LAB-12

CHI-SQUARE TEST

(INDEPENDENCE OF ATTRIBUTES AND GOODNESS OF FIT)

CHI-SQUARE TEST FOR INDEPENDENCE OF ATTRIBUTES:

Two random variables x and y are called independent if the probability distribution of one variable is not affected by the presence of another. Assume Oij is the observed frequency count of events belonging to both i-th category of x and j-th category of y. Also assume Eij to be the corresponding expected count if x and y are independent. The null hypothesis of the independence assumption is to be rejected if the p-value of the following Chi-squared test statistics is less than a given significance level α.

$$\chi^2 = \sum\left[\frac{(O_{ij} - E_{ij})^2}{E_{ij}}\right]$$

Problem 1 :The below table gives the distribution of students according to the family type and the anxiety level

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

R Code and Output:-

```
> data<-matrix(c(35,42,61,48,51,68),ncol=3,byrow=T)
> data
     [,1] [,2] [,3]
[1,]   35   42   61
[2,]   48   51   68
> chisq.test(data)

        Pearson's Chi-squared test

data:  data
X-squared = 0.53441, df = 2, p-value = 0.7655
```

Interpretation :-

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.

Problem 2: (Drug data):-

```
> data<-read.csv("C:\\Users\\aadmin\\Desktop\\chidata.csv")
> tab2=table(data$Drug,data$Age.Group)
> tab2

         21 to 55 Over 55 Under 21
  Drug A        8      12        4
  Drug B        4      18        8
  Drug C        4      11       24
> data<-read.csv("C:\\Users\\aadmin\\Desktop\\chidata.csv")
> tab2=table(data$Drug,data$Age.Group)
> tab2

         21 to 55 Over 55 Under 21
  Drug A        8      12       22
  Drug B       34      18        8
  Drug C       30      11       24
> chisq.test(tab2)

        Pearson's Chi-squared test

data:  tab2
X-squared = 23.223, df = 4, p-value = 0.0001143
```

Inference:-

Hence P value (0.0001143) < 0.05 .Hence reject our Null hypothesis. Finally conclude that There is a association between Type of Drug and Age Group.

Problem 3:

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.

R code:-

```
> library(MASS)
> tbl = table(survey$Smoke, survey$Exer)
> tbl

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

Test the hypothesis whether the students smoking habit is independent of their exercise level at .05 significance level.

```
> chisq.test(tbl)

        Pearson's Chi-squared test

data:  tbl
X-squared = 5.4885, df = 6, p-value = 0.4828

Warning message:
In chisq.test(tbl) : Chi-squared approximation may be incorrect
```

Interpretation :-

As the p-value 0.4828 is greater than the .05 significance level, we do not reject the null hypothesis that the smoking habit is independent of the exercise level of the students.

Enhanced Solution:

The warning message found in the solution above is due to the small cell values in the contingency table. To avoid such warning, we combine the second and third columns of tbl.

```
> ctbl = cbind(tbl[,"Freq"], tbl[,"None"] + tbl[,"Some"])
> ct

> ctbl
      [,1] [,2]
Heavy    7    4
Never   87  102
Occas   12    7
Regul    9    8
> chisq.test(ctbl)

        Pearson's Chi-squared test

data:  ctbl
X-squared = 3.2328, df = 3, p-value = 0.3571
```

Problem 4: (Goodness of Fit)

In the built-in data set survey, the Smoke column records the survey response about the student's smoking habit. As there are exactly four proper response in the survey: "Heavy", *"Regul" (regularly), "Occas" (occasionally) and "Never", the Smoke data is multinomial. It can be confirmed with the levels function in R.*

R code:-

```
> library(MASS)
> levels(survey$Smoke)
[1] "Heavy" "Never" "Occas" "Regul"
> smoke.freq = table(survey$Smoke)
> smoke.freq

Heavy Never Occas Regul
   11   189    19    17
```

Suppose the campus smoking statistics is as below. Determine whether the sample data insurvey supports it at .05 significance level.

