Week 3 Problem Based Learning and Practical Solutions

Jeffrey O. Hanson[1]
[1]*School of Biological Sciences, The University of Queensland, Brisbane, QLD, Australia*
Correspondance should be addressed to jeffrey.hanson@uqconnect.edu.au

21 March 2016

Contents

Problem based learning workshop

Exercise 1

In a genetics experiment on tomatoes, a dihybroid cross was made, with the frequencies of the progeny expected to be in the ratio 9:3:3:1.

The following table gives the observed frequencies

Round/Yellow	Wrinkled/Yellow	Round/Green	Wrinked/Green
56	19	17	8

1. *What are the expected proportions?*

- Remember that our expected ratio is 9:3:3:1. We can use this to calculate the expected proportions.

Round/Yellow	Wrinkled/Yellow	Round/Green	Wrinked/Green
$\frac{9}{9+3+3+1} = 0.5625$	$\frac{3}{9+3+3+1} = 0.1875$	$\frac{3}{9+3+3+1} = 0.1875$	$\frac{1}{9+3+3+1} = 0.0625$

2. *What are the expected frequencies?*

- To calculate the expected frequencies, we first need to calculate the number of individuals in the experiment.
- The total number of individuals in the experiment is $56 + 19 + 17 + 8 = 100$.
- We can then multiply the epxected proportions (which we claculated in the previous question) by the total number of observed individuals to yield the expected frequencies.

Round/Yellow	Wrinkled/Yellow	Round/Green	Wrinked/Green
$0.5625 \times 100 = 56.25$	$0.1875 \times 100 = 18.75$	$0.1875 \times 100 = 18.75$	$0.0625 \times 100 = 6.25$

3. *What is the χ^2 value?*

   - We can use the following equation to calculate the χ^2 value: $\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$.
   - In this equation, O_i is the observed frequency for the i'th combination and E_i is the expected frequency for the i'th combination.
   - We can sub in our observed and expected frequencies into the equation, and calculate the χ^2 value.

$$\chi^2 = \frac{(56 - 56.25)^2}{56.25} + \frac{(19 - 18.75)^2}{18.75} + \frac{(17 - 18.75)^2}{18.75} + \frac{(8 - 6.25)^2}{6.25}$$
$$\chi^2 = 0.001 + 0.003 + 0.163 + 0.49$$
$$\chi^2 = 0.657$$

4. *What are the degrees of freedom?*

   - Here our degrees of freedom (df) is 3. This is because it is our sample size - 1.

5. *What is the associated p-value?*

   - Given that we know the χ^2 value is 0.657 and the degrees of freedom is 1, we can calculate the p-value using the following `R` code, where `x` is the χ^2 and `d` is the degrees of freedom (df).

```
pchisq(x, df=d, lower.tail=FALSE)
```

   - We can sub our values into this code, and calculate the p-value.

```
pchisq(0.657, df=1, lower.tail=FALSE)
```

   - $P = 0.4176211$

6. *Can you reject the null hypothesis?*

   No, given that the p-value is much greater than 0.05, we cannot reject the null hypothesis.

Exercise 2

Here we have a contingency table for case and control individuals genotyped at a diallelic marker locus.

Affection status	AA	AB	BB
Case	23	47	30
Control	12	40	48

1. *Calculate the expected values for each cell?*

   - We can calculate the expected values for each cell using the equation below, where R is the number of rows and C is the number of columns. $E_ij = \frac{\sum_{r=1}^{R} O_{r,j} \times \sum_{c=1}^{C} O_{i,C}}{\sum_{r=1}^{R} \sum_{c=1}^{C} O_{r,c}}$
   - First, let's calculuate the row and column totals | Affection status | AA | AB | BB | *Row totals* | |:————————-:|:—:|:—:|:—:|:—:|————-:| | Case | 23 | 47 | 30 | 100 | | Control | 12 | 40 | 48 | 100 | | *Column totals* | 35 | 98 | 67 | *Total* = 200 |
   - Now, lets apply the equation to calculate expected values. | Affection status | AA | AB | BB | *Row totals* | |:————————-:|:—:|:—:|:—:|:—:|————-:| | Case | $\frac{35 \times 100}{200} = 17.5$ | $\frac{98 \times 100}{200} = 49$ | $\frac{67 \times 100}{200} = 33.5$ | 100 | | Control | $\frac{35 \times 100}{200} = 17.5$ | $\frac{98 \times 100}{200} = 49$ | $\frac{67 \times 100}{200} = 33.5$ | 100 | | *Column totals* | 35 | 98 | 67 | *Total* = 200 |

