DATA2002

Testing for independence

Garth Tarr

Testing for independence in 2×2 tables

Testing for independence in general tables

Testing for independence in 2×2 tables

Titanic

- The `Titanic` dataset comes preloaded in R.
- This data set provides information on the fate of passengers on the fatal maiden voyage of the ocean liner Titanic, summarized according to economic status (class), sex, age and survival.

> ⍰ Is there any evidence that being a women increased your chance of survival on the ship?

```
x = as.data.frame(Titanic)
head(x)
```

```
##   Class    Sex   Age Survived Freq
## 1   1st   Male Child       No    0
## 2   2nd   Male Child       No    0
## 3   3rd   Male Child       No   35
## 4  Crew   Male Child       No    0
## 5   1st Female Child       No    0
## 6   2nd Female Child       No    0
```

```
y.mat = xtabs(Freq ~ Sex + Survived, x)
y.mat
```

```
##          Survived
## Sex         No  Yes
##   Male    1364  367
##   Female   126  344
```

Tests for independence

- There are times where a sample may be categorised according to two or more factors.
- It is of interest to know whether the factors for the classification are independent.
- We summarise what we know in a contingency table

	Survived	Did not survive	Row total
Male	y_{11}	y_{12}	$y_{1\bullet}$
Female	y_{21}	y_{22}	$y_{2\bullet}$
Column total	$y_{\bullet 1}$	$y_{\bullet 2}$	n

Table of proportions

Let p_{ij} denote the probability of an observation falling in the $(i,j)^{th}$ category.

The marginal row and column probabilities are respectively:

$$p_{i\bullet} = \sum_{j=1}^{2} p_{ij} \quad \text{and} \quad p_{\bullet j} = \sum_{i=1}^{2} p_{ij}$$

	Survived	Did not survive	Row total
Male	p_{11}	p_{12}	$p_{1\bullet}$
Female	p_{21}	p_{22}	$p_{2\bullet}$
Column total	$p_{\bullet 1}$	$p_{\bullet 2}$	1

Independence

X and Y are said to be independent if

$$P(X = x \mid Y = y) = P(X = x) \quad \text{or} \quad P(X = x, Y = y) = P(X = x)P(Y = y).$$

In the context of a 2×2 table of proportions,

	Survived	Did not survive	Row total
Male	p_{11}	p_{12}	$p_{1\bullet}$
Female	p_{21}	p_{22}	$p_{2\bullet}$
Column total	$p_{\bullet 1}$	$p_{\bullet 2}$	1

Let X be a random variable representing survival status and let Y be a random variable representing sex. Under independence,

$$p_{11} = P(X = \text{Survived}, Y = \text{Male}) = P(X = \text{Survived})P(Y = \text{Male}) = p_{\bullet 1}p_{1\bullet}$$

Test statistic

Under the null hypothesis of independence, the expected frequencies are $e_{ij} = np_{ij} = np_{i\bullet}p_{\bullet j}$.

A large test statistic,

$$T = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(Y_{ij} - e_{ij})^2}{e_{ij}} = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(Y_{ij} - np_{i\bullet}p_{\bullet j})^2}{np_{i\bullet}p_{\bullet j}},$$

indicates that we should reject H_0.

However T includes unknown parameters $p_{i\bullet}$ and $p_{\bullet j}$.

We need to estimate $p_{i\bullet}$ and $p_{\bullet j}$ by

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

Hence, we calculate the observed test statistic,

$$t_0 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(y_{ij} - n\,\hat{p}_{i\bullet}\,\hat{p}_{\bullet j})^2}{n\,\hat{p}_{i\bullet}\,\hat{p}_{\bullet j}} = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(y_{ij} - y_{i\bullet}y_{\bullet j}/n)^2}{y_{i\bullet}y_{\bullet j}/n}.$$

Hypothesis testing workflow

The workflow for the test of independence between two variables in a 2×2 contingency table are:

