STAT 40001/STAT 50001 Statistical Computing

Lecture 12

Department of Mathematics and Statistics

Chi-square Test

- Chi-square Test for Goodness-of-Fit
- Chi-square Test for Independence
- Chi-square Test for Homogeneity

Chi-Square Test for Goodness of Fit

This test can be used when

- The data have been randomly selected.
- The sample data consist of frequency counts for each of the different categories.
- All expected frequency are at least 5. (The expected frequency for a category is the frequency that would occur if the data actually have the distribution that is being claimed. There is no requirement that the observed frequency for each category must be at lest 5.)
- Another rule of thumb is that there are more than 4 groups, the average of the expected values is at least 5, and the smallest expected value is at least 1

Chi-square test for Goodness of Fit

This test is used to see if the data "fits" a distribution. The data will be divided into k groups. (Asking if the data fits the distribution, is the same as asking if we can predict how many items will be in each group.) We are interested to test

H_0 : the data fits the proposed distribution

H_a : the data does not fit the proposed distribution

OR

H_0 : the probability of each group is the expected probability

H_a : the probability of at least one group does not match the expected prob.

The test statistic for Goodness-of-fit test is

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^{k} \frac{(Observed - Expected)^2}{Expected}$$

where k is the number of categories. For large n, the test statistic above has approximately χ^2 distribution with $k-1$ degrees of freedom. Note that when there are multiple parameters then the degrees of freedom are reduced by one for each parameter that must be estimated.

Example

Professor Jackson takes a random sample of 95 students from his large statistics class. He finds that: there are 25 freshman, 32 sophomores, 18 juniors, and 20 seniors. Test the null hypothesis that freshman, sophomores, juniors, and seniors are equally represented.
Solution: Let p_1, p_2, p_3 and p_4 are probability of freshman, sophomores, juniors and seniors respectively. We would like to test

$$H_0 : p_1 = p_2 = p_3 = p_4$$
$$H_a : \text{at least one of the probabilities is different from others}$$

```
> chisq.test(c(25,32,18,20))
        Chi-squared test for given probabilities

data:  c(25, 32, 18, 20)
X-squared = 4.9158, df = 3, p-value = 0.1781
```

Decision: the null hypothesis cannot be rejected at $\alpha = .05$.

Example

The samhda data set in the UsingR package contains information about the health behavior of school-age children. The variable amt.smoke measures how often child smoke in the previous month. There are seven levels: 1 means the child smoke everyday to 7 means not at all. In particular it is coded as Amount of days you smoked cigarettes in last 30. 1 = all 30, 2= 20-29, 3 = 10-19, 4 = 6-9, 5= 3-5, 6 = 1-2, 7=0. Values 98 and 99 indicates missing data. We investigate whether the sample proportions agree with:
$p_1 = 0.15, p_2 = 0.05, p_3 = 0.05, p_4 = 0.05, p_5 = 0.1, p_6 = 0.2, p_7 = 0.40$.
We can use the R code below to perform the test

```
> library(UsingR)
> attach(samhda)
> x=table(amt.smoke[amt.smoke<98])
> x
  1   2   3   4   5   6   7
 32   7  13  10  14  43 105
> p=c(0.15,0.05,0.05,0.05,0.10,0.20,0.40)
> chisq.test(x,p=p)
        Chi-squared test for given probabilities
data:  x
X-squared = 7.9382, df = 6, p-value = 0.2427
```

Decision: Fail to reject the null hypothesis.

Example

The pi2000 data set in the UsingR package contains first 2000 digits of π. Perform a chi-squared significance test to see if the digits appear with equal probability

```
> library(UsingR)
> x=table((pi2000))
> chisq.test(x)
        Chi-squared test for given probabilities

data:  x
X-squared = 4.42, df = 9, p-value = 0.8817
```

Decision: There is no enough evidence to reject that the digits appears with equal probability.

Example

The number of murders by day of the week in New Jersey in 2003 is shown below (Source: http://www.njsp.org)

Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
53	42	51	45	36	37	65

Perform a significance test to test the null hypothesis that a murder is equally likely to occur on any given day

```
> chisq.test(c(53,42,51,45,36,37,65))

        Chi-squared test for given probabilities

data:  c(53, 42, 51, 45, 36, 37, 65)
X-squared = 13.3191, df = 6, p-value = 0.03824
```

Decision: Reject the null hypothesis. This means the murder is not equally likely to occur on any given day of the week.