```
> smoke.prob = c(.045, .795, .085, .075)
> chisq.test(smoke.freq, p=smoke.prob)

        Chi-squared test for given probabilities

data:  smoke.freq
X-squared = 0.10744, df = 3, p-value = 0.9909
```

Interpretation :

As the p-value 0.991 is greater than the .05 significance level, we do not reject the null hypothesis that the sample data in survey supports the campus-wide smoking statistics.

Problem 5:

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?

R code:-

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

        Chi-squared test for given probabilities

data:  plants
X-squared = 1.9702, df = 3, p-value = 0.5786

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

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.

Fitting of Binomial distribution with goodness of fit:-

Problem 6 : *A survey of 320 families with 5 children each revealed the following distribution:*

Number of Boys	*5*	*4*	*3*	*2*	*1*	*0*
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.

R code:-

```
> n=5                          #Probability of 'r' male births in a family
> N=320                        #Total Number of families
> P<-0.5                       #Probability of Male Birth
> Obf<-c(14,56,110,88,40,12)   #Obsevred frequencies
> exf<-dbinom(x,n,P)*320       #Expected frequencies
> # check the Condintion Sum of Observed and Expected are Equal
> sum(Obf)
[1] 320
> sum(exf)
[1] 320
> chisq<-sum((Obf-exf)^2/exf)
> chisq
[1] 7.16
> qchisq(0.95,5)
[1] 11.0705
```

Interpretation:-

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.

Fitting of Poisson distribution with goodness of fit:-

Problem 7: 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

Solution :-

R Code:-

```
> x<-0:6
> f<-c(275,72,30,7,5,2,1)
> lambda<-(sum(f*x)/sum(f))     #mean
> expf <-dpois(x,lambda)*sum(f)    #expcted frequencies
> f1=round(expf)
> # check obserevd and Expected frequencies Total
> sum(f)
[1] 392
> sum(f1)
[1] 393
> # here subtrat '1' from expected frequencies
> #The last 3 frequencies are less than 5 so combine these frequencies in Observation and Expected
> obf<-c(275,72,30,15)
> exf<-c(242,117,28,6)
> chisq<-sum(((obf-exf)^2)/exf)
> chisq
[1] 35.45055
> qchisq(0.95,2)
[1] 5.991465
```

(One d.f. being lost because of the linear constraint $\sum O = \sum E$; 1 d.f. is lost because the parameter m has been estimated from the given data and is then used for computing the expected frequencies; 3 d.f. are lost because of pooling the last four expected frequencies which are less than five expected cell frequencies which are less than five.)

Interpretation :-

Since calculated value of χ^2 =35.45055 is much greater than 5.99,it is highly significant .Hence we conclude that poisson distribution is not good fit to the given data

Fitting of Normal distribution with goodness of fit:-

Problem : The following table displays a frequency distribution of heights of trees in a certain locality. Fit a normal distribution to the data and test the goodness of fit.

Class Interval	Frequency
13.20 – 20.90	2
20.90 – 28.60	10

28.60 – 36.30	16
36.30 – 44.00	37
44.00 – 51.70	43
51.70 – 59.40	39
59.40 – 67.10	29
67.10 – 74.80	13
74.80 – 82.50	06
82.50 – 90.20	05

Heights of Trees (in inches)

R code :

```
> midy<-seq(17.05,86.5,length=10)
> f<-c(2,10,16,37,43,39,29,13,6,5)
> mean<-sum(f*midy)/sum(f)
> sd<-sqrt(sum(f*(midy-mean)^2)/sum(f))
> l<-seq(13.2,82.5,length=10)
> l<-c(l,90.2)
> cdf<-pnorm(l,mean,sd)
> cdf<-c(0,cdf,1)
> pcf<-diff(cdf)
> f<-c(0,f,0)
> ex<-round(pcf*sum(f),4)
> fr<-data.frame(f,ex)
> obf<-c(12,16,37,43,39,29,13,11)
> exf<-c(sum(ex[c(1,2,3)]),ex[c(4:9)],sum(ex[c(10,11,12)]))
> sum(obf)
[1] 200
> sum(exf)
[1] 200
> chisq<-sum((obf-exf)^2/exf)
> chisq
[1] 2.153974
> qchisq(0.95,5)
[1] 11.0705
```

Interpretation :-

Here chi-square cal value is less than chi-square tab value then there is no evidence to reject our null hypothesis.ie the fit of normal distribution is good

Practice Problems and Challenging Experiments:-

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

Gender\flavor	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

The number of Addresses Needing Change

3. A series of traps were set in line across sand dunes and the numbers of different types of insects caught in a fixed time interval are recorded to study their movement across the dune. Following table shows the data on the movement of leafhopper (Hemiptera) across a sand dune.

Leafhopper(Hemiptera) Per trap X_i	Frequency f_i
0	6
1	8
2	12
3	4
4 or more	3

Movement of Leafhopper Across a Sand Dune

Fit Poisson distribution to the above data and test goodness of fit.