28 Chi-square test for categorical data

28.1 Introduction

The models for different level of measurements of y and x are summaried below:

Y	X: Categorical	X: Continuous
Categorical	χ^2 GOF test	-
Continuous	ANOVA	Regression

In many applications, particularly in the social sciences, data are simply classified into distinct categories. For instance, the income in Australian families might be classified by income classes (categories), kidney patients might be classified by blood groups (A, AB, O), etc.

The observed frequencies in the distinct categories are known as *categorical data*. From the data, we can obtain a histogram or a cumulative frequency diagram.

Suppose we wish to compare these frequencies to some theoretical models for the frequency function, as presented by the probability distribution or probability density functions.

Example: (Computer sold) The number of computers sold at a store in four consecutive quarterly periods were recorded as below:

i	1	2	...	111	...	168	...	221	...	Total count
A	1	1	...	0	...	0	...	0	...	110
B	0	0	...	1	...	0	...	0	...	57
C	0	0	...	0	...	1	...	0	...	53
D	0	0	...	0	...	0	...	1	...	80
y_i	A	A	...	B	...	C	...	D	...	300

Quarter	Jan-Mar (A)	Apr-June (B)	July-Sept (C)	Oct-Dec (D)	Total
No. sold	110	57	53	80	300

The store claims that twice as many computers are sold in the Jan-Mar quarter as are sold in any one of the other quarters. Is there evidence to support the stated conjecture?

Solution: Let p_i be the expected proportion of the computers sold in the i-th quarter. If the stated conjecture is right, then

$$p_1 = 2/5, \quad p_2 = 1/5, \quad p_3 = 1/5, \quad p_4 = 1/5.$$

which is a theoretical model (a probability distribution) of the stated conjecture. Note that $p_1 + p_2 + p_3 + p_4 = 1$.

To examine whether the hypothesized model appears to be a good fit to the observations, we need to compare the observed numbers O_i:

Comp. Sold (O_i): 110, 57, 53, 80,

with the expected number E_i:

Expet. Sold (E_i): 120, 60, 60, 60.

We have

$$O_i - E_i : \quad -10, \quad -3, \quad -7, \quad 20.$$

Note that the larger the expected number, the higher the variability of the observed number. Hence it is better to examine the standardized residuals $r_i = (O_i - E_i)/\sqrt{E_i}$:

$$r_i : \quad -10/\sqrt{120}, \quad -3/\sqrt{60}, \quad -7/\sqrt{60}, \quad 20/\sqrt{60}.$$

and these standardized residuals r_i behave rather like standard normal r.v. provided that the theoretical model is correct.

Generally, we say that the fit is good if the residuals r_i are between -2 and 2. In this example, the hypothesized model is not a good fit as $r_4 = 20/\sqrt{60} = 2.582 > 2$.

In general, suppose we have k observed frequencies $y_1, y_2, ..., y_k$. We say that a model (a probability distribution):

$$p_1 = p_{10},\ p_2 = p_{20},\ \cdots,\ p_k = p_{k0}, \tag{2}$$

where $p_{i0} > 0$ and $\sum_{i=1}^{k} p_{i0} = 1$, is a good fit to the observations y_i if the standardized residuals r_i as given by

$$r_i = (y_i - np_{i0})/\sqrt{np_{i0}}, \quad i = 1, 2, ..., k,$$

where $n = \sum_{i=1}^{k} y_i$, are between -2 and 2.

28.2 Theoretical explanation

There is a theoretical explanation for above statement.

Let

$$X_{ij} = \begin{cases} 1, & \text{if the } j\text{-th observation falls in the } i\text{-th category,} \\ 0, & \text{otherwise.} \end{cases}$$

Then y_i is a observed value of $S_i = \sum_{j=1}^{n} X_{ij}$. Note that, under the hypothesized model,

$$(S_i - np_{i0})/\sqrt{np_{i0}} \to_D \mathcal{N}(0, 1), \quad \text{as } n \to \infty$$

where '$\to_D$' denotes 'converge in distribution'. Therefore,

$$\Pr(|(S_i - np_{i0})/\sqrt{np_{i0}}| \geq 2) \leq 0.05$$

approximately.

This implies that the requirements in above statements are reasonable.

28.3 Chi square goodness-of-fit test (P.499-503,506-519)

With k observed frequencies $y_1, y_2, ..., y_k$, we may construct a test for the hypothesis:

$$H_0: \ p_1 = p_{10},\ p_2 = p_{20},\ \cdots,\ p_k = p_{k0}$$

where $p_{i0} > 0$ and $\sum_{i=1}^{k} p_{i0} = 1$. If we accept the null hypothesis, the proposed model $p_i = p_{i0}$ is a good fit to the observations.

With the observed frequencies y_i in k categories, we have $n = \sum_{i=1}^{k} y_i$ observations altogether, and the expected category frequencies are np_{i0} in category i under the null hypothesis H_0.

In terms of that the observed and expected category frequencies, we should reject the H_0 if

$$\chi_0^2 = \sum_{i=1}^{k} \frac{(y_i - np_{i0})^2}{np_{i0}} = \sum_{i=1}^{k} \frac{y_i^2}{np_{i0}} - n$$

is large. Note that

$$\sum_{i=1}^{k} \frac{(y_i - np_{i0})^2}{np_{i0}} = \sum_{i=1}^{k} \frac{y_i^2 - 2np_{i0}y_i + n^2p_{i0}^2}{np_{i0}} = \sum_{i=1}^{k} \frac{y_i^2}{np_{i0}} - 2\sum_{i=1}^{k} y_i + n\sum_{i=1}^{k} p_{i0} = \sum_{i=1}^{k} \frac{y_i^2}{np_{i0}} - n.$$

In particular, if $p_{i0} = 1/k$, then

$$\chi_0^2 = \frac{k}{n}\sum_{i=1}^{k} y_i^2 - n.$$

Note:

1. The test statistic $\chi_0^2 \sim \chi_{k-1-p}^2$ where p is the number of parameters estimated from the sample.

2. The df from the sample is $k-1$ because the first $k-1$ observations y_i contain all the information and the last observation is fixed by $y_k = n - \sum_{i=1}^{k-1} y_i$ adding no extra information.
3. The approximation will only be accurate if *no expected frequency* is too small (<5). Otherwise, we will pool adjacent categories so that the expected frequencies are always ≥ 5.

The five steps of the Chi-square goodness-of-fit test are:

1. **Hypotheses:** $H_0: \ p_1 = p_{10}, \ p_2 = p_{20}, \ \cdots, \ p_k = p_{k0}$ vs H_1 : at least one equality does not hold.
2. **Test statistic:** $\chi_0^2 = \sum_{i=1}^{k} \frac{(y_i - np_{i0})^2}{np_{i0}}$.
3. **Assumption:** $E_i = np_{i0} \geq 5$. Under H_0, $\chi_0^2 \sim \chi^2_{k-1-p}$ approx.
4. ***P*-value:** $\Pr(\chi^2_{k-1} \geq \chi_0)$.
5. **Conclusion:** Reject H_0 if the p-value$< \alpha$.

Calculations are summarized in the following table.

Class i	$O_i = y_i$	p_{i0}	$E_i = np_{i0}$	$\chi_{i0}^2 = (y_i - np_{i0})^2/(np_{i0})$
1	y_1	p_{10}	np_{10}	$(y_1 - np_{10})^2/(np_{10})$
2	y_2	p_{20}	np_{20}	$(y_2 - np_{20})^2/(np_{20})$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
k	y_k	p_{k0}	np_{k0}	$(y_k - np_{k0})^2/(np_{k0})$
Total	n	1	n	χ_0^2

Notes: The χ_0^2 test statistic also can be used to test whether the sample data fit a particular model for a population distribution.