- **Hypothesis:** H_0: $p_{ij} = p_{i\bullet}p_{\bullet j}, \quad i = 1, 2;\ j = 1, 2$ vs H_1: Not all equalities hold.
- **Assumptions:** independent observations and $e_{ij} = y_{i\bullet}y_{\bullet j}/n \geq 5$.
- **Test statistic:** $T = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(Y_{ij} - e_{ij})^2}{e_{ij}}$. Under H_0, $T \sim \chi^2_1$ approx.
- **Observed test statistic:** $t_0 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(y_{ij} - y_{i\bullet}y_{\bullet j}/n)^2}{y_{i\bullet}y_{\bullet j}/n}$
- **P-value:** $P(T \geq t_0) = P(\chi^2_1 \geq t_0)$
- **Decision:** Reject H_0 if the p-value $< \alpha$

Titanic

- The `Titanic` dataset comes preloaded in R.
- This data set provides information on the fate of passengers on the fatal maiden voyage of the ocean liner Titanic, summarized according to economic status (class), sex, age and survival.

> ⍰ Is there any evidence that being a women increased your chance of survival on the ship?

```
x = as.data.frame(Titanic)
head(x)
```

```
##   Class    Sex   Age Survived Freq
## 1   1st   Male Child       No    0
## 2   2nd   Male Child       No    0
## 3   3rd   Male Child       No   35
## 4  Crew   Male Child       No    0
## 5   1st Female Child       No    0
## 6   2nd Female Child       No    0
```

```
y.mat = xtabs(Freq ~ Sex + Survived, x)
y.mat
```

```
##          Survived
## Sex         No  Yes
##   Male    1364  367
##   Female   126  344
```

Titanic

The test for independence between survival and sex is:

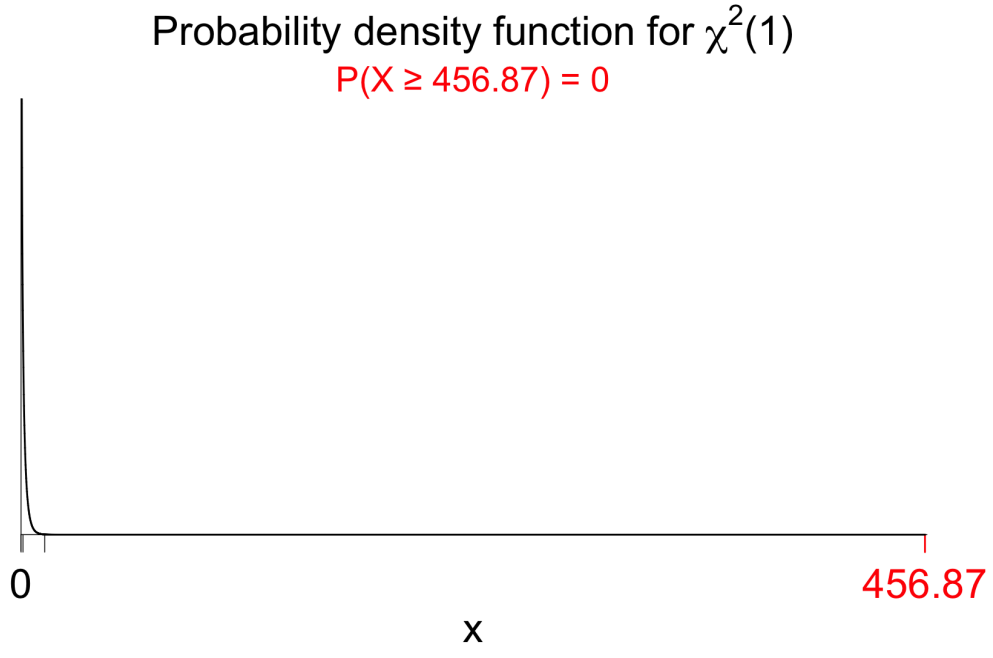
- **Hypothesis:** H_0: $p_{ij} = p_{i\bullet}p_{\bullet j}, \quad i = 1, 2;\ j = 1, 2$ vs H_1: Not all equalities hold.
- **Assumptions:** $e_{ij} = y_{i\bullet}y_{\bullet j}/n \geq 5$ (do you think we have independent observations?).
- **Test statistic:** $T = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(Y_{ij} - e_{ij})^2}{e_{ij}}$. Under H_0, $T \sim \chi^2_1$ approx.
- **Observed test statistic:** $t_0 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(y_{ij} - y_{i\bullet}y_{\bullet j}/n)^2}{y_{i\bullet}y_{\bullet j}/n} = 456.87$
- **P-value:** $P(T \geq t_0) = P(\chi^2_1 \geq 456.87) < 0.001$
- **Decision:** We reject the null hypothesis that sex is independent of survival as the p-value is very small (much smaller than 0.05). Hence, there is evidence to suggest that survival status of passengers on the Titanic is related to the sex of the passenger.

```
y.mat
```

```
##          Survived
## Sex         No  Yes
##   Male    1364  367
##   Female   126  344
```

```
chisq.test(y.mat, correct = FALSE)
```

```
##
##     Pearson's Chi-squared test
##
## data:  y.mat
## X-squared = 456.87, df = 1, p-value < 2.2e-16
```



```
r = c = 2
(yr = apply(y.mat, 1, sum))  # rowSums(y.mat)
```

```
##   Male Female
##   1731    470
```

```
(yc = apply(y.mat, 2, sum))  # colSums(y.mat)
```

```
##   No  Yes
## 1490  711
```

```
(yr.mat = matrix(yr, r, c, byrow = FALSE))
```

```
##      [,1] [,2]
## [1,] 1731 1731
## [2,]  470  470
```

```
(yc.mat = matrix(yc, r, c, byrow = TRUE))
```

```
##      [,1] [,2]
## [1,] 1490  711
## [2,] 1490  711
```

```
(ey.mat = yr.mat * yc.mat / sum(y.mat))
```

```
##           [,1]     [,2]
## [1,] 1171.8264 559.1736
## [2,]  318.1736 151.8264
```

```
# or using matrix multiplication
# ey.mat = yr %*% t(yc) / n
all(ey.mat >= 5) # check e_{ij} >= 5
```