Example

A cross between white and yellow summer squash gave progeny of the following colors

Color	White	Yellow	Green
Number of Progeny	155	40	10

Are these data consistent with the 12 : 3 : 1 ratio predicated by a certain genetic model?

Solution: We would like to test the following hypothesis

H_0 : *the model is correct with proportion of 12: 3: 1*

H_a : *The model is incorrect (the proportion are not as specified by above)*

```
> x=c(155,40,10)
> p=c(12/16,3/16,1/16)
> chisq.test(x,p=p)

        Chi-squared test for given probabilities

data:  x
X-squared = 0.6911, df = 2, p-value = 0.7078
```

Decision: We fail to reject the null hypothesis and conclude that the colors are consistent with the genetic model with given proportions.

Example

The data set Soccer in the PASWR package contains how many goals were scored in the regulation 90 minute periods of World Cup soccer matches from 1990 to 2002. Test the hypothesis that the number of goals scored has a Poisson distribution with $\lambda = 2.5$.

```
> library(PASWR)
> attach(Soccer)
> x=table(Goals)
> p=dpois(0:8,2.5)
> x
Goals
 0  1  2  3  4  5  6  7  8
19 49 60 47 32 18  3  3  1
> chisq.test(x,p=p)
Error in chisq.test(x, p = p) : probabilities must sum to 1
> p=c(dpois(0:7,2.5),1-ppois(7,2.5))
> chisq.test(x,p=p)

        Chi-squared test for given probabilities

data:  x
X-squared = 2.6769, df = 8, p-value = 0.953
```

Example

Use the chi-square goodness-of-fit test with $\alpha = 0.05$ to test the hypothesis that the SAT scores stored in the data frame Grades in the PASWR package have normal distribution.

```
> library(PASWR)
> attach(Grades)
> mu=mean(sat)
> sigma=sd(sat)
```

Because a normal distribution is continuous, it is necessary to create a categories that include all the data. We know that in normal distribution almost all should reside within 3 standard deviation of the mean.

```
> bin<-seq(mu-3*sigma,mu+3*sigma, sigma)
> bin
[1]697.824  843.4323 989.041 1134.650 1280.259 1425.867 1571.476
> x<-table(cut(sat,breaks=bin))
> x
(698,843]    (843,989]  (989,1.13e+03] (1.13e+03,1.28e+03]
  4            27            65               80
(1.28e+03,1.43e+03] (1.43e+03,1.57e+03]
     21                    3
```

Example- SAT scores

```
> x=c(4,27,65,80,21,3)
> p=c(pnorm(-2), pnorm(-1:2)-pnorm(-2:1),1-pnorm(2))
> p
[1]0.0227501 0.1359051 0.3413448 0.3413448 0.1359051 0.0227501
> chisq.test(x,p=p)
        Chi-squared test for given probabilities
data:  x
X-squared = 4.1737, df = 5, p-value = 0.5247
```

Decision: fail to reject H_0. There is no evidence to suggest that the true distribution of SAT scores is not normally distributed.

Chi-Square Test for Independence

A contingency table(or two way frequency table) is a table in which frequencies correspond to two variables; one variable is used to categorize rows, and a second variable is used to categorize columns.
The test of independence tests the null hypothesis that there is no association between the row variable and the column variable in a contingency table. In other word for null hypothesis we will use the statement that "the row and column variables are independent".
Example of Contingency table:

	Men	Women	Boys	Girls	Total
Survived	332	318	29	27	706
Died	1360	104	35	18	1517
Total	1692	422	64	45	2223