2. *Perform a χ^2 test on those values.*

   - Let's refresh our memory on what our observed frequencies are.

Affection status	AA	AB	BB
Case	23	47	30
Control	12	40	48

   - Also, let's refresh our memory on what our expected frequencies are.

Affection status	AA	AB	BB
Case	17.5	49	67
Control	17.5	49	67

   - We can calculate the χ^2 value using this equation $\chi^2 = \sum_{i=1}^{n} \frac{O_i - E_i}{E_i}$.
   - We can then calculate the χ^2 value by subbing our values into the equation.

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$
$$\chi^2 = \frac{(23 - 17.5)^2}{17.5} + \frac{(47 - 49)^2}{49} + \frac{(30 - 67)^2}{67} + \frac{(12 - 17.5)^2}{17.5} + \frac{(40 - 49)^2}{49} + \frac{(48 - 67)^2}{67}$$
$$\chi^2 = 1.729 + 0.082 + 20.433 + 1.729 + 1.653 + 5.388$$
$$\chi^2 = 31.014$$

3. *Find the associated p-value.*

   - First we need to calculate the degrees of freedom (df). Since we generated our expected values using row and column totals we calculate our degrees of freedom (df) with **one simple trick**.

$$df = (\text{number of rows} - 1) \times (\text{number of columns} - 1)$$
$$df = (2 - 1) \times (3 - 1)$$
$$df = 2$$

   - Now that we have the degrees of freedom ($df = 2$) and our χ^2 statistic (31.014) which we calculated earlier, we can calculate the p-value using the following `R` code.

```
pchisq(31.014, df=2, lower.tail=FALSE)
```

$$P = 1.842449 \times 10^{-7}$$
$$P < 0.001$$

4. Can you reject the null hypothesis at alpha = 0.05?

   - Yes. Yes, we can.

Exercise 3

1. *Are extinction events, as observed in the fossil record, random in time, or do they have cluster (eg. mass extinctions)? In other words, do species go extinct at random intervals, or species tend to go extinct at the same time? To answer this question, use χ^2 to measure the "fit" of the probability model to the data. Significant lack of fit implies rejection of the null hypothesis of "no departure from the model".*

- Here we have a table that shows the number of time intervals between extinction events. The "number of extinctions" column contains the index (starting at zero) of the extinction (eg. the first extinction has a zero, the second has a one, etc). The "number of time intervals" describes the amount of time that has passed between consecutive extinctions (using a standardised unit of time).

Number of extinctions	Number of Time Intervals
0	0
1	13
2	15
3	16
4	7
5	10
6	4
7	2
8	1
9	2
10	1
11	1
12	0
13	0
14	1
15	0
16	2
17	0
18	0
19	0
20	1

- We will use the Poisson distribution to express our null hypothesis. An expectation (or prediction) from the Poisson distribution can be calculated using the equation $Pr\{Y = k|\mu\} = \frac{e^{-\mu}\mu^k}{k!}$ assuming a given value for μ. This expectation is our expected probability. We can then multiply $Pr\{Y = k|\mu\}$ by the total amount of observed extinction intervals (76) to calculate an expected frequency.
- A different μ value will give us a different expectation for the same input. For example, if we wanted to create an expectation for 5 extinctions, we could chose $\mu = 4$ and this would give us an expected probability of $Pr\{Y = 5|\mu = 4\} = \frac{e^{-4}4^5}{5!} = \frac{0.018\times1024}{5\times4\times3\times2\times1} = 0.154$. We

can then use this to calculate the expected frequency $0.154 \times 76 = 11.704$. Alternatively, if we used $\mu = 8$ we would have an expected probability of $Pr\{Y = 5|\mu = 8\} = \frac{e^{-8}8^5}{5!} = 0.091$ and an expected frequency of $0.091 \times 76 = 6.916$. We can see that with an observed value of 10 for 5, using $\mu = 5$ gives a much closer estimate. So if we want to create a good model for our data using this distribution, we need to pick a good value for μ.