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

```
(t0 = sum((y.mat - ey.mat)^2 / ey.mat))
```

```
## [1] 456.8742
```

```
(pval = pchisq(t0, 1, lower.tail = FALSE))
```

```
## [1] 2.302151e-101
```

Testing for independence in general tables

Advertisement

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

	Positive	Negative	No opinion	Total
High income	24	46	38	108
Low income	32	22	38	92
Total	56	68	76	200

Do the data present sufficient evidence to indicate that income level and opinion are related?

General 2-way contingency tables

We now turn our attention from 2×2 table to $r \times c$ tables with classifying factors R at r levels and S at c levels. Let

$$P(R = i, S = j) = p_{ij}, \quad 1 \leq i \leq r,\ 1 \leq j \leq c.$$

Let y_{ij} be the observed count in the (i, j)th cell. Again, we will use the $\bullet$ notation to denote summation over a particular index. So that $y_{i\bullet}$ is the sum of values in the ith row and $y_{\bullet j}$ is the sum of values in the jth column.

We can summarise this data in a contingency table,

	$S = 1$	$S = 2$	$\cdots$	$S = c$	Total
$R = 1$	y_{11}	y_{12}	$\cdots$	y_{1c}	$y_{1\bullet}$
$R = 2$	y_{21}	y_{22}	$\cdots$	y_{2c}	$y_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	
$R = r$	y_{r1}	y_{12}	$\cdots$	y_{rc}	$y_{r\bullet}$
Total	$y_{\bullet 1}$	$y_{\bullet 2}$	$\cdots$	$y_{\bullet c}$	$y_{\bullet\bullet}$

Hence, $y_{\bullet\bullet} = \sum_{i=1}^{r} \sum_{j=1}^{c} y_{ij} = n$, the sample size.

General 2-way contingency tables

To make this a little more concrete, 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:

		Variable 1				
		Level 1	Level 2	...	Level c	Row total
Variable 2	Level 1	y_{11}	y_{12}	$\cdots$	y_{1c}	$y_{1\bullet}$
	Level 2	y_{21}	y_{22}	$\cdots$	y_{2c}	$y_{2\bullet}$
	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
	Level r	y_{r1}	y_{r2}	$\cdots$	y_{rc}	$y_{r\bullet}$
	Column total	$y_{\bullet 1}$	$y_{\bullet 2}$	$\cdots$	$y_{\bullet 3}$	$y_{\bullet\bullet} = n$

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

Independence

Let p_{ij} denote the probability of an observation falling in the $(i,j)^{th}$ category.