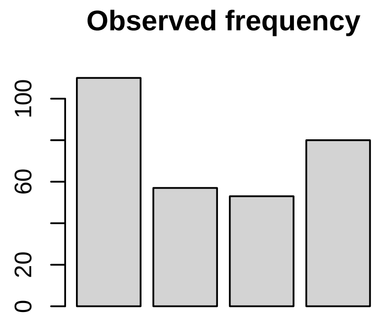
Example: (Computer sold)

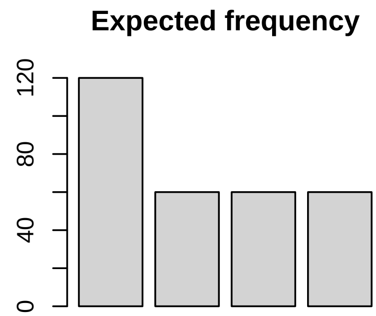
Solution: We use the following table to calculate the test statistic:

Quarter i	O_i	$E_i = np_{i0}$	$O_i - E_i$	$\frac{(O_i-E_i)^2}{E_i}$
Jan-Mar	110	2/5(300)=120	-10	$\frac{10^2}{120} = \frac{5}{6}$
Apr-June	57	1/5(300)=60	-3	$\frac{3^2}{60} = \frac{3}{20}$
July-Sept	53	1/5(300)=60	-7	$\frac{7^2}{60} = \frac{49}{60}$
Oct-Dec	80	1/5(300)=60	20	$\frac{20^2}{60} = \frac{20}{3}$
Total	300	300	0	8.47

The Chi-square goodness-of-fit test is

1. **Hypotheses:** $H_0: \ p_1 = 2/5, \ p_2 = 1/5, \ p_3 = 1/5, \ p_4 = 1/5$
   vs H_1 : at least one equality does not hold.
2. **Test statistic:** $\chi_0^2 = \sum_{i=1}^{k} \frac{(y_i - np_{0i})^2}{np_{0i}} = 8.47.$
3. **Assumption:** $E_i = np_{i0} \geq 5$. Under H_0, $\chi_0^2 \sim \chi_{k-1}^2$.
4. ***P*-value:** $0.025 < p\text{-value} \approx \Pr(\chi_{4-1}^2 \geq 8.47) < 0.05$
   $(\chi_{3,0.95}^2 = 7.815, \ \chi_{3,0.975}^2 = 9.348, \ 0.03723$ from R).
5. **Decision:** Since the p-value< 0.05, there is evidence in the data against the stated claim that twice as many computers are sold in the Jan-Mar quarter as are sold in any one of the other quarters.





In R,

```
> y=c(110,57,53,80)
> p=c(2/5,1/5,1/5,1/5)
> chisq.test(y,p=p)

        Chi-squared test for given probabilities

data:  y
X-squared = 8.4667, df = 3, p-value = 0.03729

> n=sum(y)   #checking
> k=length(y)  #no. of class
> ey=n*p
> ey
[1] 120  60  60  60
> ey>=5  #test Ei>=5
[1] TRUE TRUE TRUE TRUE
> chi2=sum((y-ey)^2/ey)
> chi2
[1] 8.466667
> p.value=1-pchisq(chi2,k-1)
> p.value
[1] 0.03729023
```

Example: (Genetic linkage) In a backcross experiment to investigate the genetic linkage between two factors A and B in a species of flower, some researchers classified 400 offspring by phenotype as follows:

AB	Ab	aB	ab
128	86	74	112

(a) Under the 'no linkage' model, the four phenotypes are equally likely. Show that this model is a poor fit.

(b) If linkage is in the 'coupling phase', the probabilities of the four phenotypes are

AB	Ab	aB	ab
$\frac{1}{2}(1-p)$	$\frac{1}{2}p$	$\frac{1}{2}p$	$\frac{1}{2}(1-p)$

where p is the 'recombination fraction' and is estimated by the overall proportion of Ab and aB. Show that this model fits the data well.

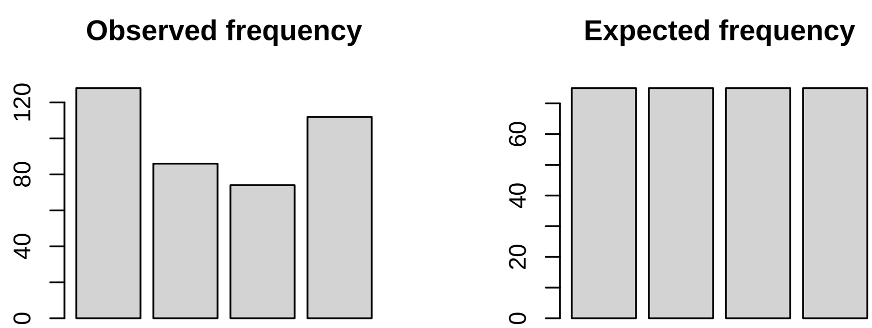
Solution:

(a) Under the 'no linkage' model, we complete the following table:

Type	O_i	$E_i = np_i$	$O_i - E_i$	$\frac{(O_i-E_i)^2}{E_i}$
AB	128	$400 \times \frac{1}{4} = 100$	128 - 100 = 28	$\frac{(28)^2}{100} = 7.84$
Ab	86	$400 \times \frac{1}{4} = 100$	86 - 100 = -14	$\frac{(-14)^2}{100} = 1.96$
aB	74	$400 \times \frac{1}{4} = 100$	74 - 100 = -26	$\frac{(-26)^2}{100} = 6.76$
ab	112	$400 \times \frac{1}{4} = 100$	112 - 100 = 12	$\frac{(12)^2}{100} = 1.44$
Total	400	400	0	$\chi_0^2 = 18.00$

Then the *Chi-square test* for the *proportions of phenotypes* is

1. **Hypotheses:** $H_0 : p_i = \frac{1}{4}$
   vs H_1 : at least one equality does not hold.
2. **Test statistic:** $\chi_0^2 = \sum_{i=1}^{k} \frac{(y_i - np_{0i})^2}{np_{0i}} = 18.$
3. **Assumption:** $E_i = np_{i0} \geq 5$. Under H_0, $\chi_0^2 \sim \chi_{k-1}^2$.
4. ***P*-value:** $\Pr(\chi_{4-1}^2 > 18) < 0.01$ ($\chi_{3,0.99}^2 = 9.21$, (0.0004 from R)).
5. **Decision:** Since the p-value is < 0.05, we reject H_0 and conclude that there is strong evidence in the data against H_0 that the four phenotypes are equally likely.



In R,

```
> y=c(128,86,74,112)
> p=c(1/4,1/4,1/4,1/4)
> chisq.test(y,p=p)

        Chi-squared test for given probabilities

data:  y
X-squared = 18, df = 3, p-value = 0.0004398
```

```
> n=sum(y)   #checking
> k=length(y)
> ey=n*p
> ey
[1] 100 100 100 100
> ey>=5  #test Ei>=5
[1] TRUE TRUE TRUE TRUE
> chi2=(y-ey)^2/ey
> chi2
[1] 7.84 1.96 6.76 1.44
> chi2=sum(chi2)
> chi2
[1] 18
> p.value=1-pchisq(chi2,k-1)
> p.value
[1] 0.0004398497
```

(b) Under the 'coupling phase' linkage model, we estimate the probability p by the sample proportion

$$\hat{p} = \frac{86 + 74}{400} = 0.4.$$

Hence the four probabilities are

$\frac{1}{2}(1 - 0.4) = 0.3$, $\frac{1}{2}0.4 = 0.2$, $\frac{1}{2}0.4 = 0.2$ and $\frac{1}{2}(1 - 0.4) = 0.3$.