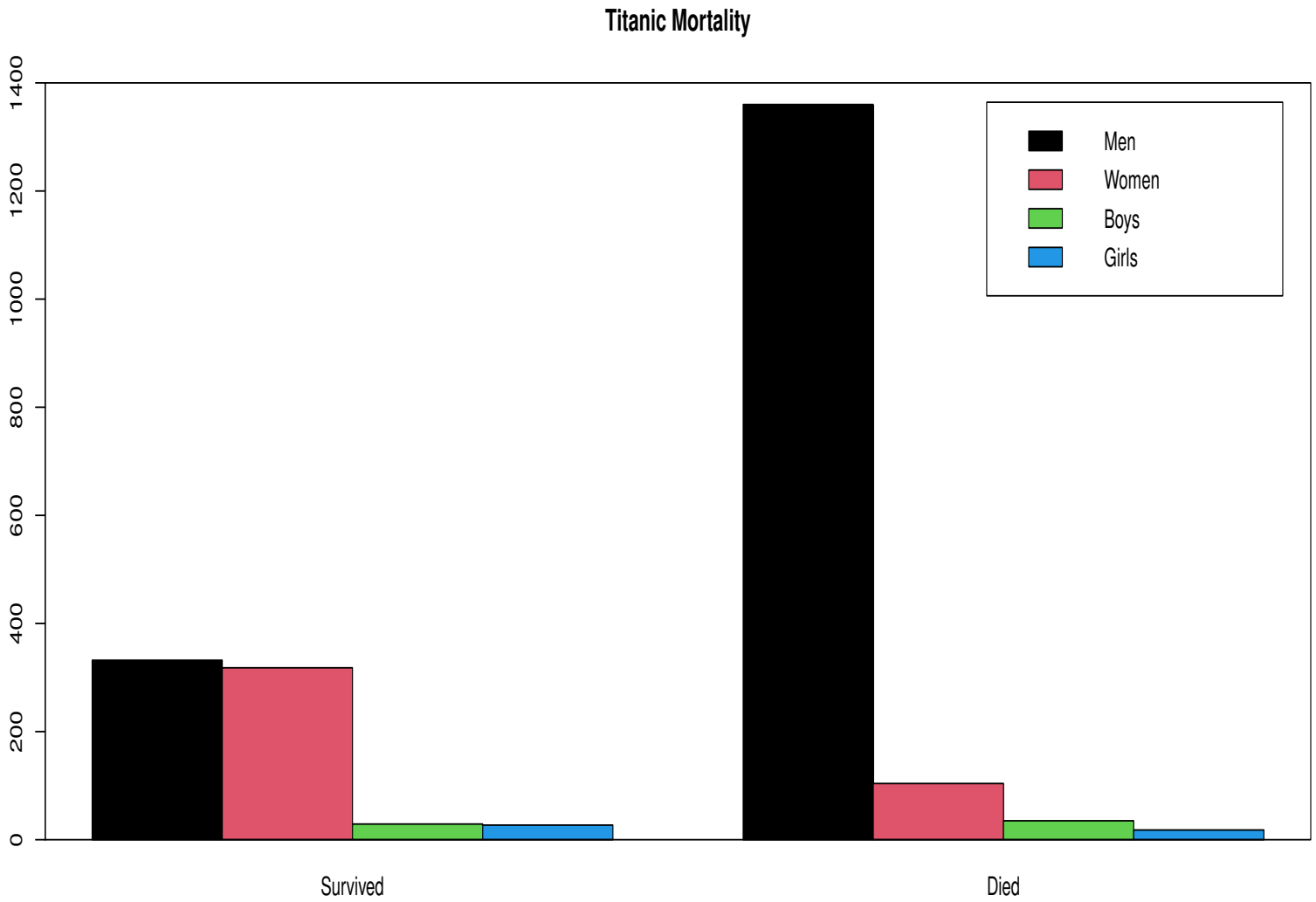
Table: Titanic Mortality

Titanic Mortality

Graphical Display:

```
Survived=c(332, 318, 29, 27)
names(Survived)=c("Men","Women","Boys", "Girls")
Died=c(1360, 104, 35, 18)
names(Died)=c("Men","Women","Boys", "Girls")
Titanic=cbind(Survived, Died)
barplot(Titanic,beside=T, legend= rownames(Titanic),
col=c(1,2,3,4), ylim=c(0,1400), main="Titanic Mortality")
box()
```

