

# Tutorial 9 Solutions

## Question 1

(14.2) Previous enrolment records at a large university indicate that of the total number of persons who apply for admission, 60% are admitted unconditionally, 5% are conditionally admitted, and the remainder are refused admission. Of 500 applicants to date for next year, 329 were admitted unconditionally, 43 were conditionally admitted, and the remainder were not admitted. Do the data indicate a departure from previous admission rates?

(a) Carry out the test using  $\alpha = 0.05$ .

If we let the subscripts 1 represent unconditional admission, 2 represent conditional admission, and 3 represent refusal, the hypotheses we are interested in testing are:

$$H_0 : p_1 = 0.60, p_2 = 0.05, p_3 = 0.35 \quad \text{vs} \quad H_1 : \text{At least one } p_i \neq p_{i,0}.$$

The observed and expected counts are as follows:

	Unconditional admission	Conditional admission	Refused
Observed	329	43	128
Expected	$500(0.6) = 300$	$500(0.05) = 25$	$500(0.35) = 175$

The test statistic is:

$$V = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \quad \underset{\sim}{\text{approx}} \quad \chi_{k-1}^2,$$

where  $O_i$  is the observed frequencies,  $E_i$  is the expected frequencies (under  $H_0$ ), and  $k$  is the number of categories. The degrees of freedom  $k - 1$  arises from the restriction that  $\sum_{i=1}^k p_i = 1$ . Plugging in the values from the table, the value of our test statistic is:

$$v = \frac{(329 - 300)^2}{300} + \frac{(43 - 25)^2}{25} + \frac{(128 - 175)^2}{175} = 28.386,$$

on  $3 - 1 = 2$  degrees of freedom. As this is always an upper-tailed test, we reject  $H_0$  if  $V > \chi_{k-1, \alpha}^2$ .

```
qchisq(0.05, df=3-1, lower.tail=FALSE)
```

```
## [1] 5.991465
```

Since  $28.386 > 5.991$ , we reject the null hypothesis in favour of the alternative. We conclude that there is evidence to support the claim that current admission rates differ from previous admission rates.

(b) Find the  $p$ -value associated with the test in (a).


The (approximate)  $p$ -value is found as:

```
pchisq(28.386, df=3-1, lower.tail=FALSE)
```

```
## [1] 6.855804e-07
```

Since  $6.856 \times 10^{-7} < 0.05$ , we reject the null hypothesis once again and make the same conclusions as in (a).

## Question 2

(14.8)  The Mendelian theory states that the number of a type of peas that fall into the classifications:

- Round and yellow
- Wrinkled and yellow
- Round and green
- Wrinkled and green

should be in the ratio of 9:3:3:1. Suppose that 100 such peas revealed 56, 19, 17, and 8 in the respective categories. Are the data consistent with the model? Use  $\alpha = 0.05$ . (Note: the expression 9:3:3:1 means that 9/16 of the peas should be round and yellow, 3/16 should be wrinkled and yellow, etc.)

Let R denote round, Y denote yellow, W denote wrinkled, and G denote green. The hypotheses we are interested in testing are:

$$H_0 : \text{Ratio is 9:3:3:1} \quad \text{vs} \quad H_1 : H_0 \text{ untrue,}$$

or equivalently,

$$H_0 : p_{RY} = 9/16, p_{WY} = 3/16, p_{RG} = 3/16, p_{WG} = 1/16 \quad \text{vs} \quad H_1 : \text{At least one } p_i \neq p_{i,0}.$$

We carry out this test using the `chisq.test()` function in **R**.


```
observed_peas <- c(56, 19, 17, 8)
expected_probabilities <- c(9/16, 3/16, 3/16, 1/16)

chisq.test(observed_peas, p=expected_probabilities)
```

```
##
## Chi-squared test for given probabilities
##
## data:  observed_peas
## X-squared = 0.65778, df = 3, p-value = 0.8831
```

From the output above, the  $p$ -value is given as 0.8831. Since  $0.8831 > 0.05$ , we fail to reject the null hypothesis. We conclude that there is insufficient evidence to support the claim that the 9:3:3:1 ratio does not hold.

## Question 3

(14.12)  The numbers of accidents experienced by machinists were observed for a fixed period of time, with the results as shown in the table below. Test at the 5% significance level that the data come from a Poisson

distribution.