Then we complete the following table:

Type	O_i	$E_i = n\hat{p}_i$	$O_i - E_i$	$\frac{(O_i - E_i)^2}{E_i}$
AB	128	$400 \times \frac{3}{10} = 120$	128- 120 = 8	$\frac{(8)^2}{120} = 0.533$
Ab	86	$400 \times \frac{2}{10} = 80$	86 - 80 = 6	$\frac{(6)^2}{80} = 0.450$
aB	74	$400 \times \frac{2}{10} = 80$	74 - 80 = -6	$\frac{(-6)^2}{80} = 0.450$
ab	112	$400 \times \frac{3}{10} = 120$	112- 120 = -8	$\frac{(-8)^2}{120} = 0.533$
Total	400	400	0	$\chi_0^2 = 1.967$

The *Chi-square test* for the *proportions of phenotypes* is:

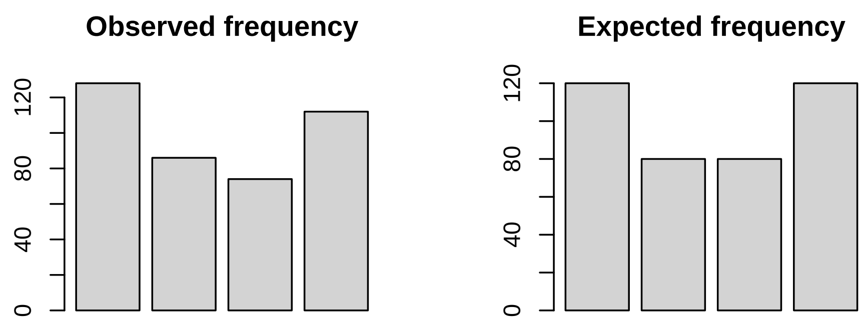
1. **Hypotheses:** $H_0 : p_1 = 0.3,\ p_2 = 0.2,\ p_3 = 0.2,\ p_4 = 0.3$
   vs H_1 : at least one equality does not hold.

2. **Test statistic:** $X^2 = \sum_{i=1}^{g} \frac{(y_i - n\hat{p}_i)^2}{n\hat{p}_i} = 1.967.$

3. **Assumption:** $E_i = n\hat{p}_{i0} \geq 5$. Under H_0, $\chi_0^2 \sim \chi^2_{k-1-1}$.

4. ***P*-value:** $\Pr(\chi^2_{4-1-1} > 1.967) > 0.1$ ($\chi_{2,0.9} = 4.605$, 0.3741 (from R)).

5. **Decision:** Since the p-value is > 0.05, we accept H_0 and conclude that the data is consistent with the 'coupling phase' linkage model.



In R,

```
> par=(y[2]+y[3])/n
> p=c((1-par)/2,par/2,par/2,(1-par)/2)
> p
[1] 0.3 0.2 0.2 0.3
> chisq.test(y,p=p)  #ignore df & p-value; incorrect

        Chi-squared test for given probabilities

data:  y
X-squared = 1.9667, df = 3, p-value = 0.5794

> ey=n*p   #checking
> ey
[1] 120  80  80 120
> ey>=5  #test Ei>=5
[1] TRUE TRUE TRUE TRUE
> chi2=(y-ey)^2/ey
> chi2
```

```
[1] 0.5333333 0.4500000 0.4500000 0.5333333
> chi2=sum(chi2)
> chi2
[1] 1.966667
> p.value=1-pchisq(chi2,k-1-1)
> p.value
[1] 0.3740621
```

Note that the p-value of **`chisq.test`** is *incorrect* as the deg. of freedom is not subtracted by one when a parameter is estimated from the data.

29 Chi-square test for discrete distribution

Suppose we have a sample $x_1, x_2, ..., x_n$. We want to test whether the sample is taken from a population with a given distribution function $F_0(x|\theta_1, \theta_2, ..., \theta_h)$ where θ_l are parameters of the distribution.

We may count the frequencies y_i for each value of x_j and calculate expected probabilities p_i using $F_0(x|\theta_1, \theta_2, ..., \theta_h)$. This is a *general* Chi-square goodness-of-fit test.

However the model parameters $\theta_1, \theta_2, ..., \theta_h$ are usually unknown and have to be estimated from the sample.

In this case, the expected probabilities p_i are replaced by their estimates $\hat{p}_i$. Then the test statistic is

$$\chi_0^2 = \sum_{i=1}^{k} \frac{(y_i - n\hat{p}_i)^2}{n\hat{p}_i},$$

and the p-value is

$$p\text{-value} = \Pr(\chi^2_{k-h-1} \geq \chi_0^2).$$

29.1 Poisson distribution

Example: (Suicides) The number of suicides Y per month was checked over a 5 year period, with results shown as follow:

y	0	1	2	3 or more
Freq.	33	17	7	3

We want to test whether the random variable Y has a Poisson distribution.

Solution: Since λ is unknown, we estimate λ by the sample mean $\hat{\lambda} = \bar{x} = 40/60 = 2/3$.

The calculations are summarized in the following table:

x	Obs. f. y_i	Prod. $x\ y_i$	Exp. prob. $\hat{p}_i = \hat{\lambda}^x e^{-\hat{\lambda}}/x!$	Exp. f. $n\hat{p}_i$	Chi-square $\frac{(y_i - n\hat{p}_i)^2}{n\hat{p}_i}$
0	33	0	$\frac{0.667^0 e^{-0.667}}{0!} = 0.5134$	$60(0.5134) = 30.81$	$\frac{(2.195)^2}{30.81} = 0.156$
1	17	17	$\frac{0.667^1 e^{-0.667}}{1!} = 0.3423$	$60(0.3423) = 20.54$	$\frac{(-3.537)^2}{20.54} = 0.609$
2	7	14	$\frac{0.667^2 e^{-0.667}}{2!} = 0.1141$	$60(0.1141) = 6.85$	$\frac{(0.154)^2}{6.85} = 0.003$
≥ 3	3	9	1-0.513 -0.342 -0.114=0.0302	$60(0.0302) = 1.81$	$\frac{(1.187)^2}{1.81} = 0.778$
Sum	60	40	1.0000	60.00	1.547

Note that $E_{\geq 3} = 1.8 < 5$ which violates the assumption. We combine the last two classes so that

$$O_{\geq 2} = 7{+}3 = 10,\ E_{\geq 2} = 6.85{+}1.81 = 8.66,\ \chi^2_{\geq 2} = \frac{(10 - 8.66)^2}{8.66} = 0.207$$

and $\chi^2_0 = 0.156 + 0.609 + 0.207 = 0.972.$

The Chi-square goodness-of-fit test for the Poisson distribution is

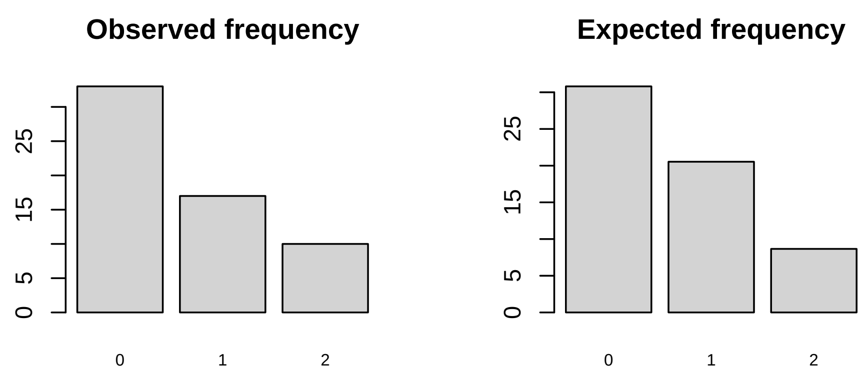
1. **Hypotheses:** H_0 : The data follow a Poisson dist.
   vs H_1 : The data do not follow a Poisson dist.

