The University of Sydney
School of Mathematics and Statistics

Solutions to Weekly Independent Exercises 7 (Week 13)

STAT5002: Introduction to Statistics — Semester 1, 2025

Lecturers: T. Cui

1. A company produces bags of lollies of different colours. The company advertises that the percentages of each colour should on average be as given by the following table:

Colour	Blue	Orange	Green	Yellow	Red	Brown
Percentage	24	20	16	14	13	13

A random sample of 60 such lollies gives the following frequencies:

Colour	Blue	Orange	Green	Yellow	Red	Brown
Frequency	9	8	12	15	10	6

Does this seem to be consistent with the advertised percentages? Explain with an appropriate hypotheis test by following the steps below. The R output below should be useful for this. **Note:** the R function `outer(u, v, fun)` returns a matrix whose `(i,j)`-th element is `fun(u[i],v[j])`).

```
Of = c(9, 8, 12, 15, 10, 6)
p0 = c(.24, .2, .16,.14, .13, .13)
Ef = p0*sum(Of)
Ef
```

```
## [1] 14.4 12.0  9.6  8.4  7.8  7.8
```

```
((Of-Ef)^2)/Ef
```

```
## [1] 2.0250000 1.3333333 0.6000000 5.1857143 0.6205128 0.4153846
```

```
qchisq.values=outer(c(.9, .95, .98, .99), c(4,5,6), qchisq)
rownames(qchisq.values)=c("90%", "95%", "98%", "99%")
colnames(qchisq.values)=c("4 df", "5 df", "6 df")
qchisq.values
```

```
##          4 df      5 df     6 df
## 90%  7.779440  9.236357 10.64464
## 95%  9.487729 11.070498 12.59159
## 98% 11.667843 13.388223 15.03321
## 99% 13.276704 15.086272 16.81189
```

(a) State the statistical model we assume to perform an appropriate χ^2 test.

Solution: We model the data as being obtained by drawing a ticket from a box 60 times randomly with replacement, where the tickets have one of the colours "Blue", "Orange", "Green", "Yellow", "Red" or "Brown" written on them. The proportions of each type of ticket are (respectively) in the vector $\boldsymbol{p} = (p_1, \ldots, p_6)$ where each $p_i \geq 0$ and $p_1 + \cdots + p_6 = 1$.

(b) State the appropriate null and alternative hypotheses.

Solution:

- $H_0 \colon \boldsymbol{p} = (0.24, 0.2, 0.16, 0.14, 0.13, 0.13)$.
- $H_1 \colon \boldsymbol{p} \neq (0.24, 0.2, 0.16, 0.14, 0.13, 0.13)$.

(c) Describe the test statistic we use in terms of $O_1, \ldots, O_6$, the (respective) observed frequencies of the colours "Blue", "Orange", "Green", "Yellow", "Red" and "Brown".

Solution: We use Pearson's statistic $\sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i}$ where

- $E_1 = 60 \times 0.24 = 14.4$
- $E_2 = 60 \times 0.2 = 12$
- $E_3 = 60 \times 0.16 = 9.6$
- $E_4 = 60 \times 0.14 = 8.4$
- $E_5 = 60 \times 0.13 = 7.8$
- $E_6 = 60 \times 0.13 = 7.8$

Note that this can (in principle) be specified before we see the data (knowing only the planned sample size).

(d) What is the (approximate) distribution of the test statistic if the null hypothesis is true? Hence specify the rejection region if the test is conducted at the

- 10%
- 5%
- 2%
- 1%

level of significance.

Solution: The approximate distribution of the test statistic is χ^2_5 (the chi-squared distribution with $6 - 1 = 5$ degrees of freedom. We reject if the test statistic is large enough. Thus using the R output above, the rejection region is

- $(9.236, \infty)$ if the test is conducted at the 10% level of significance
- $(11.070, \infty)$ if the test is conducted at the 5% level of significance
- $(13.388, \infty)$ if the test is conducted at the 2% level of significance
- $(15.086, \infty)$ if the test is conducted at the 1% level of significance

Note that these can also be specified (in principle) before the data is observed.

(e) Determine the value taken by the test statistic and hence declare at which of the level(s) 10%, 5%, 2% or 1% the result is significant.