- We will use the following equation to give us a decent μ value, where T_i is the i'th extinction interval.

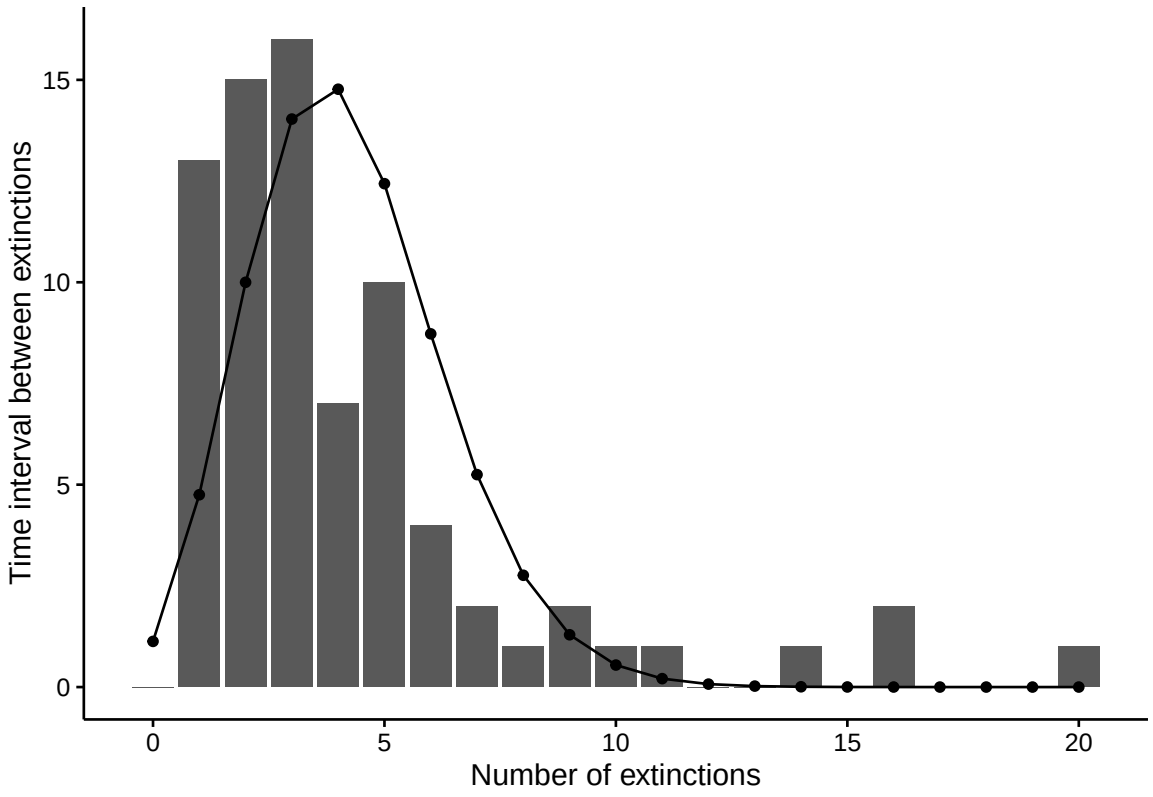
$$\mu = \frac{\sum_{i=1}^{N}(i-1)T_i}{\sum_{i=1}^{N} T_i}$$
$$\mu = \frac{(0 \times 0) + (1 \times 13) + (2 \times 15) + (3 \times 16) + (4 \times 7)...}{0 + 13 + 15 + 16 + 7...}$$
$$\mu = 4.21$$

