Homework 11

1

- H0 : P(brown) = 0.13, P(yellow) = 0.14, P(Red) = 0.13, P(Blue) = 0.24, P(Orange) = 0.20, P(Green) = 0.16
- H1 : The proportions of candies are different from that of the above

```
observed <- c(121, 84, 118, 226, 226, 123)
p_expected <- c(0.13, 0.14, 0.13, 0.24, 0.20, 0.16)

sum_observed <- sum(observed)
expected <- sum_observed*p_expected

#likelihood ratio version
G2 <- 2*sum(observed*log(observed/expected))
1 - pchisq(G2, df=5)
```

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

```
x2 <- sum((observed-expected)^2/expected)
1 - pchisq(x2, df=5)
```

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

In both cases the p value is way less than 0.05, thus we can reject the null hypothesis. The proportions are different from the ones given above.

2

a

```
fx <- log10(1+1/(1:9))

sum(fx)
```

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

sum of all probabilities in a pmf is 1. Thus this function does signify a pmf

b

- H0 : p(1) = 0.3, p(2) = 0.17, p(3) = 0.12, p(4) = 0.09, p(5) = 0.07 p(6) = 0.06, p(7) = 0.05, p(8) = 0.05, p(9) = 0.04
- H1 : The probabilities are different from the ones mentioned above

```
observed <- c(107, 55, 39, 22, 13, 18, 13, 23, 15)
expected <- sum(observed)*fx

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

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

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

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

The probability calculated by the likelihood version is less than 0.05, where as the probability caculated by pearson's version is more than 0.05. Thus the conclusion depends on the choice of the test.

Thus we reject the null hypothesis under the likelihood test. And thus the probabilities are different.

3

- H0 : patients of Hodgkin's disease with different histological type is independent of its reponse to treatment
- H1 : patients of Hodgkin's disease with different histological type are not independent of its reponse to treatment

```
observed <- c(74, 18, 12, 68, 16, 12,
              154, 54, 58, 18, 10, 44)

N <- sum(observed)
P.LP <-sum(c(74, 8, 12))/N
P.NS <- sum(c(68, 16, 12))/N
P.MC <- sum(c(154, 54, 58))/N
P.LD <- sum(c(18, 10, 44))/N
P.Pos <- sum(c(74, 68, 154, 18))/N
P.Par <- sum(c(18, 16, 54, 10))/N
P.None <- sum(c(12, 12, 58, 44))/N

colums <- c(P.Pos, P.Par, P.None)
```

```r
expected <- N*c(P.LP*colums, P.NS*colums, P.MC*colums, P.LD*colums)

df <- (3-1)*(4-1)

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

```
## [1] 0
```

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

```
## [1] 9.547918e-15
```

Both the test give very low probabilities, thus we can reject the null hypothesis ans thus it can be said that patients of Hodgkin's disease with different histological type are not independent of its reponse to treatment.

4

Note : The data is imported in my workspace, but doesn't showup in the pdf for some reason. Tried to resolve this issue but it didn't work.

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

```
## Warning in file(file, "rt"): URL 'http://www.football-data.co.uk/
## mmz4281/1415/E0.csv': status was 'Couldn't resolve host name'
```

```
## Error in file(file, "rt"): cannot open the connection to 'http://www.football-data.co.uk/
```

```r
data <- data[1:380,]
```

```
## Error in data[1:380, ]: object of type 'closure' is not subsettable
```

a

```r
max(data$FTHG)
```

```
## Error in data$FTHG: object of type 'closure' is not subsettable
```

```r
observed <- c()
for(i in 0:8){
    observed <- c(observed, sum(data$FTHG == i))
}
```

```
## Error: object of type 'closure' is not subsettable
```

```r
games <- sum(observed)
goals <- sum(c(0:8)*observed)
average <- sum(c(0:8)*observed)/sum(observed)
```

```r
round(games*dpois(0:20,average), 1)
```

```
##  [1] NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN
## [18] NaN NaN NaN NaN
```

```r
#6 categories above 5
# 5 or more
#eliminate 3

expected <- rep(NA, 6)
expected[1:5] = games * dpois(0:4, average)
expected[6] = games * (1 - ppois(4, average))

observed <-c(92, 119, 102, 46, 12, 9)
G2 <- 2*sum(observed*log(observed/expected))
1 - pchisq(G2, df=4)
```