Accidents per machinist	Frequency of observation (number of machinists)
0	296
1	74
2	26
3	8
4	4
5	4
6	1
7	0
8	1

Let  $Y$  be the number of accidents experienced by machinists. The hypotheses we wish to test are:

$$H_0 : Y \sim \text{Poisson}(\lambda) \quad \text{vs} \quad H_1 : Y \text{ is not Poisson}(\lambda).$$

We begin by entering our data.

```
accidents <- data.frame(
  number = 0:8,
  observed_count = c(296, 74, 26, 8, 4, 4, 1, 0, 1)
)
```

We must make an estimate of  $\lambda$ . We recall that the MLE of  $\lambda$  was given by  $\hat{\lambda} = \bar{Y}$ . In the setting of tabulated count data,

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i = \frac{1}{n} \sum_{j=1}^k Y_j \cdot n_{y_j},$$

i.e. the sum can be obtained by multiplying the group values by their group counts and summing those values together. We calculate  $\hat{\lambda}$  as follows:

```
lambda_hat <- with(accidents, (1 / sum(observed_count)) * sum(number * observed_count))

lambda_hat
```

```
## [1] 0.4830918
```

In the above, I wrap my calculations using the `with()` function so that I do not need to keep typing `accidents$column` to access the columns. Next, we create a column for the expected counts. The expected counts are calculated as:

$$\hat{\mathbf{E}}(n_i) = n\hat{p}_i = 414 \cdot \frac{\hat{\lambda}^{y_i} e^{-\hat{\lambda}}}{y_i!}$$

We use `dpois()` to obtain our Poisson values (d for density, though technically, this is a discrete distribution so it is mass).

```
accidents <- accidents |>
  transform(expected_count = sum(observed_count) * dpois(number, lambda=lambda_hat))

accidents
```

##	number	observed_count	expected_count
## 1	0	296	2.553855e+02
## 2	1	74	1.233746e+02
## 3	2	26	2.980064e+01
## 4	3	8	4.798814e+00
## 5	4	4	5.795670e-01
## 6	5	4	5.599681e-02
## 7	6	1	4.508600e-03
## 8	7	0	3.111525e-04
## 9	8	1	1.878940e-05

We notice that for  $Y \geq 3$ , the expected counts drops below 5. As stated on page 715 of the textbook, a rule of thumb is given that we should aim for expected cell counts to be at least five (though they can be as low as one). As such, we should collapse all rows for  $Y \geq 3$  to be represented by  $Y = 3$  by summing their expected counts together.

We can do this by first changing the values under the `number` column – if the `number` is greater or equal to three, we change the value to three, otherwise it remains the same. Then we convert it to a factor.

```
accidents <- accidents |>
  transform(number = ifelse(number >= 3, 3, number)) |>
  transform(number = as.factor(number))

accidents
```

##	number	observed_count	expected_count
## 1	0	296	2.553855e+02
## 2	1	74	1.233746e+02
## 3	2	26	2.980064e+01
## 4	3	8	4.798814e+00
## 5	3	4	5.795670e-01
## 6	3	4	5.599681e-02
## 7	3	1	4.508600e-03
## 8	3	0	3.111525e-04
## 9	3	1	1.878940e-05

Next, we collapse all rows that have the same value of `number` by taking the sums of the values of the other columns. Note that the first three rows of the data will remain unchanged after taking their group column sums since they only have one row for each group.

```
accidents <- aggregate(. ~ number, data=accidents, sum)

accidents
```

##	number	observed_count	expected_count
## 1	0	296	255.385505
## 2	1	74	123.374640
## 3	2	26	29.800638
## 4	3	18	5.439217

We will not use the `chisq.test()` function for this question. While the resulting value of the test statistic will be correct, **R** does not know that we made an additional estimate (the MLE for  $\lambda$ ), which would decrease our degrees of freedom by one. As such, the reported degrees of freedom will be incorrect, as will the  $p$ -value. Therefore, we will perform this test manually.

The value of the test statistic is computed as follows:

```
V <- with(accidents, sum((observed_count - expected_count)^2 / expected_count))
```

```
V
```

```
## [1] 55.71012
```

The degrees of freedom is given by  $k - 2 = 4 - 2 = 2$ , for we lose one degree of freedom by estimating  $\lambda$  via its MLE, and we have the constraint that the Poisson probabilities sum to one.

The  $p$ -value of this test is:

```
pchisq(V, df=4-2, lower.tail=FALSE)
```

```
## [1] 7.992856e-13
```

Since the  $p$ -value is less than 0.05, we reject the null hypothesis in favour of the alternative. We conclude that there is sufficient evidence to support the claim that the data do not come from a Poisson distribution.