In the general case we can summarise the probabilities as depicted below:

	$S=1$	$S=2$	$\cdots$	$S=c$	
$R=1$	p_{11}	p_{12}	$\cdots$	p_{1c}	$p_{1\bullet}$
$R=2$	p_{21}	p_{22}	$\cdots$	p_{2c}	$p_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	
$R=r$	p_{r1}	p_{12}	$\cdots$	p_{rc}	$p_{r\bullet}$
	$p_{\bullet 1}$	$p_{\bullet 2}$	$\cdots$	$p_{\bullet c}$	1

The marginal row and column probabilities are respectively:

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

Estimating p_{ij} under independence

Suppose we have a completely random sample of size n classified by R and S. We want to test

$$H_0\text{: } R \text{ and } S \text{ are independent.}$$

If R and S are independent, by definition of independence,

$$p_{ij} = P(R = i, S = j) = P(R = i)P(S = j)$$

This equation allows us to estimate the marginal probabilities $P(R = i)$ and $P(S = j)$ via

$$P(R = i) = p_{i\bullet} = \sum_{j=1}^{c} p_{ij} \text{ estimated by } \frac{1}{n}\sum_{j=1}^{c} y_{ij} = \frac{y_{i\bullet}}{n} \quad \text{and}$$

$$P(S = j) = p_{\bullet j} = \sum_{i=1}^{r} p_{ij} \text{ estimated by } \frac{1}{n}\sum_{i=1}^{r} y_{ij} = \frac{y_{\bullet j}}{n}.$$

Hence, under the independence hypothesis, the expected frequency in the (i, j)th cell is

$$e_{ij} = n\hat{p}_{ij} = n\hat{p}_{i\bullet}\hat{p}_{\bullet j} = \frac{y_{i\bullet}y_{\bullet j}}{n}$$

Advertisement

	Positive	Negative	No opinion	Total
High income	p_{11}	p_{12}	p_{13}	$p_{1\bullet}$
Low income	p_{21}	p_{22}	p_{23}	$p_{2\bullet}$
Total	$p_{\bullet 1}$	$p_{\bullet 2}$	$p_{\bullet 3}$	1

Recall: X and Y are said to be independent if

$$P(X = x \mid Y = y) = P(X = x) \quad \text{or} \quad P(X = x, Y = y) = P(X = x)P(Y = y)$$

Let X be a random variable representing opinion and let Y be a random variable representing income. Under independence,

$$p_{12} = P(X = \text{Negative}, Y = \text{High income}) = P(X = \text{Negative})P(Y = \text{High income}) = p_{\bullet 2}p_{1\bullet}$$

Test statistic

Under H_0 of independence, the expected frequencies are $e_{ij} = np_{ij} = np_{i\bullet}p_{\bullet j}$ for $i = 1, 2, \ldots, r$ and $j = 1, 2, \ldots, c$. Hence

$$T = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(Y_{ij} - e_{ij})^2}{e_{ij}} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(Y_{ij} - n\,p_{i\bullet}\,p_{\bullet j})^2}{n\,p_{i\bullet}\,p_{\bullet j}}$$

will be large if we should reject H_0.

However T includes unknown parameters $p_{i\bullet}$ and $p_{\bullet j}$. We estimate $p_{i\bullet}$ and $p_{\bullet j}$ with

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

Hence, we may calculate the observed test statistic as

$$t_0 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(y_{ij} - n\,\hat{p}_{i\bullet}\,\hat{p}_{\bullet j})^2}{n\,\hat{p}_{i\bullet}\,\hat{p}_{\bullet j}} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(y_{ij} - y_{i\bullet}y_{\bullet j}/n)^2}{y_{i\bullet}y_{\bullet j}/n}.$$

Degrees of freedom

	Positive	Negative	No opinion	Total
High income	y_{11}	y_{12}	y_{13}	$y_{1\bullet}$
Low income	y_{21}	y_{22}	y_{23}	$y_{2\bullet}$
Total	$y_{\bullet 1}$	$y_{\bullet 2}$	$y_{\bullet 3}$	n

- The degrees of freedom for a $r \times c$ table is $(r-1)(c-1)$.

$$(rc-1)-(r-1)-(c-1) = rc-r-c+1 = (r-1)(c-1)$$

- The degrees of freedom for a 2×3 table is 2.