Graphical Display

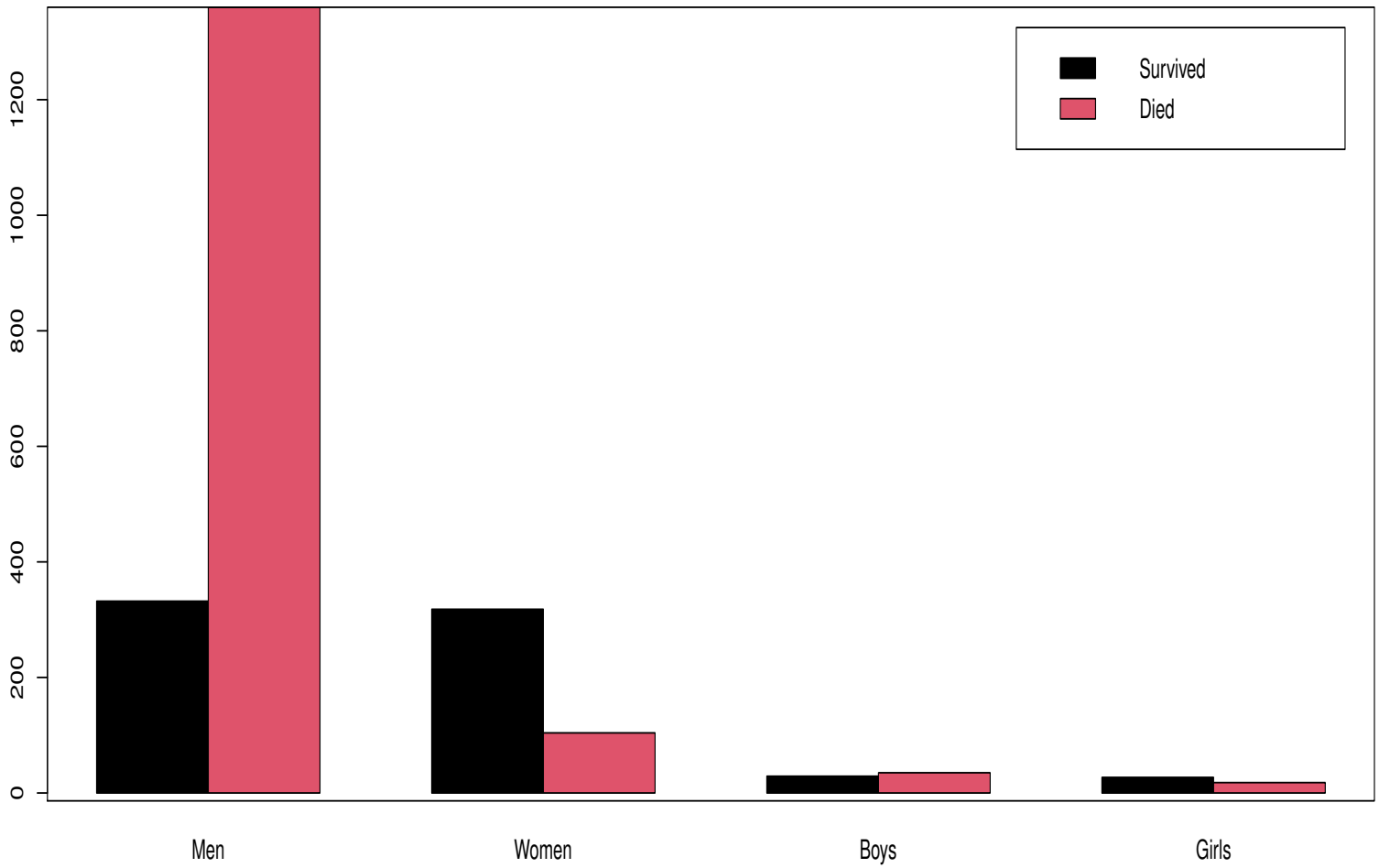


Titanic Data

	Men	Women	Boys	Girls	Total
Survived	332	318	29	27	706
Died	1360	104	35	18	1517
Total	1692	422	64	45	2223

```
Men=c(332, 1360)
Women=c(318, 104)
Boys=c(29, 35)
Girls=c(27,18)
names(Men)=c("Survived","Died")
names(Women)=c("Survived","Died")
names(Boys)=c("Survived","Died")
names(Girls)=c("Survived","Died")
Titanic=cbind(Men, Women, Boys, Girls)
barplot(Titanic,beside=T,legend= rownames(Titanic),col=c(1,2))
box()
```

Graphical Display



Chi-Square Test for Independence

This test can be used when

- the data comes from one population.
- the data have been randomly selected.
- The cells have counts or frequencies. It doesn't work if the data is percentages or relative frequencies!
- The sample data consist of frequency counts for each of the different categories.
- All expected frequency are at least 5. (The expected frequency for a category is the frequency that would occur if the data actually have the distribution that is being claimed. There is no requirement that the observed frequency for each category must be at lest 5.)

