Statistics for Engineers (MAT2001)- Lab Experiment-VIII: Chi-square Test

1 Chi-square test for independence of attributes

```
chisq.test(x, y = NULL, p )
```

- x - a numeric vector or matrix. x and y can also both be factors.
- y - a numeric vector; ignored if x is a matrix. If x is a factor, y should be a factor of the same length.
- p - a vector of probabilities of the same length as x. An error is given if any entry of p is negative.

The below table gives the distribution of students according to the family type and the anxiety level

Family	Anxiety level		
	Low	Normal	High
Joint family	35	42	61
Nuclear family	48	51	68

```
family <- matrix(c(35, 42, 61, 48, 51, 68), nrow = 2, byrow = TRUE)
family
```

```
##      [,1] [,2] [,3]
## [1,]   35   42   61
## [2,]   48   51   68
```

```
chisq.test(family)
```

```
##
##  Pearson's Chi-squared test
##
## data:  family
## X-squared = 0.53441, df = 2, p-value = 0.7655
```

Here P value $0.7655 > 0.05$. Hence there is no evidence to reject the null hypothesis. So we consider the anxiety level and family type as independent.

In the built-in data set survey, the Smoke column records the students smoking habit, while the Exer column records their exercise level. The allowed values in Smoke are "Heavy", "Regul" (regularly), "Occas" (occasionally) and "Never". As for Exer, they are "Freq" (frequently), "Some" and "None". We can tally the students smoking habit against the exercise level with the table function in R. The result is called the contingency table of the two variables. Test the hypothesis whether the students smoking habit is independent of their exercise level at 5% significance level.

```
library(MASS)
data <- table(survey$Smoke, survey$Exer)
data
```

```
##
##         Freq None Some
##   Heavy    7    1    3
##   Never   87   18   84
##   Occas   12    3    4
##   Regul    9    1    7
```

```
chisq.test(data)
```

```
## Warning in chisq.test(data):  Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  data
## X-squared = 5.4885, df = 6, p-value = 0.4828
```

```
new_data <- cbind(data[, "Freq"], data[, "None"]+data[, "Some"])
new_data
```

```
##       [,1] [,2]
## Heavy    7    4
## Never   87  102
## Occas   12    7
## Regul    9    8
```

```
chisq.test(new_data)
```

```
##
##  Pearson's Chi-squared test
##
## data:  new_data
## X-squared = 3.2328, df = 3, p-value = 0.3571
```

A biologist is conducting a plant breeding experiment in which plants can have one of four phenotypes. If these phenotypes are caused by a simple Mendelian model, the phenotypes should occur in a 9:3:3:1 ratio. She raises 41 plants with the following phenotypes.
Phenotype: 1 2 3 4
count: 20 10 7 4
Should she worry that the simple genetic model doesn't work for her phenotypes?

```
plants <- c(20, 10 ,7 ,4)
chisq.test(plants, p=c(9/16, 3/16, 3/16, 1/16))

## Warning in chisq.test(plants, p = c(9/16, 3/16, 3/16, 1/16)):  Chi-squared
approximation may be incorrect

##
##  Chi-squared test for given probabilities
##
## data:  plants
## X-squared = 1.9702, df = 3, p-value = 0.5786
```

The Chi-squared distribution is only an approximation to the sampling distribution of our test statistic, and the approximation is not very good when the expected cell counts are too small. This is the reason for the warning. Here the probability value p is greater than alpha level (0.05), so we do not reject the null hypothesis.

A survey of 320 families with 5 children each revealed the following distribution:
Number of Boys: 5 4 3 2
No of Girls: 0 1 2 3 4 5
No of families: 14 56 110 88 40 12
Is this result consistent with the hypothesis that male and female births are equally possible?
Solution: Let us setup the null hypothesis that the data are consistent with the hypothesis of equal probability for male and female births.

```
x <- c(5, 4, 3, 2 ,1)
n=5
N=320
p <- 0.5
cbf <- c(14, 56, 110, 88, 40, 12)
exf <- dbinom(x, n, p)*N
exf

## [1]  10  50 100 100  50
```

```
chisq <- sum((cbf-exf)^2/exf)

## Warning in cbf - exf:  longer object length is not a multiple of
shorter object length
## Warning in (cbf - exf)^2/exf:  longer object length is not a multiple
of shorter object length

chisq

## [1] 7.16

qchisq(0.95, 5)

## [1] 11.0705
```

Calculated value of chi-square is less than the tabulated value,it is not significant at 5% level of significance and hence the null hypothesis of equal probability for male and female births.

Fit a Poisson distribution to the following data and test the goodness of fit
X: 0 1 2 3 4 5 6
f: 275 72 30 7 5 2 1

```
x = 0:6
f <- c(275, 72, 30, 7, 5, 2, 1)
f

## [1] 275  72  30   7   5   2   1

N <- sum(f)
lambda <- (sum(f*x))/N
exf <- dpois(x, lambda)*N
exf <- round(exf)
exf

## [1] 242 117  28   5   1   0   0

new_obs <- c(75, 72, 30, 15)
new_exf <- c(242, 117, 28, 6)
chisq <-  sum((new_obs-new_exf)^2/new_exf)
chisq

## [1] 146.1944

qchisq(0.95, 2)

## [1] 5.991465
```

Since calculated value of $x = 146.1944$ is much greater than 5.99, it is highly significant Hence we conclude that poisson distribution is not good fit to the given data

Experiment

1. The following data come from a hypothetical survey of 920 people (Men, Women) preference of one of the three ice cream flavors. Is there any association between gender and preference for ice cream flavors.

	Chocolate	Vanilla	Strawberry
Men	100	120	60
Women	350	320	150

2. As a part of quality improvement project focused on a delivery of mail at a department office within a large company, data were gathered on the number of different addresses that had to be changed so that the mail could be redirected to thee correct mail stop. Table shows the frequency distribution. Fit binomial distribution and test goodness of fit
   x: 0 1 2 3 4
   fx: 5 20 45 20 10