Hypothesis testing workflow

The workflow for the test of independence between the two factors are:

- **Hypothesis:** H_0: $p_{ij} = p_{i\bullet}p_{\bullet j}, i = 1, 2, \ldots, r, j = 1, 2, \ldots, c$ vs H_1: Not all equalities hold.
- **Assumptions:** independent observations and $e_{ij} = y_{i\bullet}y_{\bullet j}/n \geq 5$.
- **Test statistic:** $T = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(y_{ij} - e_{ij})^2}{e_{ij}}$. Under H_0, $T \sim \chi^2_{(r-1)(c-1)}$ approx.
- **Observed test statistic:** $t_0 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(y_{ij} - y_{i\bullet}y_{\bullet j}/n)^2}{y_{i\bullet}y_{\bullet j}/n}$.
- **P-value:** $P(T \geq t_0) = P(\chi^2_{(r-1)(c-1)} \geq t_0)$
- **Decision:** Reject H_0 if the p-value $< \alpha$

As noted in the assumptions, this test should only really be applied when all cells have expected counts greater than 5, $e_{ij} \geq 5$. If there are expected cell counts less than 5, consider using simulation or Fisher's exact test.

Advertisement

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

	Positive	Negative	No opinion	Total
High income	24	46	38	108
Low income	32	22	38	92
Total	56	68	76	200

Do the data present sufficient evidence to indicate that income level and opinion are related?