2. **Test statistic:** $\chi_0^2 = \sum_{i=1}^{k} \frac{(y_i - n\hat{p}_i)^2}{n\hat{p}_i} = 0.972.$

3. **Assumption:** $E_i = n\hat{p}_{i0} \geq 5$. Under H_0, $\chi_0^2 \sim \chi^2_{k-1-h}$ where h is the number of estimated parameters.

4. ***P*-value:** $\Pr(\chi^2_{3-1-1} \geq 0.972) > 0.1$ ($\chi^2_{1,0.90} = 2.706$; 0.4615 from R).

5. **Decision:** Since the p-value > 0.05, we accept H_0. The data are consistent with the claim that the random variable Y has a Poisson distribution.



In R,

```
> y=c(33,17,7,3)
> x=c(0,1,2,3)
> n=sum(y)
> k=length(y)
> lam=sum(y*x)/n
> lam
[1] 0.6666667
> p=lam^x*exp(-1*lam)/factorial(x)
> p[4]=1-sum(p[1:3])
```

```
> p
[1] 0.51341712 0.34227808 0.11409269 0.03021211
> ey=n*p
> ey
[1] 30.805027 20.536685  6.845562  1.812727
> ey>=5  #Ei>=5 not all satisfied
[1] TRUE TRUE TRUE FALSE
> yr=c(y[1:2],y[3]+y[4])
> yr
[1] 33 17 10
> eyr=c(ey[1:2],ey[3]+ey[4])
> eyr
[1] 30.805027 20.536685  8.658288
> pr=c(p[1:2],p[3]+p[4])
> pr
[1] 0.5134171 0.3422781 0.1443048
> kr=length(yr)
> xr=x[1:kr]
> xr
[1] 0 1 2
> chi2=(yr-eyr)^2/eyr
> chi2
[1] 0.1564000 0.6090632 0.2079153
> chi2=sum(chi2)
> chi2
[1] 0.9733785
> p.value=1-pchisq(chi2,kr-1-1)
> p.value
[1] 0.323839
> chisq.test(yr,p=pr)  #ignore df & p-value; incorrect
```

```
        Chi-squared test for given probabilities

data:  yr
X-squared = 0.9734, df = 2, p-value = 0.6147

> par(mfrow=c(2,2))
> barplot(yr,names.arg=xr,col="lightgray",
  main="Observed frequency")
> barplot(eyr,names.arg=xr,col="lightgray",
  main="Expected frequency")
```

29.2 Binomial distribution

Example: (Sales volumes) A salesperson makes five calls per day. A sample of 200 days gives the frequencies of sales volumes Y listed below:

Number of sales x	0	1	2	3	4	5
Observed frequency (days) y_i	10	38	69	63	18	2

Test if Y follows a binomial distribution.

Solution: The probability p_0 in the binomial distribution is estimated to be

$$\hat{p}_0 = \frac{447}{200(5)} = 0.447.$$

The calculations are summarized in the following table:

Sales x	Obs. f. y_i	Prod. $x \times y_i$	Exp. prob. $\hat{p}_i = {}_5C_i\hat{p}_0^i(1-\hat{p}_0)^{5-i}$	Exp. f. $E_i = n\hat{p}_i$	Chi-square $\frac{(y_i - n\hat{p}_i)^2}{n\hat{p}_i}$
0	10	0	$C_5^0 0.45^0 0.553^5 = 0.0517$	$200(0.0517) = 10.34$	$\frac{(-0.343)^2}{10.34} = 0.011$
1	38	38	$C_5^1 0.45^1 0.553^4 = 0.2090$	$200(0.2090) = 41.80$	$\frac{(-3.803)^2}{41.80} = 0.346$
2	69	138	$C_5^2 0.45^2 0.553^3 = 0.3379$	$200(0.3379) = 67.58$	$\frac{(1.420)^2}{67.58} = 0.030$
3	63	189	$C_5^3 0.45^3 0.553^2 = 0.2731$	$200(0.2731) = 54.63$	$\frac{(8.374)^2}{54.63} = 1.284$
4	18	72	$C_5^4 0.45^4 0.553^1 = 0.1104$	$200(0.1104) = 22.08$	$\frac{(-4.078)^2}{22.08} = 0.753$
5	2	10	$C_5^5 0.45^5 0.553^0 = 0.0178$	$200(0.0178) = 3.57$	$\frac{(-1.569)^2}{3.57} = 0.690$
Sum	200	447	1.0000	200.00	3.114

Note that $E_6 = 3.57 < 5$ which violates the assumption. We combine the last two classes so that

$$O_{\geq 4} = 18{+}2 = 20,\ E_{\geq 4} = 22.08{+}3.57 = 25.65,\ \chi^2_{\geq 4} = \frac{(20-25.65)^2}{25.65} = 1.245$$

and $\chi_0^2 = 0.011 + 0.346 + 0.030 + 1.284 + 1.245 = 2.916.$

The Chi-square goodness-of-fit test for binomial distribution is

1. **Hypothesis:** H_0 : The data follow a binomial dist.
   vs H_1: The data do not follow a binomial dist.
2. **Test statistic:** $\chi_0^2 = \sum_i \frac{(y_i - n\hat{p}_i)^2}{n\hat{p}_i} = 2.916$
3. **Assumption:** $E_i = n\hat{p}_i \geq 5$ and $\chi_0^2 \sim \chi_{k-1-h}^2$ where h is the number of parameters estimated.
4. ***P*-value:** $\Pr(\chi_3^2 > 2.916) > 0.1$ ($\chi_{3,0.9}^2 = 6.251,\ 0.405$ from R).
5. **Conclusion:** Since the p-value > 0.05, we accept H_0. The data are consistent with H_0 that the data follow a binomial distribution.

In R,

```
> y=c(10,38,69,63,18,2)
> x=c(0,1,2,3,4,5)
> n=sum(y)
> m=max(x)
> k=length(y)
> prob=sum(y*x)/(m*n)
> prob
[1] 0.447
> p=dbinom(x,m,prob)
> p
[1] 0.05171609 0.20901529 0.33790175 0.27313216 0.11038885
    0.01784587
> ey=n*p
> ey
```

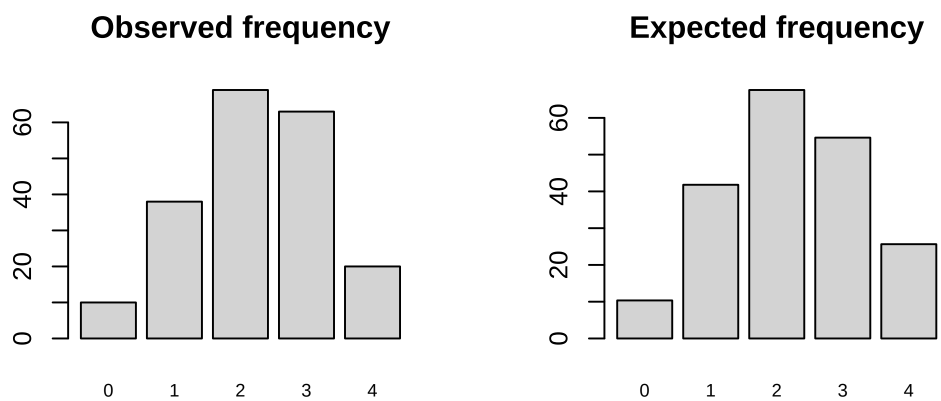
```
[1] 10.343217 41.803058 67.580350 54.626431 22.077771  3.56917
> ey>=5  #Ei>=5 not all satisfied
[1] TRUE TRUE TRUE TRUE TRUE FALSE
> yr=c(y[1:4],y[5]+y[6])
> yr
[1] 10 38 69 63 20
> eyr=c(ey[1:4],ey[5]+ey[6])
> eyr
[1] 10.34322 41.80306 67.58035 54.62643 25.64694
> pr=c(p[1:4],p[5]+p[6])
> pr
[1] 0.05171609 0.20901529 0.33790175 0.27313216 0.12823472
> kr=length(yr)
> xr=x[1:kr]
> xr
[1] 0 1 2 3 4
> chi2=(yr-eyr)^2/eyr
> chi2
[1] 0.01138893 0.34598539 0.02982237 1.28356649 1.24334413
> chi2=sum(chi2)
> chi2
[1] 2.914107
> p.value=1-pchisq(chi2,kr-1-1)
> p.value
[1] 0.4050586
> chisq.test(yr,p=pr)  #ignore df & p-value; incorrect

        Chi-squared test for given probabilities

data:  yr
```

```
X-squared = 2.9141, df = 4, p-value = 0.5723

> par(mfrow=c(2,2))
> barplot(yr,names.arg=xr,col="lightgray",
  main="Observed frequency")
> barplot(eyr,names.arg=xr,col="lightgray",
  main="Exected frequency")
```



