HW11

Sumeet Mishra

4/12/2019

1. (5 points) Trosset Section 13.4 Exercise 2

Null hypothesis is M&M milk chocolate is mixed in proportions as claimed by Mars company that is {0.13, 0.14, 0.13, 0.24, 0.20, 0.16} Alternate hypothesis is M&M milk chocolate is not mixed in proportions as claimed by the Mars company.

Assume, significance level, $\alpha = 0.05$

```
observed=c(121, 84, 118, 226, 226, 123)
obs_sum=sum(observed)
prop=c(0.13, 0.14, 0.13, 0.24, 0.20, 0.16)
expected=rep(prop*obs_sum)
```

The pvalue in likelyhood ratio test is-

```
G2 = 2 * sum(observed * log(observed/expected))
pvalue= 1 - pchisq(G2, df=5)
pvalue
```

```
## [1] 1.141029e-05
```

The pvalue in Pearson's test is-

```
X2 = sum((observed - expected)^2 / expected)
pvalue= 1 - pchisq(X2, df=5)
pvalue
```

```
## [1] 1.860203e-05
```

As the pvalue of both the tests is $< \alpha$, we reject null hypothesis and conclude that the M&M milk chocolates are not mixed in proportions as claimed by the Mars company.

2. (7 points) Trosset Section 13.4 Exercise 5

(a)

```
fx=c()
for (x in 1:9) {
fx[x]=log10(1+(1/x)) }
sum=sum(fx)
sum
```

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

```
fx
```

```
## [1] 0.30103000 0.17609126 0.12493874 0.09691001 0.07918125 0.06694679
## [7] 0.05799195 0.05115252 0.04575749
```

Since sum of all the probabilities is 1 and each probability is in between 0 and 1, we can say that the PMF is valid.

(b)

Null hypothesis is The leading digits of the town populations follow Benford's law. Alternate hypothesis is The leading digits of the town populations does not follow Benford's law.

Assume, significance level, $\alpha = 0.05$

```
observed=c(107, 55, 39, 22, 13, 18, 13, 23, 15)
obs_sum=sum(observed)
expected=rep(fx*obs_sum)
```

The pvalue in likelyhood ratio test is-

```
G2 = 2 * sum(observed * log(observed/expected))
pvalue= 1 - pchisq(G2, df=8)
pvalue
```

```
## [1] 0.04919622
```

As the pvalue of the Likelihood ratio test is $< \alpha = 0.05$, we reject null hypothesis and conclude that the leading digits of the town populations does not follow Benford's law.

The pvalue in Pearson's test is-

```
X2 = sum((observed - expected)^2 / expected)
pvalue= 1 - pchisq(X2, df=8)
pvalue
```

```
## [1] 0.06399094
```

As the pvalue of the Pearson test is $> \alpha = 0.05$, we fail to reject null hypothesis and conclude that the leading digits of the town populations follow Benford's law.

3. (5 points) Trosset Section 13.4 Exercise 11

Null hypothesis is The patient's response to treatment for Hodgkin's disease is independent of histological type. Alternate hypothesis is The patient's response to treatment for Hodgkin's disease is varies by histological type. Assume, significance level, $\alpha = 0.05$

```
observed1=c(74, 18, 12, 68, 16, 12, 154, 54, 58, 18, 10, 44)
N=sum(observed1)
LP=(74+18+12)/N
NS=(68+16+12)/N
MC=(154+54+58)/N
LD=(18+10+44)/N
positive=c(74+68+154+18)/N
partial=c(18+16+54+10)/N
none=c(12+12+58+44)/N
expected1=c(LP*positive*N,LP*partial*N,LP*none*N,NS*positive*N,NS*partial*N, NS*none*N, MC*positive*N, 
data.frame(observed1, expected1=round(expected1, 1))
```

```
##    observed1 expected1
## 1         74      60.7
## 2         18      18.9
## 3         12      24.4
## 4         68      56.0
## 5         16      17.5
## 6         12      22.5
## 7        154     155.2
## 8         54      48.5
## 9         58      62.3
## 10        18      42.0
## 11        10      13.1
## 12        44      16.9
```

The pvalue in likelyhood ratio test is-

```
df=(4-1)*(3-1)
G2 = 2 * sum(observed1 * log(observed1/expected1))
pvalue= 1 - pchisq(G2, df=df)
pvalue
```

```
## [1] 9.139356e-13
```

The pvalue in Pearson's test is-

```
X2 = sum((observed1 - expected1)^2 / expected1)
pvalue= 1 - pchisq(X2, df=df)
pvalue
```

```
## [1] 2.520206e-14
```

As the pvalue of both the tests is $< \alpha = 0.05$, we reject null hypothesis and conclude that the patient's response to treatment for Hodgkin's disease varies by histological type.