Advertisement

The test for independence between factors of *income level* and *opinion* is

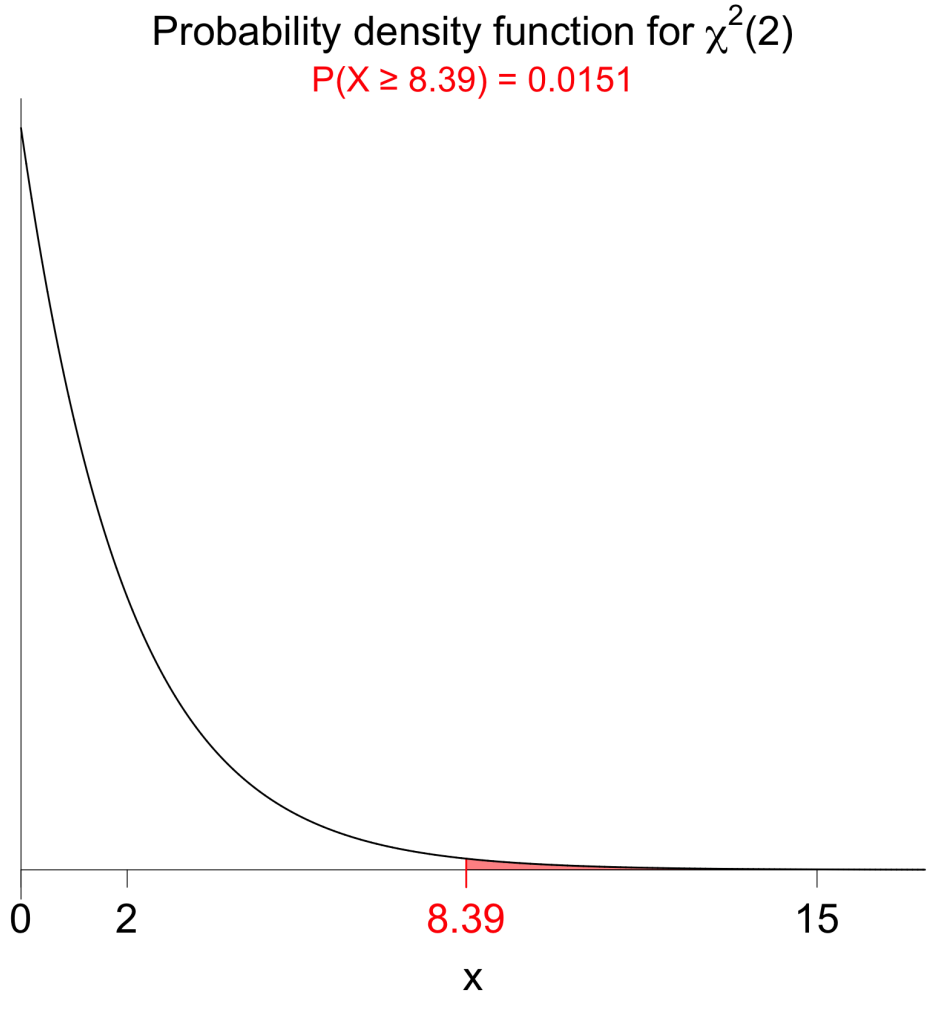
- **Hypothesis:** H_0: $p_{ij} = p_{i\bullet}p_{\bullet j}$, $i = 1, 2$; $j = 1, 2, 3$, vs H_1: Not all equalities hold **or** H_0: income level is independent of opinion vs H_1: income level is not independent of opinion.
- **Assumptions:** independent observations (satisfied, they were randomly sampled)
  $e_{ij} = y_{i\bullet}y_{\bullet j}/n \geq 5$ (satisfied, checked with calculations).
- **Test statistic:** $T = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(y_{ij} - e_{ij})^2}{e_{ij}}$. Under H_0, $T \sim \chi^2_2$ approx.
- **Observed test statistic:** $t_0 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(y_{ij} - y_{i\bullet}y_{\bullet j}/n)^2}{y_{i\bullet}y_{\bullet j}/n} = 8.39.$
- **P-value:** $P(T \geq t_0) = P(\chi^2_2 \geq 8.39) = 0.015$
- **Decision:** Since the p-value is less than 0.05, the data provide evidence against H_0. There is evidence to suggest that there is an association between income level and opinion.

```r
y = c(24, 32, 46, 22, 38, 38)
n = sum(y)
c = 3
r = 2
y.mat = matrix(y, nrow = r, ncol = c)
colnames(y.mat) = c("Positive", "Negative", "Nc
rownames(y.mat) = c("High income", "Low income"
y.mat
```

```
##              Positive Negative No opinion
## High income        24       46         38
## Low income         32       22         38
```

```r
chisq.test(y.mat, correct = FALSE)
```

```
##
##     Pearson's Chi-squared test
##
## data:  y.mat
## X-squared = 8.3871, df = 2, p-value = 0.01509
```



```r
(yr = apply(y.mat, 1, sum))  # rowSums(y.mat)
```

```
## High income  Low income
##          108          92
```

```r
(yc = apply(y.mat, 2, sum))  # colSums(y.mat)
```

```
##   Positive   Negative No opinion
##         56         68         76
```

```r
(yr.mat = matrix(yr, r, c, byrow = FALSE))
```

```
##      [,1] [,2] [,3]
## [1,]  108  108  108
## [2,]   92   92   92
```

```r
(yc.mat = matrix(yc, r, c, byrow = TRUE))
```

```
##      [,1] [,2] [,3]
## [1,]   56   68   76
## [2,]   56   68   76
```

```r
# matrix mult: ey.mat = yr %*% t(yc) / n
(ey.mat = yr.mat * yc.mat / sum(y.mat))
```

```
##       [,1]  [,2]  [,3]
## [1,] 30.24 36.72 41.04
## [2,] 25.76 31.28 34.96
```

```r
all(ey.mat>=5) # check all e_{ij} >= 5
```

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

```r
(t0 = sum((y.mat - ey.mat)^2 / ey.mat))
```

```
## [1] 8.387123
```

```r
(pval = pchisq(t0, (r - 1) * (c - 1),
               lower.tail=FALSE))
```

```
## [1] 0.01509244
```

References

For further details see Larsen and Marx (2012), section 10.5.

Franke, T. M., T. Ho, and C. A. Christie (2012). "The Chi-Square Test: Often Used and More Often Misinterpreted". In: *American Journal of Evaluation* 33.3, pp. 448-458. DOI: 10.1177/1098214011426594. URL: https://journals-sagepub-com.ezproxy.library.sydney.edu.au/doi/10.1177/1098214011426594.

Larsen, R. J. and M. L. Marx (2012). *An Introduction to Mathematical Statistics and its Applications*. 5th ed. Boston, MA: Prentice Hall. ISBN: 978-0-321-69394-5.