30 Chi-square test for continuous distribution

30.1 Test procedures

With a given data set, the observed frequencies y_i and the expected frequencies np_i can be calculated in the following steps:

Step 1: Divide the x-axis into k intervals $I_1, I_2, ..., I_k$ such that for each interval, the expected frequency $E_i = n\hat{p}_i$ is at least 5. Determine the frequencies y_i of sample values x_j in the intervals I_i.

Step 2: Using $F_0(x)$, compute the probability p_i of the population falling in the I_i. Then np_i is the number of sample values theoretically expected in I_i if the hypothesis is true.

Hence the Chi-square test for a continuous distribution is:

1. **Hypotheses:** $H_0: \; F(x) = F_0(x) \quad \text{vs} \quad H_1: F(x) \neq F_0(x)$.
2. **Test statistic:** $\chi_0^2 = \sum_{i=1}^{k} \frac{(y_i - n\hat{p}_i)^2}{n\hat{p}_i}$.
3. **Assumption:** $E_i = n\hat{p}_i \geq 5$. Under H_0, $\chi_0^2 \sim \chi_{k-1-h}^2$ approx.
4. **P-value:** $\Pr(\chi_{k-1-h}^2 \geq \chi_0^2)$
5. **Decision:** Reject H_0 if the p-value $< \alpha$.

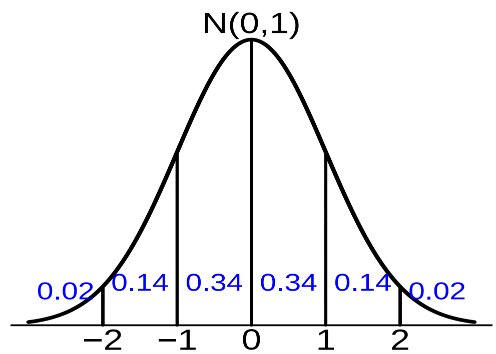
Note: h is the number of parameters in the distribution which are estimated from the data. This test can be used to test for the *normality* assumption of the residuals.

30.2 Number of intervals for normal distribution

In practice, one can use any number of intervals. The number of intervals should be chosen to comply with the rule of *five*. Moreover because the true mean μ and variance σ^2 are to be estimated from the data, the degree of freedom is $k-3$. Hence k has to be at least 4 and $n \geq 32$.

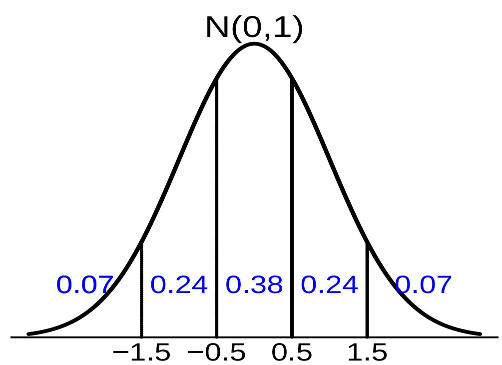
1. When $n \geq 220$ (note: $220(0.0228) = 5.016 > 5$),

Interval	Probability
$Z \leq -2$	0.0228
$-2 < Z \leq -1$	0.1359
$-1 < Z \leq 0$	0.3413
$0 < Z \leq 1$	0.3413
$1 < Z \leq 2$	0.1359
$Z > 2$	0.0228



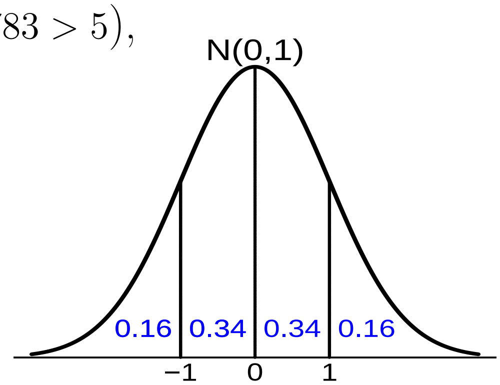
2. When $80 \leq n < 220$ (note: $80(0.0668) = 5.344 > 5$),

Interval	Probability
$Z \leq -1.5$	0.0668
$-1.5 < Z \leq -0.5$	0.2417
$-0.5 < Z \leq 0.5$	0.3829
$0.5 < Z \leq 1.5$	0.2417
$Z > 1.5$	0.0668



3. When $32 \leq n < 80$ (note: $32(0.1587) = 5.0783 > 5$),

Interval	Probability
$Z \leq -1$	0.1587
$-1 < Z \leq 0$	0.3413
$0 < Z \leq 1$	0.3413
$Z > 1$	0.1587

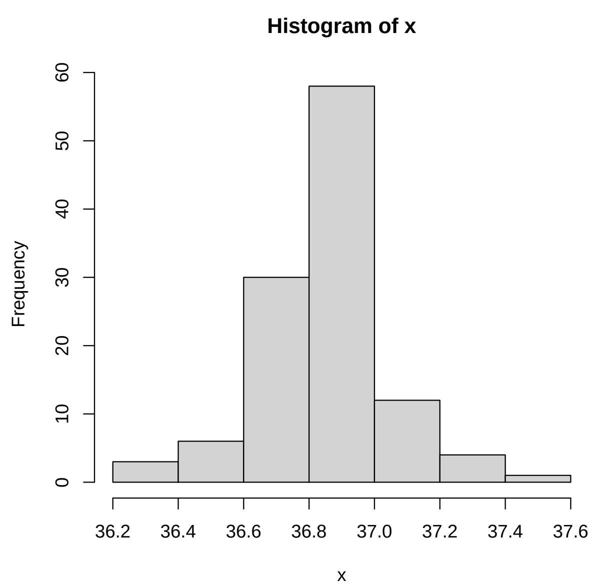


Example: The data set `beav1$temp` contains measurements of body temperature (in degree Celsius) over a certain time period.

```
> beav1=read.csv("data/beav1.csv")
> attach(beav1)
> x=sort(beav1$temp)
```

```
> x
  [1] 36.33 36.34 36.35 36.42 36.50 36.54 36.55 36.55| 36.59 36.62 36.62 36.64
 [13] 36.65 36.67 36.67 36.69 36.69 36.69 36.70 36.71 36.71 36.72 36.73 36.74
 [25] 36.75 36.75 36.75 36.75 36.76 36.76 36.77 36.77| 36.78 36.78 36.79 36.79
 [37] 36.80 36.80 36.80 36.81 36.81 36.82 36.82 36.82 36.83 36.83 36.84 36.84
 [49] 36.85 36.85 36.85 36.85 36.86 36.86 36.87 36.87 36.87 36.87 36.88 36.88
 [61] 36.88 36.88 36.89 36.89 36.89 36.89 36.89 36.89 36.89 36.91 36.91 36.91
 [73] 36.92 36.92 36.92 36.93 36.93 36.93 36.93 36.94 36.94 36.94 36.94 36.95
 [85] 36.95 36.96| 36.97 36.97 36.97 36.98 36.98 36.99 36.99 36.99 37.00 37.00
 [97] 37.00 37.01 37.02 37.05 37.07 37.09 37.10 37.10 37.15| 37.18 37.20 37.20
[109] 37.20 37.21 37.23 37.24 37.25 37.53
```

```
> hist(x,col="lightgray")
```



Is there evidence in the data that the observation is not from a normal population?