4.

```
EPL201415 = read.csv("http://www.football-data.co.uk/mmz4281/1415/E0.csv")
```

(a) Home Goals

Null hypothesis is the home team goals follows a poisson distribution. Alternate hypothesis is the home team goals does not follow a poisson distribution. Assume, significance level, $\alpha = 0.05$

```
homegoals = EPL201415$FTHG[1:380]
data.frame(table(homegoals))
```

```
##   homegoals Freq
## 1         0   92
## 2         1  119
## 3         2  102
## 4         3   46
## 5         4   12
## 6         5    5
## 7         6    3
## 8         8    1
```

```
games=length(homegoals)
goals=sum(homegoals)
ave=goals/games
expected = games * dpois(0:20,ave)
data.frame(goals=0:20,expected=round(expected, 1))
```

```
##    goals expected
## 1      0     87.1
## 2      1    128.3
## 3      2     94.5
## 4      3     46.4
## 5      4     17.1
## 6      5      5.0
## 7      6      1.2
## 8      7      0.3
## 9      8      0.0
## 10     9      0.0
## 11    10      0.0
```

```
## 12    11      0.0
## 13    12      0.0
## 14    13      0.0
## 15    14      0.0
## 16    15      0.0
## 17    16      0.0
## 18    17      0.0
## 19    18      0.0
## 20    19      0.0
## 21    20      0.0
```

As the expected values is less than 5 for goals 6 and above, we combine 5 or more goals.

```
observed=c(92,119,102,46,12,9)
expected=rep(NA, 6)
expected[1:5]=games * dpois(0:4,ave)
expected[6]=games * (1 - ppois(4,ave))
data.frame(observed, expected=round(expected, 1))
```

```
##   observed expected
## 1       92     87.1
## 2      119    128.3
## 3      102     94.5
## 4       46     46.4
## 5       12     17.1
## 6        9      6.6
```

The pvalue in likelyhood ratio test is-

```
df=4
G2 = 2 * sum(observed * log(observed/expected))
pvalue= 1 - pchisq(G2, df=df)
pvalue
```

```
## [1] 0.4015313
```

The pvalue in Pearson's test is-

```
X2 = sum((observed - expected)^2 / expected)
pvalue= 1 - pchisq(X2, df=df)
pvalue
```

```
## [1] 0.4131955
```

As the pvalue of both the tests is $> \alpha = 0.05$, we fail to reject null hypothesis and conclude that the home team goals follows a poisson distribution.

(b) Away Team goals

Null hypothesis is the away team goals follows a poisson distribution. Alternate hypothesis is the away team goals does not follow a poisson distribution. Assume, significance level, $\alpha = 0.05$

```
awaygoals = EPL201415$FTAG[1:380]
data.frame(table(awaygoals))
```

```
##   awaygoals Freq
## 1         0  132
## 2         1  134
## 3         2   73
## 4         3   32
```

```
## 5          4    7
## 6          5    1
## 7          6    1
```

```
games=length(awaygoals)
goals=sum(awaygoals)
ave=goals/games
expected = games * dpois(0:20, ave)
data.frame(goals=0:20, expected=round(expected, 1))
```

```
##    goals expected
## 1      0    127.5
## 2      1    139.2
## 3      2     76.0
## 4      3     27.7
## 5      4      7.6
## 6      5      1.7
## 7      6      0.3
## 8      7      0.0
## 9      8      0.0
## 10     9      0.0
## 11    10      0.0
## 12    11      0.0
## 13    12      0.0
## 14    13      0.0
## 15    14      0.0
## 16    15      0.0
## 17    16      0.0
## 18    17      0.0
## 19    18      0.0
## 20    19      0.0
## 21    20      0.0
```

As the expected values is less than 5 for goals 5 and above , we combine 4 or more goals.

```
observed=c(132,134,73,32,9)
expected=rep(NA, 5)
expected[1:4]=games * dpois(0:3,ave)
expected[5]=games * (1 - ppois(3, ave))
data.frame(observed, expected=round(expected, 1))
```

```
##   observed expected
## 1      132    127.5
## 2      134    139.2
## 3       73     76.0
## 4       32     27.7
## 5        9      9.6
```

The pvalue in likelyhood ratio test is-

```
df=3
G2 = 2 * sum(observed * log(observed/expected))
pvalue= 1 - pchisq(G2, df=df)
pvalue
```

```
## [1] 0.7637317
```

The pvalue in Pearson's test is-

```
X2 = sum((observed - expected)^2 / expected)
pvalue= 1 - pchisq(X2, df=df)
pvalue
```

```
## [1] 0.7566041
```

As the pvalue of both the tests is $> \alpha = 0.05$, we fail to reject null hypothesis and conclude that the away team goals follows a poisson distribution.