- Note that if we were feeling particularly clever, we could optimise our μ value to give us the best possible model assuming a Poisson distribution.
- Let's make a quick plot to see how our model compares to the data.

```
# store data in data.frame
ext.DF=data.frame(n.extinctions=0:20,
    n.intervals=c(0,13,15,16,7,10,4,2,1,2,1,1,0,0,1,0,2,0,0,0,1))

# make column with prediction
ext.DF$predicted.probs <- dpois(ext.DF$n.extinctions, 4.21)
ext.DF$predicted.freqs <- ext.DF$predicted.probs * sum(ext.DF[[2]])

# make plot
library(ggplot2)
ggplot(data=ext.DF) +
    geom_bar(aes(x=n.extinctions, y=n.intervals), stat='identity') +
    geom_line(aes(x=n.extinctions, y=predicted.freqs)) +
    geom_point(aes(x=n.extinctions, y=predicted.freqs)) +
    theme_classic() +
    xlab('Number of extinctions') +
    ylab('Time interval between extinctions')
```



- Now that we can see our data. Let's check to see if the Poisson is appropriate here. Some general rules of thumb for the Poisson distribution are:
  - No expected frequencies should be less than 1.
  - No more than 20% of the expected frequencies should be less than 5.
- **The data breaks one of our rules of thumb for the Poisson distribution**. So, we could either transform the data so that it obeys these rules or pick a different distribution (eg. negative binomial). Here, we will group the data together so that it obeys these rules. Note that the system of grouping we apply can affect the p-value we calculate at the end. So if you have to group your own data, you should think about how to sensibly group the data. The table below shows the groupings we will use.

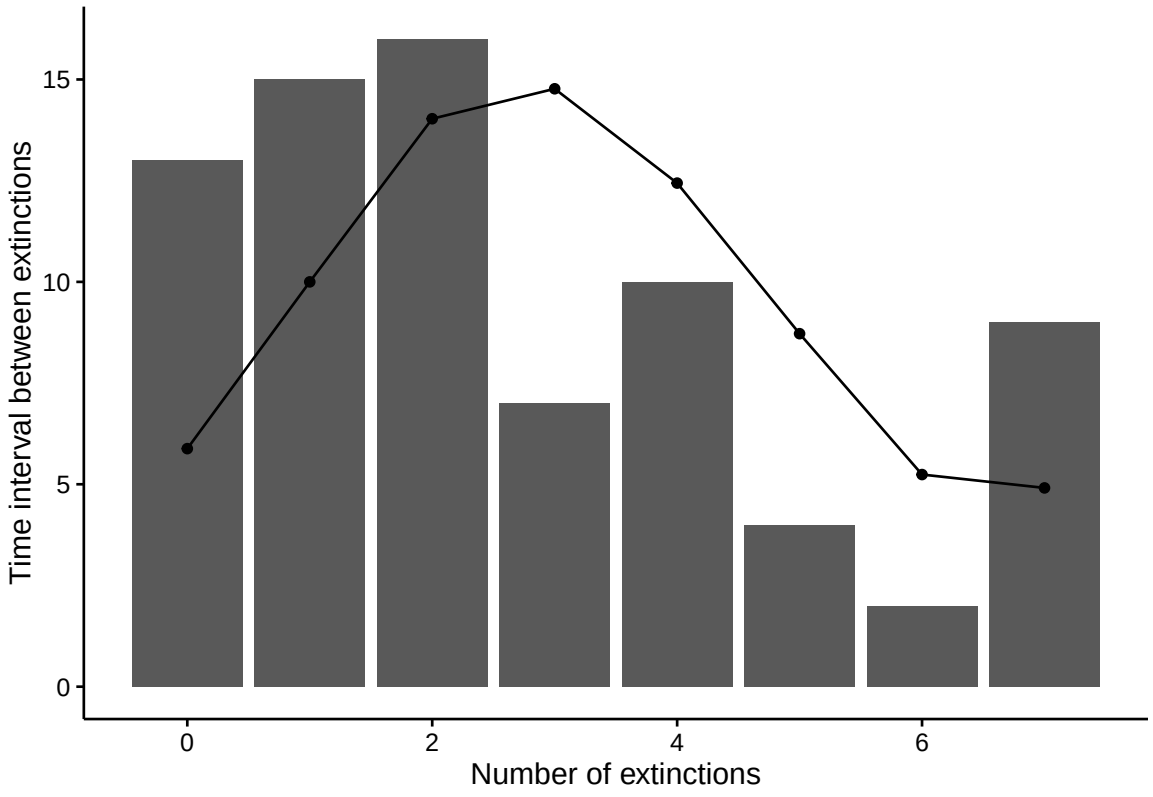
Number of extinctions	Number of Time Intervals	Predicted Time Intervals
0 or 1	13	5.88
2	15	10
3	16	14.03
4	7	14.77
5	10	12.44
6	4	8.72
7	2	5.24
≥ 8	9	4.91