Solution Let X be the random variable.

Note that the mean μ and variance σ^2 are unknown and are estimated by the sample mean $\bar{x} = 36.8622$ and the sample variance $s^2 = 0.1934^2$ respectively. Since $n = 114$, we choose $k = 5$ intervals.

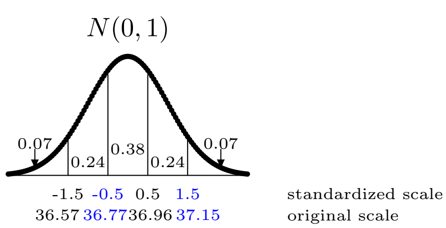
Calculation of the test statistic is summarized in the following table:

Interval i	Stand. int. i	O.f. y_i	Exp. prob. p_i	Exp.f. np_i	Chi-square $\frac{(y_i-np_i)^2}{np_i}$
$X \leq 36.57$	$Z \leq$-1.5	8	Φ(-1.5)=.067	114(.067)= 7.6	$\frac{(8-7.6)^2}{7.6} = 0.02$
$36.57 < X \leq 36.77$	-1.5< $Z \leq$-0.5	22	Φ(-.5)-Φ(-1.5)=.242	114(.242)=27.6	$\frac{(22-27.6)^2}{27.6} = 1.12$
$36.77 < X \leq 36.96$	-0.5< $Z \leq$0.5	55	Φ(.5)-Φ(-.5)=.383	114(.383)=43.7	$\frac{(55-43.7)^2}{43.7} = 2.95$
$36.96 < X \leq 37.15$	0.5< $Z \leq$1.5	20	Φ(1.5)-Φ(.5)=.242	114(.242)=27.6	$\frac{(20-27.6)^2}{27.6} = 2.07$
$X > 37.15$	$Z >$1.5	9	1-Φ(1.5)=.067	114(.067)= 7.6	$\frac{(9-7.6)^2}{7.6} = 0.25$
Total		114	1.000	114.0	6.41

Note: the class marks for **x** within which y_i are counted are

$\min(x) = 36.33$, $\bar{x}-1.5s = 36.86-1.5(0.19) = 36.57$, $\bar{x}-0.5s = 36.86-0.5(0.19) = 36.77$,

$\max(x) = 37.53$, $\bar{x}+0.5s = 36.86+0.5(0.19) = 36.96$, $\bar{x}+1.5s = 36.86+1.5(0.19) = 37.15$.



The Chi-square goodness-of-fit test for a normal distribution is

1. **Hypotheses:** $H_0: \ X \sim \mathcal{N}(\mu, \sigma)$ vs $H_1: \ X \not\sim \mathcal{N}(\mu, \sigma^2)$.
2. **Test statistic:** $\chi_0^2 = \sum_{i=1}^{k} \frac{(y_i - n\hat{p}_i)^2}{n\hat{p}_i} = 6.41.$
3. **Assumption:** $E_i = n\hat{p}_{i0} \geq 5$. Under H_0, $\chi_0^2 \sim \chi^2_{k-1-h}$ approx.

4. **P-value:** $\Pr(\chi_2^2 \geq 6.41) \in (0.025, 0.05)$

$$(\chi^2_{2,0.95} = 5.991,\ \chi^2_{2,0.975} = 7.378,\ 0.04049 \text{ from R})$$

5. **Decision:** Since the p-value < 0.05 H_0, we reject H_0. There is sufficient evidence in the data against the null hypothesis that the data follow a normal distribution.

In R,

Step 1. Divide the range of `x` into 5 intervals according to $n = 114$. Estimate μ and σ^2 by the sample mean and sample s.d..

```
> n=length(x)
> k=6
> if (n<80) k=4 else if (n<220) k=5  #find the no. of class, k
> k
[1] 5
> xm=mean(x)
> xsd=sd(x)
> c(xm,xsd)
[1] 36.8621930  0.1934217
> zlow=-2   #set lowest to be 1st class mark
> if (n<80) zlow=-1 else if (n<220) zlow=-1.5 #find 1st class mark
> zlow
[1] -1.5
> int=c()  #set vector of class marks for x
> int=xm+(zlow+0:3)*xsd
> int
[1] 36.57206 36.76548 36.95890 37.15233
```

Step 2. Count the observed frequency in each interval.

```
> y=c()   #count class freq.
> y[1]=length(x[x<=int[1]])
> y[2]=length(x[x<=int[2]])-length(x[x<=int[1]])
> y[3]=length(x[x<=int[3]])-length(x[x<=int[2]])
> y[4]=length(x[x<=int[4]])-length(x[x<=int[3]])
```

```
> y[5]=length(x[x>int[4]])
> y
[1]  8 22 55 20  9
> sum(y)  #check if sum to n
[1] 114
```

Step 3. Calculate the expected probability in each interval.

```
> p=c()   #calculate expected prob.
> p[1]=pnorm(zlow)
> p[2:4]=pnorm(zlow+1:3)-pnorm(zlow+0:2)
> p[5]=1-pnorm(zlow+3)
> p
[1] 0.0668072 0.2417303 0.3829249 0.2417303 0.0668072
> sum(p)  #check if sum to 1
[1] 1
```

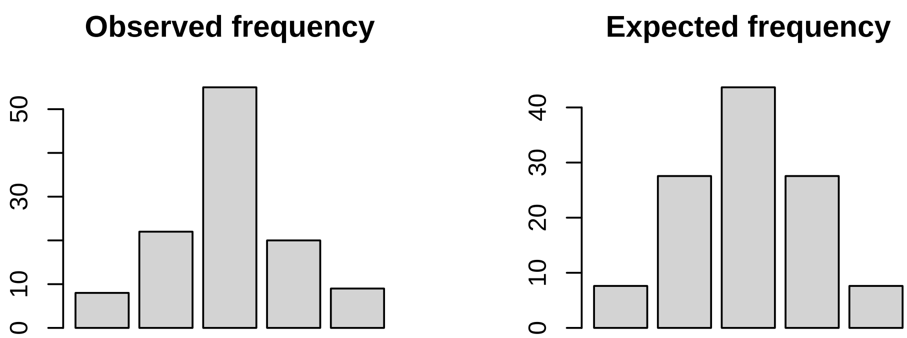
Step 4. Calculate the Chi-square test statistic and the p-value.

```
> ey=n*p
> ey
[1]  7.616021 27.557258 43.653441 27.557258  7.616021
> chi=(y-ey)^2/(ey)
> chi
[1] 0.01935918 1.12068919 2.94923822 2.07249047 0.25149590
> chi=sum(chi)
> chi
[1] 6.413273
> p.value=1-pchisq(chi,k-1-2)
> p.value
[1] 0.04049258
```

Step 5. Decision. Since the p-value is less than 0.05, we reject H_0. There are sufficient evidence in the data against H_0 and hence we conclude that the data do not follow a normal distribution.

Bar charts that compare observed `y` and expected `ey` frequencies are

```
> par(mfrow=c(2,2))
> barplot(y,col="lightgray",main="Observed frequency")
> barplot(ey,col="lightgray",main="Expected frequency")
```



31 Tests for homogeneity and independence (P.519-528)

31.1 Tests for homogeneity

Suppose that several samples are taken from some independent populations, each of which is categorized according to the same set of variables. We want to test whether the probability distributions (proportions) of the categories are the same over different populations.