Solution: Using the R output above, the statistic takes the value (approximately, with some rounding)

$$2.025 + 1.333 + 0.6 + 5.186 + 0.621 + 0.415 = 10.18\,.$$

Thus the result is significant at the 10% level, but not the 5% level.

2. Some dogs resemble their owners. The article Roy and Christenfeld (2004) describes an experiment attempting to see if this tendency is due to the owner seeking out a dog that resembles them, or if the dog and owner get "more similar-looking" over time. Since purebred dogs have a more predictable appearance, the researchers thought this tendency might be more prevalent for purebred dogs.

From the paper: "Owners were approached at random and asked if they would be willing to help...with a psychology experiment examining the relation between owners and their dogs. The pictures were taken so that the background was different for dog and owner. This ensured that raters would not be able to match dog and owner by simply comparing the backgrounds in the photographs...Owners...were asked to indicate the breed of their dog... We constructed triads of pictures, each consisting of one owner, that owner's dog, and one other dog photographed at the same park. Each set of 15 pictures was viewed by 28 ... judges. Each judge was instructed to identify which of the two possible dogs belonged to each person. A dog was regarded as resembling its owner if a majority of judges matched the pair."

The results are summarised in the table below.

No. judges matching	> 14	$= 14$	< 14	Total
Purebred	16	0	9	25
Nonpurebred	7	4	9	20
Total	23	4	18	45

(a) There is a chi-squared test we can perform to explore this question. Which type exactly?

Solution: The appropriate test is a **test of independence**. This is where we have a random sample of individuals which is classified according to two factors ("Row" and "Column"). As a consequence the row and column totals are only known after the data is collected.

(b) Specify a box model for describing the "full model".

Solution: Under the full model there is a single box containing N equally likely tickets. There are 6 types of tickets.

- Np_{11} tickets marked "Purebred/> 14 matches" (with proportion p_{11})
- Np_{12} tickets marked "Purebred/$= 14$ matches" (with proportion p_{12})
- Np_{13} tickets marked "Purebred/< 14 matches" (with proportion p_{13})
- Np_{21} tickets marked "Nonpurebred/> 14 matches" (with proportion p_{21})
- Np_{22} tickets marked "Nonpurebred/$= 14$ matches" (with proportion p_{22})
- Np_{23} tickets marked "Nonpurebred/< 14 matches" (with proportion p_{23})

and $p_{11} + p_{12} + p_{13} + p_{21} + p_{22} + p_{23} = 1$. These p_{ij}s thus give the probability that a single draw is of the corresponding type.

This is to say, there is no structural relationship between the Rows factor ("Purebred or Nonpurebred") and the Columns factor (> 14, $= 14$ or < 14 matches), the probabilities of each of the 6 categories are unrestricted (except that they must add to 1: there are thus 5 "free" parameters).

Data is then obtained by taking 45 random draws with replacement from this box, the counts of each type are written in the appropriate cell of the table.

(c) Specify a restricted version of the above model corresponding to the null hypothesis.

Solution: The null hypothesis of independence says that we actually have two boxes:

- a "Rows box" containing L tickets with one of two things written on them:
  - $Lp_{1\bullet}$ marked "Purebred" (with proportion $p_{1\bullet}$)
  - $Lp_{2\bullet}$ marked "Nonpurebred" (with proportion $p_{2\bullet}$)

  and $p_{1\bullet} + p_{2\bullet} = 1$ (only 1 "free" parameter).
- a "Columns" box containing M tickets with one of three things written on them:
  - $Mp_{\bullet 1}$ marked > 14 (with proportion $p_{\bullet 1}$)
  - $Mp_{\bullet 2}$ marked $= 14$ (with proportion $p_{\bullet 2}$)
  - $Mp_{\bullet 3}$ marked < 14 (with proportion $p_{\bullet 3}$)

  and $p_{\bullet 1} + p_{\bullet 2} + p_{\bullet 3} = 1$ (only 2 "free" parameters).

There are thus 3 "free parameters" under the null hypothesis.

Data is then obtained in a two-step process:

1. a random sample of size 45 is taken with replacement from the Rows box;
2. a second random sample of size 45 is taken with replacement from the Columns box;

These are formed into ordered pairs, and the counts of each possible pair are written in the appropriate cell of the table.