Number of extinctions	Number of Time Intervals	Predicted Time Intervals

- Now, let's plot the grouped data.

```
# store data in data.frame
ext2.DF=data.frame(n.extinctions=0:7,
    n.extinctions.label=c('0 or 1', 2:7, '8+'),
    n.intervals=c(13,15,16,7,10,4,2,9),
    predicted.freqs=c(5.88,10,14.03,14.77,12.44,8.72,5.24,4.91))

ggplot(data=ext2.DF) +
    geom_bar(aes(x=n.extinctions, y=n.intervals), stat='identity') +
    geom_line(aes(x=n.extinctions, y=predicted.freqs)) +
    geom_point(aes(x=n.extinctions, y=predicted.freqs)) +
    theme_classic() +
    xlab('Number of extinctions') +
    ylab('Time interval between extinctions')
```



- Now, let's use the χ^2 test to see if our model adequately describes the data, and in turn, see if extinctions cluster in time.
  - First we will have to calculate our test statistic.

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$
$$\chi^2 = \frac{13 - 5.88}{5.88} + \frac{15 - 10}{10} + \frac{16 - 14.03}{14.03} + \frac{7 - 14.77}{14.77} + \frac{10 - 12.44}{12.44} + \frac{4 - 8.72}{8.72} + \frac{2 - 5.24}{5.24} + \frac{9 - 4.91}{4.91}$$
$$\chi^2 = 23.929$$

- Second, we will need to calculate the degrees of freedom (df). This can be calculated as the number of samples in our grouped data (8) minus 2. Thus our $df = 6$.
- Finally, at long last, we can calculate our p-value using the `R` code.

```
pchisq(23.929, df=6, lower.tail=FALSE)
```

- Our p-value is 0.0005382. Therefore the predictions from our model are significantly different to the observed data, and so we can reject the null hypothesis. Thus we do not yet have evidence to suggest that extinctions cluster in time.

R practical session

P1. Example 1: mendelian ratios

```
Observed <- c(56,19,17,8)                # enter observed counts
# calculate expected counts
ratio <- c(9,3,3,1)                    # expected ratios
probs <- ratio/sum(ratio)                # convert ratios to proportions
Expected <- sum(Observed)*probs      # distribute data across those proportions
Expected                                 # look at the expected counts
```

```
## [1] 56.25 18.75 18.75  6.25
```

```
# compare observed and expected using Pearson's chi-square statistic
x2 <- sum((Observed-Expected)^2/Expected)    # Pearson's chi-square statistic
x2                                       # look at Pearson's test-statistic
```

```
## [1] 0.6577778
```

```
pchisq(x2, df=3, lower.tail=FALSE)       # estimate p-value
```

```
## [1] 0.8830872
```