Chi-square test for Independence

This test is used to see if the row and column variables are independent. We are interested to test

$$\begin{aligned} H_0 &: \text{the two variables are independent} \\ H_a &: \text{the two variables are dependent} \end{aligned}$$

The test statistic is

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(Observed - Expected)^2}{Expected}$$

where r is the number of rows and c is the number of columns. Under the assumption of independence, and when the observations in the sufficiently large the statistic above has approximately χ^2 distribution with $(r-1)(c-1)$ degrees of freedom.

Example

Table below summarizes the fate of the passengers and crew when the Titanic sank on Monday, April 15, 1912.

	Men	Women	Boys	Girls	Total
Survived	332	318	29	27	706
Died	1360	104	35	18	1517
Total	1692	422	64	45	2223

Using a 0.05 significance level, test the claim that when the Titanic sank, whether someone survived or died is independent of whether the person is a man, woman, boy, or girl.

```
> Survived=c(332,318,29,27)
> Died=c(1360,104,35,18)
> table=rbind(Survived, Died)
> data=data.frame(table)
> data
           X1  X2 X3 X4
Survived  332 318 29 27
Died     1360 104 35 18
> chisq.test(data)
        Pearson's Chi-squared test
data:  data
X-squared = 507.0797, df = 3, p-value < 2.2e-16
```

Decision: Reject the null hypothesis. It appears that whether a person survived/died in the the Titanic and whether that person is a man, woman, boy , or a girl are dependent variables

Example

The "samhda" data set in the "UsingR "package contains information about the health behavior of school-age children. The variable "amt.smoke" measures how often child smoke in the previous month. There are seven levels: 1 means the child smoke everyday to 7 means not at all. In particular it is coded as Amount of days you smoked cigarettes in last 30. 1 = all 30, 2= 20-29, 3 = 10-19, 4 = 6-9, 5= 3-5, 6 = 1-2, 7=0. Values 98 and 99 indicates missing data. The variable gender contains the gender information of the participant. Are the two variables gender and smoking are independent? Is smoking dependent on gender? We want to test

$$H_0 \quad : \quad \text{smoking and gender are independent}$$
$$H_a \quad : \quad \text{smoking and gender are not independent}$$

Example

```
> library(UsingR)
> attach(samhda)
> table=xtabs(~gender+amt.smoke, subset=amt.smoke<98&gender!=7)
> table
      amt.smoke
gender  1  2  3  4  5  6  7
     1 16  3  5  6  7 24 64
     2 16  4  8  4  7 19 40
> chisq.test(table)
        Pearson's Chi-squared test
data:  table
X-squared = 4.1468, df = 6, p-value = 0.6568
Warning message:
In chisq.test(table):Chi-squared approximation may be incorrect
```

The warning message is due to some expected counts being small.
We can request simulate data and calculate the p- value whether the warning is serious

Example

Checking whether the warning is serious

```
> chisq.test(table,simulate.p.value=TRUE)

 Pearson's Chi-squared test with simulated p-value
                         (based on 2000 replicates)

data:  table
X-squared = 4.1468, df = NA, p-value = 0.6492
```

The p-value is not change significantly.

Example

Does happiness depends on gender? Table below summarize the GSS survey results about happiness of twenty-six year old people.

Gender	Very happy	Pretty happy	Not too happy
Male	110	277	50
Female	163	302	63

```
> HA <- c(110,277,50,163,302,63)
> HAT <- matrix(data=HA,nrow=2,byrow=TRUE)
> dimnames(HAT) <- list(Gender=c("Male","Female"),
+ Category=c("Very Happy","Pretty Happy", "Not too happy"))
> HAT
        Category
Gender   Very Happy Pretty Happy Not To Happy
  Male          110           277           50
  Female        163           302           63
> chisq.test(HAT)
        Pearson's Chi-squared test
data:  HAT
X-squared = 4.3215, df = 2, p-value = 0.1152
```

Chi-square test of homogeneity

The test of homogeneity is performed to check whether the distribution of the data in each categories are equivalent (same). This test can be used when