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

```r
X2 <- sum((observed-expected)^2/expected)
1 - pchisq(X2,df=4)
```

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

Here both the tests show a result greater than 0.05. Thus we can say that the distribution follows a poisson distribution.

b

```r
max(data$FTAG)
```

```
## Error in data$FTAG: object of type 'closure' is not subsettable
```

```r
observed <- c()
for(i in 0:6){
    observed <- c(observed, sum(data$FTAG == i))
}
```

```
## Error: object of type 'closure' is not subsettable
```

```r
games <- sum(observed)
goals <- sum(c(0:6)*observed)
average <- sum(c(0:6)*observed)/sum(observed)

round(games*dpois(0:20,average), 1)
```

```
##  [1] NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN
## [18] NaN NaN NaN NaN
```

```r
#5 categories above 5
# 4 or more
#eliminate 2

expected <- rep(NA, 5)
expected[1:4] = games * dpois(0:3, average)
expected[5] = games * (1 - ppois(3, average))

observed <- c(132, 134, 73, 32, 9)
G2 <- 2*sum(observed*log(observed/expected))
1 - pchisq(G2, df=3)
```

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

```r
X2 <- sum((observed-expected)^2/expected)
1 - pchisq(X2,df=3)
```

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

Here both the tests give a result that is more than 0.05. Thus we can say that the given distribution does follow a poisson distribution.

c

```r
combined <- data$FTHG+data$FTAG
```

```
## Error in data$FTHG: object of type 'closure' is not subsettable
```

```r
observed <- c()
for(i in 0:9){
    observed <- c(observed, sum(combined == i))
}
```

```
## Error in eval(expr, envir, enclos): object 'combined' not found
```

```r
games <- sum(observed)
goals <- sum(c(0:9)*observed)
average <- sum(c(0:9)*observed)/sum(observed)

round(games*dpois(0:20,average), 1)
```

```
##  [1] NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN
## [18] NaN NaN NaN NaN
```

```r
#6 categories above 5
# 5 or more
#eliminate 3

expected <- rep(NA, 8)
expected[1:7] = games * dpois(0:6, average)
```

```
expected[8] = games * (1 - ppois(6, average))

observed <- c(31, 77, 88, 85, 56, 27, 9, 7)

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

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

```
X2 <- sum((observed-expected)^2/expected)
1 - pchisq(X2,df=6)
```

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

Here both the tests give a result that is more than 0.05. Thus we can say that the given distribution does follow a poisson distribution.

5

- H0 : anger is independent of getting heart disease
- H1 : anger is not independent of getting heart disease

a

```
observed <- c(3057, 4621, 606, 53, 110, 27)
N <- sum(3110, 4731, 633)

P.No <- sum(3057+4621+606)/N
P.Yes <- sum(53+110+27)/N
P.Low <- sum(3057+53)/N
P.M <- sum(4621+110)/N
P.H <- sum(606+27)/N

rows <- c(P.Low, P.M, P.H)
expected <- N*c(P.No*rows, P.Yes*rows)

df <- (3-1)*(2-1)

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

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

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

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

Both the test show very low probabilities, thus we reject the null hypothesis and so we can say that anger is not independent of getting heart disease.

b

The analysis only prove that anger varies with getting heart disease as using chi squared test, but we cannot conclude that anger affects the chance of getting a heart disease. For this, we need to control other factors like his medical conditions, age, cholestrol level, etc.