Example: (Voters) A survey of voter sentiment was conducted in Labor and Liberal to compare the fraction of voters favouring a new tax reform package. Random samples of 100 voters were polled in each of the two parties, with results as follows:

	Approve	Not approve	No comment	Total
Labor	62	29	9	100
Liberal	47	46	7	100
Total	109	75	16	200

Do the data present sufficient evidence to indicate that the fractions of voters favouring the new tax reform package differ in Labor and Liberal?

Solution: Let $p_{1j}, j = 1, 2, 3$, denote the proportions of `Approve, Not approve` and `No comment` for the new tax reform package in `Labor` repespectively.

Similarly, let $p_{2j}, j = 1, 2, 3$, denote the proportions of `Approve, Not approve` and `No comment` for the new tax reform package in `Liberal` repespectively.

The question becomes to test the hypothesis:

$$H_0 : \; p_{1j} = p_{2j}, \quad j = 1, 2, 3 \quad \text{vs} \quad H_1 : \; \text{Not all equalities hold.}$$

In general, suppose the data is presented in the following *contingency table*:

```
                              Categories

                           1      2  ...    c     Total

                   1   y_11   y_12 ... y_1c      y_1.
                   2   y_21   y_22 ... y_2c      y_2.
                   .      .      . ...    .
     Populations   .      .      . ...    .
                   .      .      . ...    .
                   r   y_r1   y_r2 ... y_rc      y_r.

       Total           y_.1   y_.2 ... y_.c        n
```

Note: there are rc categories and either row or column totals are fixed (n is also fixed).

The expected frequency falling in the jth category of the ith population should be $y_{i.}\, p_{ij}$ where p_{ij} denote the probability of an observation from ith population falling into jth category.. Hence we should reject H_0 if

$$\chi_0^{\prime 2} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(y_{ij} - y_{i.}\, p_{ij})^2}{y_{i.}\, p_{ij}}$$

is large.

However $\chi_0^{\prime 2}$ includes unknown parameters p_{ij}. It can be proved that under the H_0 of homogeneity across populations, the maximum likelihood estimates of the parameters p_{ij} are given by

$$\hat{p}_{ij} = \hat{p}_{.j} = y_{.j}/n, \quad i = 1, 2, ..., r,$$

which is the pooled sample proportion of the j-th category. Hence, instead of the $\chi_0'^2$, we may use the test statistic

$$\chi_0^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(y_{ij}-y_{i.}\hat{p}_{ij})^2}{y_{i.}\hat{p}_{ij}} = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(y_{ij}-y_{i.}y_{.j}/n)^2}{y_{i.}y_{.j}/n}.$$

The five steps of the test for homogeneity are:

1. **Hypotheses:** $H_0: \ p_{1j} = p_{2j} = ... = p_{rj} \quad j = 1, 2, ..., c$ vs
   H_1 : Not all equalities hold.
2. **Test statistic:** $\chi_0^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(y_{ij}-y_{i.}y_{.j}/n)^2}{y_{i.}y_{.j}/n}$
3. **Assumption:** $E_{ij}=y_{i.}y_{.j}/n \geq 5$. Under H_0, $\chi_0^2 \sim \chi^2_{(r-1)(c-1)}$ approximately.
4. **P-value:** $\Pr(\chi^2_{(r-1)(c-1)} \geq \chi_0^2)$.
5. **Decision:** Reject H_0 if the p-value $< \alpha$.

Example: (Voters)

Calculation is done by completing the following table:

		Approve ($j=1$)	Not approve ($j=2$)	No comment ($j=3$)	row total $y_{i\cdot}$
Labor	$O_{ij}=y_{ij}$	62	29	9	100
($i=1$)	$E_{ij}=\frac{y_{i\cdot}y_{\cdot j}}{n}$	$\frac{109\cdot 100}{200}=54.5$	$\frac{75\cdot 100}{200}=37.5$	$\frac{16\cdot 100}{200}=8$	100
	$\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$	$\frac{(7.5)^2}{54.5}=1.032$	$\frac{(-8.5)^2}{37.5}=1.927$	$\frac{(1.0)^2}{8.0}=0.125$	
Liberal	$O_{ij}=y_{ij}$	47	46	7	100
($i=2$)	$E_{ij}=\frac{y_{i\cdot}y_{\cdot j}}{n}$	$\frac{109\cdot 100}{200}=54.5$	$\frac{75\cdot 100}{200}=37.5$	$\frac{16\cdot 100}{200}=8$	100
	$\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$	$\frac{(-7.5)^2}{54.5}=1.032$	$\frac{(8.5)^2}{37.5}=1.927$	$\frac{(-1.0)^2}{8.0}=0.125$	6.168
Col. total	$\sum_i O_{ij}$	109	75	16	200
	$\sum_i E_{ij}$	109	75	16	200

The test for homogeneity across the two populations is

1. **Hypotheses:** $H_0: \; p_{1j}=p_{2j} \quad j=1,2,3$
   vs H_1 : Not all equalities hold.
2. **Test statistic:** $\chi_0^2=\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(y_{ij}-y_{i.}y_{.j}/n)^2}{y_{i.}y_{.j}/n}=6.1676.$
3. **Assumption:** Under H_0 and with $E_{ij}=y_{i.}y_{.j}/n \geq 5$, $\chi_0^2 \sim \chi^2_{(r-1)(c-1)}$ approximately.
4. ***P*-value:** $\Pr(\chi_2^2 \geq 6.1676) \in (0.025, 0.05)$
   ($\chi^2_{2,0.95}=5.991$; $\chi^2_{2,0.975}=7.378$; 0.045786 from R).

5. **Decision:** Since the p-value < 0.05, the data present sufficient evidence to indicate that the proportions of voters favoring the new tax reform package is different in Labor and Liberal.

In R,

```
> y=c(62,47,29,46,9,7)
> n=sum(y)
> n
[1] 200
> c=3
> r=2
> y.mat=matrix(y,r,c)  #default is to fill by col.
> y.mat
     [,1] [,2] [,3]
[1,]   62   29    9
[2,]   47   46    7
> chisq.test(y.mat)

        Pearson's Chi-squared test

data:  y.mat
X-squared = 6.1676, df = 2, p-value = 0.04579

> yr=apply(y.mat,1,sum)   #checking
> yr
[1] 100 100
> yc=apply(y.mat,2,sum)
> yc
[1] 109  75  16
> yr.mat=matrix(yr,r,c,byrow=F)
```

```
> yr.mat
     [,1] [,2] [,3]
[1,]  100  100  100
[2,]  100  100  100
> yc.mat=matrix(yc,r,c,byrow=T)
> yc.mat
     [,1] [,2] [,3]
[1,]  109   75   16
[2,]  109   75   16
> ey.mat=yr.mat*yc.mat/n
> ey.mat
     [,1] [,2] [,3]
[1,] 54.5 37.5    8
[2,] 54.5 37.5    8
> ey.mat>=5  #test Eij>=5
     [,1] [,2] [,3]
[1,] TRUE TRUE TRUE
[2,] TRUE TRUE TRUE
> chi=(y.mat-ey.mat)^2/ey.mat
> chi
        [,1]     [,2]  [,3]
[1,] 1.03211 1.926667 0.125
[2,] 1.03211 1.926667 0.125
> chi=sum(chi)
> chi
[1] 6.167554
> p.value=1-pchisq(chi,(r-1)*(c-1))
> p.value
[1] 0.04578601
```

31.2 Tests for independence

Many times a sample may be categorized according to two or more *factors* and it is of interest to know whether the factors for the classification are independent.

Example: (Advertisement) 200 randomly sampled people are classified according to their sexes and their reactions to an advertisement for a product (positive, negative, no opinion).