- The data is multinomial data in a contingency table or two way cross-classification table
- the data have been randomly selected.
- The cells have counts or frequencies. It doesn?t work if the data is percentages or relative frequencies!
- The sample data consist of frequency counts for each of the different categories.
- All expected frequency are at least 5. (The expected frequency for a category is the frequency that would occur if the data actually have the distribution that is being claimed. There is no requirement that the observed frequency for each category must be at lest 5.)
- Either the row totals or column totals are fixed.
- The data comes from multiple samples which are independent

Chi-square test of homogeneity

This test is used to see if the row and column variables are independent. We are interested to test

$$H_0 \quad : \quad \text{The distributions in all categories are the same}$$
$$H_a \quad : \quad \text{The distributions in all categories are not the same}$$

The test statistic is

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(Observed - Expected)^2}{Expected}$$

where r is the number of rows and c is the number of columns. Under the assumption of independence, and when the observations in the sufficiently large the statistic above has approximately χ^2 distribution with $(r-1)(c-1)$ degrees of freedom.

Example

In a double blind randomized drug trial, 400 male patients with mild dementia were randomly divided into two groups of 200. One group was given placebo and the other received experimental drug. Table below summarize the finding of the study after three months

Treatment	Improve	No change	Worse
Drug	67	76	57
Placebo	48	73	79

Are the proportions in the three categories the same for the two treatments?

Example

```
> DT <- c(67,76,57,48,73,79)
> DTT <- matrix(data=DT,nrow=2,byrow=TRUE)
> dimnames(DTT) <- list(Treatment=c("Drug","Placebo"),
+ Category=c("Improve","No Change", "Worse"))
> DTT
         Category
Treatment Improve No Change Worse
  Drug          67        76    57
  Placebo       48        73    79
> chisq.test(DTT)

        Pearson's Chi-squared test

data:  DTT
X-squared = 6.7584, df = 2, p-value = 0.03408
```

Decision: Reject the null hypothesis. There is sufficient evidence to suggest that not all of the probabilities for both populations with respect to each of the 3 categories are equal.

Example

Stanford University Medical Center conducted a study whether the antidepressant Celexa can help stop compulsive shopping. Twenty-four compulsive shoppers participated in the study: 12 were given placebo and 12 a dosage of Celexa daily for seven weeks. Below is the summary of the study

Treatment	much worse	worse	same	much improved	very much improved
Celexa	0	2	3	5	2
Placebo	0	2	8	2	0

Does this indicates that two samples have different distributions?
We want to test

$$H_0 \quad : \quad \text{the two distributions are the same}$$
$$H_a \quad : \quad \text{the two distributions are different}$$

Remark: It is recommended that you creta three categories namely: worse, same and better, instead of five categories.

Example

```
> celexa<-c(2,3,7)
> placebo<-c(2,8,2)
> data=rbind(celexa, placebo)
> colnames(data)=c("worse", "same", "better")
> data
        worse same better
celexa      2    3      7
placebo     2    8      2
> chisq.test(data)
        Pearson's Chi-squared test
data:  data
X-squared = 5.0505, df = 2, p-value = 0.08004
Warning message:
In chisq.test(data): Chi-squared approximation may be incorrect
```

The warning message is due to some expected counts being small.
We can request simulate data and calculate the p- value whether the warning is serious

Example

```
> celexa<-c(2,3,7)
> placebo<-c(2,8,2)
> data=rbind(celexa, placebo)
> colnames(data)=c("worse", "same", "better")
> data
        worse same better
celexa      2    3      7
placebo     2    8      2
> chisq.test(data,simulate.p.value=TRUE)

Pearson's Chi-squared test with simulated p-value
                  (based on 2000 replicates)
data:  data
X-squared = 5.0505, df = NA, p-value = 0.09695
```

In both cases, the p-value is greater than 0.05 so we fail to reject the null hypothesis and conclude that both treatment have the same distribution.

Example

The dataset "survey" in MASS package contains the responses of 237 students to a variety of questions: Consider the variables "Exer" and "Smoke".The Smoke column records the students smoking habits while the Exer column records their exercise level. Test the hypothesis whether the students smoking habit is independent of their exercise level at .05 significance level.

```
> data = data.frame(survey$Smoke,survey$Exer)
> data = table(survey$Smoke,survey$Exer)
> barplot(data, beside=T, col=c(1,2,3,4)
> legend("center", legend = rownames(data), fill=c(1,2,3,4))
> chisq.test(data)
```