P2. Example 2: poisson counts

```
Observed <- c(16,19,9,4,2)              # enter observed counts
# calculate expected counts
lambda <- (0:4 %*% Observed)/sum(Observed)  #  poisson mean (lambda) of data
probs <- c(dpois(0:3, lambda), 1-sum(dpois(0:3, lambda)))   # expected proportions
Expected <- sum(Observed)*probs     # distribute data across those proportions
Expected                                # look at the expected counts
```

```
## [1] 15.990951 18.229684 10.390920  3.948550  1.439895
```

```
# compare observed and expected using Pearson's chi-square statistic
x2 <- sum((Observed-Expected)^2/Expected)   # Pearson's chi-square statistic
x2                                      # look at Pearson's test-statistic
```

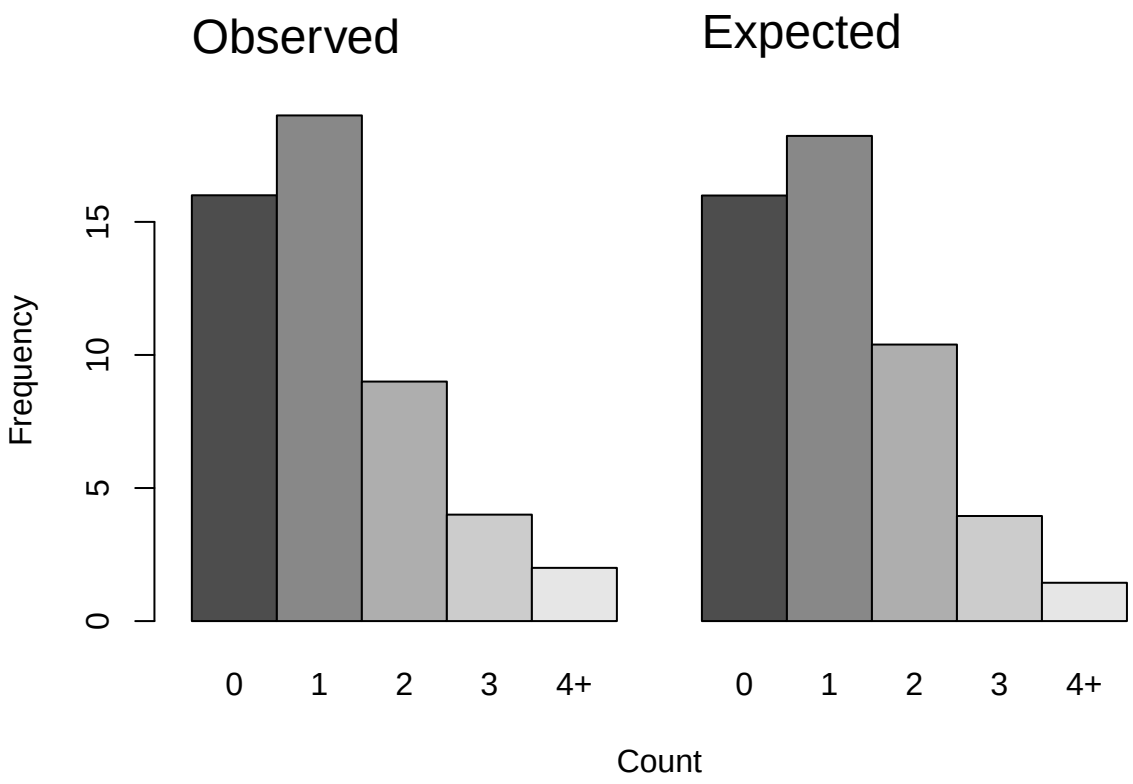
```
## [1] 0.4372888
```

```
pchisq(x2, df=3, lower.tail=FALSE)      # estimate p-value
```

```
## [1] 0.9324362
```