	Positive	Negative	No opinion	Total
M	24	46	38	108
F	32	22	38	92
Total	56	68	76	200

Do the data present sufficient evidence to indicate that the sexes and opinions are related?

Solution: In general, suppose a sample of size n is classified into categories according to two factors and the data is presented in a *contingency table* as follows:

```
                        Factor 1

                    1      2   ...   c     Total

                1  y_11   y_12 ... y_1c     y_1.
                2  y_21   y_22 ... y_2c     y_2.
                .     .      . ...   .
      Factor 2  .     .      . ...   .
                .     .      . ...   .
                r  y_r1   y_r2 ... y_rc     y_r.

       Total       y_.1   y_.2 ... y_.c      n
```

We want to know whether the two facters are independent or related.

Note: we have rc categories and neither row nor column totals are fixed (but n is fixed).

Let p_{ij} denote the probability of an observation falling in the (i,j)th category. The *marginal* row and column probabilities are respectively:

$$p_{i.} = \sum_{j=1}^{c} p_{ij} \quad \text{and} \quad p_{.j} = \sum_{i=1}^{r} p_{ij}.$$

Under H_0 of independence, the expected frequency should be $n\,p_{ij} = n\,p_{i.}\,p_{.j}$. Hence

$$\chi_0'^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(y_{ij} - n\,p_{i.}\,p_{.j})^2}{n\,p_{i.}\,p_{.j}}$$

will be large if we should reject H_0.

However $\chi_0'^2$ includes unknown parameters $p_{i.}$ and $p_{.j}$. It can be proved that the maximum likelihood estimates of $p_{i.}$ and $p_{.j}$ under the H_0 of independence between the two factors are given by the pooled sample proportions respectively as

$$\hat{p}_{i.} = y_{i.}/n, \qquad \hat{p}_{.j} = y_{.j}/n.$$

Hence, instead of the $\chi_0'^2$, we may use the test statistic

$$\chi_0^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(y_{ij} - n\,\hat{p}_{i.}\,\hat{p}_{.j})^2}{n\,\hat{p}_{i.}\,\hat{p}_{.j}} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(y_{ij} - y_{i.}y_{.j}/n)^2}{y_{i.}y_{.j}/n}.$$

The five steps of the test for independence between the two factors are:

1. **Hypotheses:** $H_0: p_{ij} = p_{i.}p_{.j}, \quad i = 1, 2, ..., r; j = 1, 2, ..., c$
   vs $H_1:$ Not all equalities hold.
2. **Test statistic:** $\chi_0^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(y_{ij} - y_{i.}y_{.j}/n)^2}{y_{i.}y_{.j}/n}$
3. **Assumption:** $E_{ij} = y_{i.}y_{.j}/n \geq 5$. Under H_0, $\chi_0^2 \sim \chi^2_{(r-1)(c-1)}$ approximately.
4. ***P*-value:** $\Pr(\chi^2_{(r-1)(c-1)} \geq \chi_0^2)$
5. **Decision:** Reject H_0 if the p-value $< \alpha$.

Example: (Advertisement)

Solution: Let $p_{1j}, j = 1, 2, 3$, denote the probability of a `male` having the opinions: `positive`, `negative` and `no opinion` respectively.

Similarly, let $p_{2j}, j = 1, 2, 3$, denote the probability of a `female` having the opinions: `positive`, `negative` and `no opinion` respectively.

Calculation is done by completing the following table:

		Positive ($j=1$)	Negative ($j=2$)	No opinion ($j=3$)	Row prob. $p_{i\cdot}$
Male	$O_{ij}=y_{ij}$	24	46	38	108
($i=1$)	$E_i=np_{i\cdot}p_{\cdot j}$	$200(0.28)(0.54)=30.24$	$200(0.34)(0.54)=36.72$	$200(0.38)(0.54)=41.04$	$\frac{108}{200}=0.54$
	$\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$	$\frac{(-6.24)^2}{30.24}=1.288$	$\frac{(9.28)^2}{36.72}=2.345$	$\frac{(-3.04)^2}{41.0}=0.225$	
Female	$O_{ij}=y_{ij}$	32	22	38	92
($i=2$)	$E_{ij}=np_{i\cdot}p_{\cdot j}$	$200(0.28)(0.46)=25.76$	$200(0.34)(0.46)=31.28$	$200(0.38)(0.46)=34.96$	$\frac{92}{200}=0.46$
	$\frac{(O_i-E_i)^2}{E_i}$	$\frac{(6.24)^2}{25.76}=1.512$	$\frac{(-9.28)^2}{31.28}=2.753$	$\frac{(3.04)^2}{35.0}=0.264$	8.387
	$\sum_j O_{ij}$	56	68	76	200
	$\sum_j E_{ij}$	56	68	76	200
	Col. prob. $p_{\cdot j}$	$\frac{56}{200}=0.28$	$\frac{68}{200}=0.34$	$\frac{76}{200}=0.38$	

The test for independence between factors of 'gender' and 'opinion' is

1. **Hypotheses:** $H_0: p_{ij}=p_{i\cdot}p_{\cdot j}, \quad i=1,2; j=1,2,3,$
   vs $H_1:$ Not all equalities hold.
2. **Test statistic:** $\chi_0^2=\sum_{i=1}^{2}\sum_{j=1}^{3}\frac{(y_{ij}-np_{i\cdot}p_{\cdot j})^2}{np_{i\cdot}p_{\cdot j}}=8.39.$
3. **Assumption:** Under H_0 and with $E_{ij}=np_{i\cdot}p_{\cdot j}\geq 5$, $\chi_0^2\sim\chi^2_{(r-1)(c-1)}$ approximately.
4. ***P*-value:** $\Pr(\chi_2^2\geq 8.39)\in(0.01,0.025)$
   ($\chi^2_{2,0.975}=7.378$; $\chi^2_{2,0.99}=9.210$; 0.015 from R).
5. **Decision:** Since the p-value <0.05, the data are against H_0. There is strong evidence in the data that the factor of 'sex' and 'opinion' are related.

In R,

```
> y=c(24,32,46,22,38,38)
> n=sum(y)
> n
[1] 200
> c=3
> r=2
> y.mat=matrix(y,r,c)
> y.mat
     [,1] [,2] [,3]
[1,]   24   46   38
[2,]   32   22   38

        Pearson's Chi-squared test

data:  y.mat
X-squared = 8.3871, df = 2, p-value = 0.01509

> pr=apply(y.mat,1,sum)/n  #checking
> pr
[1] 0.54 0.46
> pc=apply(y.mat,2,sum)/n
> pc
[1] 0.28 0.34 0.38
> pr.mat=matrix(pr,r,c,byrow=F)
> pr.mat
     [,1] [,2] [,3]
[1,] 0.54 0.54 0.54
[2,] 0.46 0.46 0.46
> pc.mat=matrix(pc,r,c,byrow=T)
```

```
> pc.mat
     [,1] [,2] [,3]
[1,] 0.28 0.34 0.38
[2,] 0.28 0.34 0.38
> ey.mat=n*pr.mat*pc.mat
> ey.mat
      [,1]  [,2]  [,3]
[1,] 30.24 36.72 41.04
[2,] 25.76 31.28 34.96
> ey.mat>=5  #test Eij>=5
     [,1] [,2] [,3]
[1,] TRUE TRUE TRUE
[2,] TRUE TRUE TRUE
> chi=(y.mat-ey.mat)^2/ey.mat
> chi
         [,1]     [,2]      [,3]
[1,] 1.287619 2.345272 0.2251852
[2,] 1.511553 2.753146 0.2643478
> chi=sum(chi)
> chi
[1] 8.387123
> p.value=1-pchisq(chi,(r-1)*(c-1))
> p.value
[1] 0.01509244
```