12} \le 684 = min\{1314, 684\},$$
$$max\{0, 46 + 676 - 1360\} = 0 \le n_{21} \le 46 = min\{46, 676\},$$
$$max\{0, 1314 + 676 - 1360\} = 630 \le n_{22} \le 676 = min\{1314, 676\}.$$

Question 3

Let T be the set of 2×2 tables with the same row and column totals as the Aspirin data. The number of tables in set T is $46 - 0 + 1 = 684 - 638 + 1 = 46 - 0 + 1 = 676 - 430 + 1 = 47$.

Let x be the (1, 1) cell count, where x is an integer between 0 and 46.
It follows that the (1, 2) cell count should be $(684 - x)$ and the (2, 1) cell count should be $(46 - x)$.
Thus the (2, 2) cell count should be $(x + 630) = (676 - (46 - x)) = (1314 - (684 - x))$.

The analytic expression for T is :

$$T = \{(x, 684 - x, 46 - x, x + 630) : x \in 0, 1, ..., 46\}.$$

which is equivalent to the other three cells.

Question 4

```
## 2 iterations: deviation 0
```

28	656
18	658

The estimated expected cell values under the saturated log-linear model is same as the observed counts in the original table, as it has 0 degrees of freedom.

Question 5

Under the log-linear model of independence, we can use the following formula to manually compute the expected cell values:

$$m_{ij} = \frac{n_{i+}n_{+j}}{n_{++}}$$

I choose to fit the log linear model for easier computation

2 iterations: deviation 0

23.13529	660.8647
22.86471	653.1353

Question 6

Our null hypothesis is $H_0 : X_1 \perp\!\!\!\perp X_2$. Pearson's chi-square statistic is

$$\chi^2 = \sum \frac{(observed - Expected)^2}{expected}$$

In the context of our problem, we have:

$$\chi^2 = \frac{(n_{11} - m_{11})^2}{m_{11}} + \frac{(n_{12} - m_{12})^2}{m_{12}} + \frac{(n_{21} - m_{21})^2}{m_{21}} + \frac{(n_{22} - m_{22})^2}{m_{22}} = 2.13$$

Our test statistic follows an asymptotic χ^2 distirbution degree of freedom= 1, the p value is 0.14 which is greater than 0.05. We make the conclusion that we fail to reject the null hypothesis. The independence model fits the data well.

Question 7

Our null hypothesis is $H_0 : X_1 \perp\!\!\!\perp X_2$.

likelihood ratio statistic G^2 is:

$$G^2 = 2[n_{11}log(\frac{n_{11}}{m_{11}}) + n_{12}log(\frac{n_{12}}{m_{12}}) + n_{21}log(\frac{n_{21}}{m_{21}}) + n_{22}log(\frac{n_{22}}{m_{22}})] = 2.15$$

Our test statistic follows an asymptotic χ^2 distirbution degree of freedom= 1, the p value is 0.14 which is very close to what we had above. We make the conclusion that we fail to reject the null hypothesis. The independence model fits the data well.

Question 8

Our cross product as an odds ratio: $\alpha = \frac{n_{11}/n_{21}}{n_{12}/n_{22}}$.

Null hypothesis: there is **no** association between Aspirin Use and Myocardial Infarction, i.e, $H_0 : \alpha = 1$. Alternative hypothesis: there **is** an association between Aspirin Use and Myocardial Infarction, i.e, $H_A : \alpha \neq 1$.

```
##
##  Fisher's Exact Test for Count Data
##
## data:  as.array
## p-value = 0.1768
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.8236804 3.0256940
```

```
## sample estimates:
## odds ratio
##   1.559785
```

The estimated exact p-value is 0.1768, which is greater than 0.05, hence we fail to reject the null hypothesis.

Question 9

The three test statistics we obtained above suggest that we have no evidence to reject the null, thus I chose the log-linear model of independence. Under this model, the logit of "Aspirin Use" does not depend on "Myocardial Infarction" because the two variables are assumed independent, so we have:

$$log\left(\frac{P(\text{Aspirin Use} = \text{"Placebo"}|\text{Myocardial Infarction})}{P(\text{Aspirin Use} = \text{"Aspirin"}|\text{Myocardial Infarction})}\right)$$

$$= log(P(Aspirin = 1|MI)) - log(P(Aspirin = 2|MI)) = \beta_0 = \mu_1(1) - \mu_1(2)$$

```
## $`(Intercept)`
## [1] 4.811572
##
## $`1`
## [1]  0.005882421 -0.005882421
##
## $`2`
## [1] -1.676095  1.676095
```

The two terms are related to the independence model where $\mu_1(1) = -\mu_1(2) = 0.0058$. And $\mu_1(1) - \mu_1(2) = 0.0117$. As a result, regardless of the value of `Myocardial Infarction`, the logit of the probability that `Aspirin Use = "Placebo"` is roughly 0.01176.

Question 10

The p-values obtained from Pearson's chi-square statistic, likelihood ratio statistics G^2, and Fisher's Exact Test are: 0.1444434, 0.1428159, and 0.1768, respectively, which are all greater than a significance level of 0.05.

Therefore, we can conclude that there is no evidence to reject the null hypothesis which states that there is no association between "Aspirin Use" and "Myocardial Infarction".

Appendix

```r
aspirin = data.frame(cbind(c(28,18), c(656, 658)))
rownames(aspirin) = c("Placebo", "Aspirin")
colnames(aspirin) = c("Yes", "No")
#aspirin
aspirin1 = aspirin
aspirin1[, "Row Total"] = c(sum(aspirin["Placebo",]), sum(aspirin["Aspirin",]))
aspirin1["Column Total", ] = c(sum(aspirin$Yes), sum(aspirin$No), sum(aspirin))
```

```r
as.array = array(c(28,18, 656, 658), c(2, 2))
saturated.loglin = loglin(as.array,margin = list(c(1,2)), fit = T, param = T)
```

```
## 2 iterations: deviation 0
```

```r
indep.loglin = loglin(as.array, margin = list(1,2), fit = T, param = T)
```

```
## 2 iterations: deviation 0
```

```r
#indep.loglin$fit
#manual.fit = array(c(46*684/1360, 46*676/1360, 1314*684/1360, 1314*676/1360),c(2, 2))

p.pearson = 1 - pchisq(indep.loglin$pearson - saturated.loglin$pearson,
                       indep.loglin$df - saturated.loglin$df)

p.lrt = 1 - pchisq(indep.loglin$lrt - saturated.loglin$lrt,
                   indep.loglin$df - saturated.loglin$df)
#p.lrt

#fisher.test(as.array, simulate.p.value = T, B = 1e5)
manual.odds = (as.array[1,1]/as.array[2,1])/(as.array[1,2]/as.array[2,2])
```