## Question 4

(14.14) A study was conducted to determine the effect of early child care on infant-mother attachment patterns. In the study, 93 infants were classified as either secure or anxious. In addition, the infants were classified according to the average number of hours per week that they spent in child care.

Attachment Pattern	Hours in child care		
	Low (0-3 Hours)	Moderate (4-19 Hours)	High (20-54 Hours)
Secure	24	35	5
Anxious	11	10	8

- (a) Do the data indicate a dependence between attachment patterns and the number of hours spent in child care? Test using  $\alpha = 0.05$ .

The hypotheses we wish to test are:

$H_0$  : There is no dependence between attachment patterns hours spent in child care

vs

$H_1$  : There is a dependence between attachment patterns and hours spent in child care.

We start by calculating the row and column totals in the given table.

Attachment Pattern	Hours in child care			Total
	Low (0-3 Hours)	Moderate (4-19 Hours)	High (20-54 Hours)	
Secure	24	35	5	64
Anxious	11	10	8	29
Total	35	45	13	93

We then create a second table of the expected counts, whose cells are computed using:

$$\widehat{\mathbf{E}(n_{ij})} = \frac{r_i c_j}{n} \equiv \frac{\text{Row total} * \text{Column total}}{\text{Total sample size}}$$

Attachment Pattern	Hours in child care			Total
	Low (0-3 Hours)	Moderate (4-19 Hours)	High (20-54 Hours)	
Secure	$\frac{64 * 35}{93} = 24.09$	$\frac{64 * 45}{93} = 30.97$	$\frac{64 * 13}{93} = 8.95$	64
Anxious	$\frac{29 * 35}{93} = 10.91$	$\frac{29 * 45}{93} = 14.03$	$\frac{29 * 13}{93} = 4.05$	29
Total	35	45	13	93

For a contingency table containing  $r$  rows and  $c$  columns, the test statistic is computed as:

$$V = \sum_{j=1}^c \sum_{i=1}^r \frac{(n_{ij} - \widehat{\mathbf{E}(n_{ij})})^2}{\widehat{\mathbf{E}(n_{ij})}}.$$

Informally, the test statistic is computed by iterating over each cell in the table and calculating a similar statistic used in the chi-square goodness of fit test:

$$V = \sum_{\text{All cells}} \frac{(O - E)^2}{E},$$


where  $O$  is the observed count of the cell and  $E$  is the expected count of the cell. Under the null hypothesis, the distribution of  $V$  is chi-squared with degrees of freedom  $(r-1)(c-1)$ .

Plugging in the numbers from the two tables above, the value of the test statistic is found to be 7.267. The degrees of freedom is  $(r-1)(c-1) = (2-1)(3-1) = 2$ . As this is an upper-tailed test, we reject the null hypothesis if  $V > \chi_{2,0.05}^2$ .

```
qchisq(0.05, df=2, lower.tail=FALSE)
```

```
## [1] 5.991465
```

Since  $7.267 > 5.991$ , we reject the null hypothesis in favour of the alternative. We conclude that there is sufficient evidence to support the claim that there is a dependence between attachment patterns and child care hours.

(b)  Repeat (a) using **R**.

We begin by entering the data from the table.

```
infants <- matrix(
  c(24, 35, 5, 11, 10, 8), byrow=TRUE, nrow=2, ncol=3,
  dimnames = list(c("Secure", "Anxious"), c("Low", "Moderate", "High")))
)
```

infants

```
##           Low Moderate High
## Secure    24         35    5
## Anxious   11         10    8
```

For this question, we can perform the chi-square test simply by supplying our data to the `chisq.test()` function.

```
chisq.test(infants)
```

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

```
##
```

```
## Pearson's Chi-squared test
```

```
##
```

```
## data:  infants
```

```
## X-squared = 7.267, df = 2, p-value = 0.02642
```

(We can safely ignore the warning given in the output.) Since the  $p$ -value is less than 0.05, we reject the null hypothesis. We conclude that there is sufficient evidence to support the claim that there is a dependence between attachment patterns and hours spent in child care.

Note that if we had swapped the dimensions of the contingency table, the result would still be the same. (Again, we can safely ignore the warning given in the output.)

```
infants2 <- t(infants)
```

infants2

```
##           Secure Anxious
## Low          24        11
## Moderate     35        10
## High          5         8
```

```
chisq.test(infants2)
```

```
## Warning in chisq.test(infants2): Chi-squared approximation may be incorrect
```

```
##
```

```
## Pearson's Chi-squared test
```

```
##
```

```
## data:  infants2
```

```
## X-squared = 7.267, df = 2, p-value = 0.02642
```