(c) Total Goals

Null hypothesis is the total team goals follows a poisson distribution. Alternate hypothesis is the total team goals does not follow a poisson distribution. Assume, significance level, $\alpha = 0.05$

```
totalgoals = EPL201415$FTAG + EPL201415$FTHG
totalgoals = totalgoals[1:380]
data.frame(table(totalgoals))
```

```
##    totalgoals Freq
## 1           0   31
## 2           1   77
## 3           2   88
## 4           3   85
## 5           4   56
## 6           5   27
## 7           6    9
## 8           7    3
## 9           8    3
## 10          9    1
```

```
games=length(totalgoals)
goals=sum(totalgoals)
ave=goals/games
expected = games * dpois(0:20, ave)
data.frame(goals=0:20, expected=round(expected, 1))
```

```
##    goals expected
## 1      0     29.2
## 2      1     74.9
## 3      2     96.1
## 4      3     82.2
## 5      4     52.7
## 6      5     27.1
## 7      6     11.6
## 8      7      4.2
## 9      8      1.4
## 10     9      0.4
## 11    10      0.1
## 12    11      0.0
## 13    12      0.0
## 14    13      0.0
## 15    14      0.0
## 16    15      0.0
## 17    16      0.0
## 18    17      0.0
## 19    18      0.0
## 20    19      0.0
```

```
## 21    20       0.0
```

As the expected values is less than 5 for goals 7 and above, we combine 7 or more goals.

```
observed=c(31,77,88,85,56,27,9,7)
expected=rep(NA, 8)
expected[1:7]=games * dpois(0:6,ave)
expected[8]=games * (1 - ppois(6,ave))
data.frame(observed,expected=round(expected, 1))
```

```
##   observed expected
## 1       31     29.2
## 2       77     74.9
## 3       88     96.1
## 4       85     82.2
## 5       56     52.7
## 6       27     27.1
## 7        9     11.6
## 8        7      6.1
```

The pvalue in likelyhood ratio test is-

```
df=6
G2 = 2 * sum(observed * log(observed/expected))
pvalue= 1 - pchisq(G2, df=df)
pvalue
```

```
## [1] 0.9282545
```

The pvalue in Pearson's test is-

```
X2 = sum((observed - expected)^2 / expected)
pvalue= 1 - pchisq(X2, df=df)
pvalue
```

```
## [1] 0.9329698
```

As the pvalue of both the tests is $> \alpha = 0.05$, we fail to reject null hypothesis and conclude that the total goals follows a poisson distribution.

5.

(a)

Null hypothesis is anger is not asociated with heart disease. Alternate hypothesis is anger is asociated with heart disease. Assume, significance level, $\alpha = 0.05$

```
observed=c(53, 3110-53, 110, 4731-110, 27, 633-27)
n=sum(observed)
la=3110/n
ma=4731/n
ha=633/n
hd=(53+110+27)/n
no_hd=c(3057+4621+606)/n
expected=c(la*hd*n, la*no_hd*n,ma*hd*n, ma*no_hd*n,ha*hd*n, ha*no_hd*n)
data.frame(observed, expected=round(expected, 1))
```

```
##   observed expected
## 1       53     69.7
## 2     3057   3040.3
```

```
## 3      110    106.1
## 4     4621   4624.9
## 5       27     14.2
## 6      606    618.8
```

The pvalue in likelyhood ratio test is-

```
df=(3-1)*(2-1)
G2 = 2 * sum(observed * log(observed/expected))
G2
```

```
## [1] 13.99914
```

```
pvalue= 1 - pchisq(G2, df=df)
pvalue
```

```
## [1] 0.0009122731
```

The pvalue in Pearson's test is-

```
X2 = sum((observed - expected)^2 / expected)
X2
```

```
## [1] 16.07676
```

```
pvalue= 1 - pchisq(X2, df=df)
pvalue
```

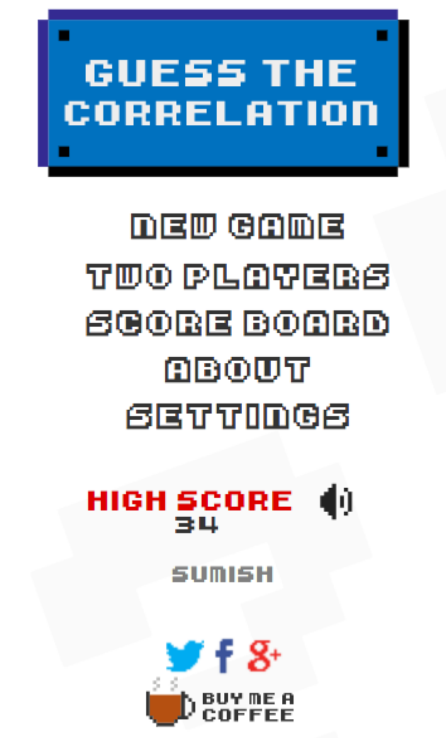
```
## [1] 0.0003228312
```

As the pvalue of both the tests is $< \alpha = 0.05$, we reject null hypothesis and conclude that anger is associated with heart disease.

(b) This analysis alone does not prove that anger affects the chance of getting heart disease because from the data we see that 2% of the people with low and moderate anger got heart disease whereas 4% of people with high anger got heart disease which shows samples are not generated uniformly. Moreover, the chi-squared test does not consider factors such as gender, age, patients health history, genetic factors which can heavily influence the chance of developing heart disease. We are not sure if these factors are considered when generating the samples.

6.

```
library(webshot)
webshot('GTC.png')
```


GUESS THE CORRELATION
NEW GAME
TWO PLAYERS
SCORE BOARD
ABOUT
SETTINGS
HIGH SCORE
34
SUMISH
BUY ME A COFFEE