P3. Barplot comparing observed and expected

```
# align observed and expected data
PoissonModelData <- as.matrix(cbind(Observed,Expected))
# Display Observed and Expected as two barplots side-by-side
Xlocations <- barplot(PoissonModelData, #enter two aligned columns of numbers
beside=T,# put barplots side-by-side on the same graph
names.arg=c("0","1","2","3","4+","0","1","2","3","4+"),# X-axis tick labels
ylab="Frequency",                                   # label the vertical axis
xlab="Count")                                       # label the horizontal axis
par(xpd=T)                                  # allow printing outside of plot area
text(1,22,"Observed",cex=1.5,adj=0)                 # Title for first barplot
text(7,22,"Expected",cex=1.5,adj=0)                 # Title for second barplot
```



P4. Example 3: contingency table analysis

```
Observed <- matrix(c(5,8,9,12,2,4),nrow=2,ncol=3,byrow=TRUE)
Observed # have a look at the table of observed counts
```

```
##      [,1] [,2] [,3]
## [1,]    5    8    9
## [2,]   12    2    4
```

```
# calculate expected counts
SexCount    <- matrix( rowSums(Observed),nrow=2,ncol=1 ) # calculate row totals
StatusCount <- matrix( colSums(Observed),nrow=1,ncol=3 ) # column totals
Expected <- (SexCount %*% StatusCount )/sum(Observed) # expected counts
Expected                                  # look at the expected counts
```

```
##      [,1] [,2] [,3]
## [1,] 9.35  5.5 7.15
## [2,] 7.65  4.5 5.85
```

```r
# compare observed and expected using Pearson's chi-square statistic
x2 <- sum((Observed-Expected)^2/Expected)   # Pearson's chi-square statistic
x2                                         # look at Pearson's test-statistic
```

```
## [1] 8.086293
```

```r
pchisq(x2, df=2, lower.tail=FALSE)      # estimate p-value
```

```
## [1] 0.01754219
```

P5. Log-linear modelling

```r
Observed <- c(5,8,9,12,2,4)                          # enter observed counts
Sex <- c("F","F","F","M","M","M")              # enter the genders
Status <- c("Reg","Occ","Non","Reg","Occ","Non")    # enter 'smoking status'
table <- data.frame(Observed,Sex,Status)         # collect data as a data.frame
attach(table)                                    # make it the default data set
```

```
## The following objects are masked _by_ .GlobalEnv:
##
##     Observed, Sex, Status
```

```r
# Fit a loglinear regression model
library(MASS) # load the library that contains the 'loglm' function
# regress 'Status' and 'Sex' onto 'Observed'
fit <- loglm(Observed ~ Sex + Status)
anova(fit)                                       # look at the test statistics
```

```
## Call:
## loglm(formula = Observed ~ Sex + Status)
##
## Statistics:
##                       X^2 df   P(> X^2)
## Likelihood Ratio 8.397656  2 0.01501316
## Pearson          8.086293  2 0.01754219
```

P6. Generalized-linear modelling

```r
# regress 'Status', 'Sex' and their interaction onto 'Observed'
fit <- glm(Observed ~ Sex * Status, family="poisson")
anova(fit,test="Chisq")              # look at the test statistics
```

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: Observed
##
## Terms added sequentially (first to last)
##
##
##            Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                           5    10.6466
## Sex         1   0.4007         4    10.2459  0.52674
## Status      2   1.8483         2     8.3977  0.39688
## Sex:Status  2   8.3977         0     0.0000  0.01501 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(fit)                                # look at regression coefficients
```

```
##
## Call:
## glm(formula = Observed ~ Sex * Status, family = "poisson")
##
## Deviance Residuals:
## [1]  0  0  0  0  0  0
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)      2.1972     0.3333   6.592 4.35e-11 ***
## SexM            -0.8109     0.6009  -1.349   0.1772
## StatusOcc       -0.1178     0.4859  -0.242   0.8085
## StatusReg       -0.5878     0.5578  -1.054   0.2920
## SexM:StatusOcc  -0.5754     0.9930  -0.579   0.5623
## SexM:StatusReg   1.6864     0.8028   2.101   0.0357 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance:  1.0647e+01  on 5  degrees of freedom
## Residual deviance: -4.4409e-16  on 0  degrees of freedom
## AIC: 33.688
##
## Number of Fisher Scoring iterations: 3
```