That is, the Row for each observation is "statisically independent" of its Column.

Equivalently, a "bigger box" of all possible pairs is constructed. This bigger box has LM possible pairs:

- $LMp_{1\bullet}p_{\bullet 1}$ equal to ("Purebred", > 14)
- $LMp_{1\bullet}p_{\bullet 2}$ equal to ("Purebred", $= 14$)
- $LMp_{1\bullet}p_{\bullet 3}$ equal to ("Purebred", < 14)
- $LMp_{2\bullet}p_{\bullet 1}$ equal to ("Nonpurebred", > 14)
- $LMp_{2\bullet}p_{\bullet 2}$ equal to ("Nonpurebred", $= 14$)
- $LMp_{2\bullet}p_{\bullet 3}$ equal to ("Nonpurebred", < 14)

This is a restricted version of the full model whereby the probabilities of the categories are given by the products:

- $p_{11} = p_{1\bullet}p_{\bullet 1}$
- $p_{12} = p_{1\bullet}p_{\bullet 2}$
- $p_{13} = p_{1\bullet}p_{\bullet 3}$
- $p_{21} = p_{2\bullet}p_{\bullet 1}$
- $p_{22} = p_{2\bullet}p_{\bullet 2}$
- $p_{23} = p_{2\bullet}p_{\bullet 3}$

(d) According to the theory, what is the approximate distribution of Pearson's statistic when the null hypothesis is true?

Solution: There are 5 free parameters under the full model and 3 free parameters under the null hypothesis. The difference $5 - 3 = 2$ is the degrees of freedom of the approximating chi-squared distribution. Alternatively, there are $r = 2$ rows, $c = 3$ columns, and the degrees of freedom of the approximating chi-squared distribution is $(r - 1)(c - 1) = 1 \times 2 = 2$.

(e) Determine appropriate chi-squared critical values for the test if it is conducted at the

- 10%
- 5%

- 2%
- 1%

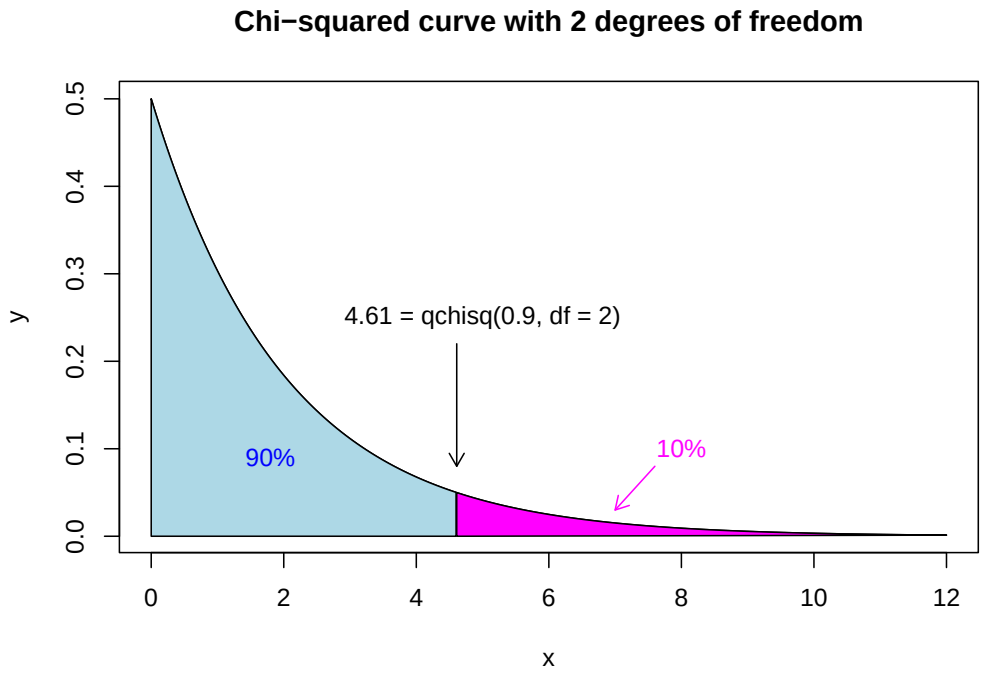
level of significance.

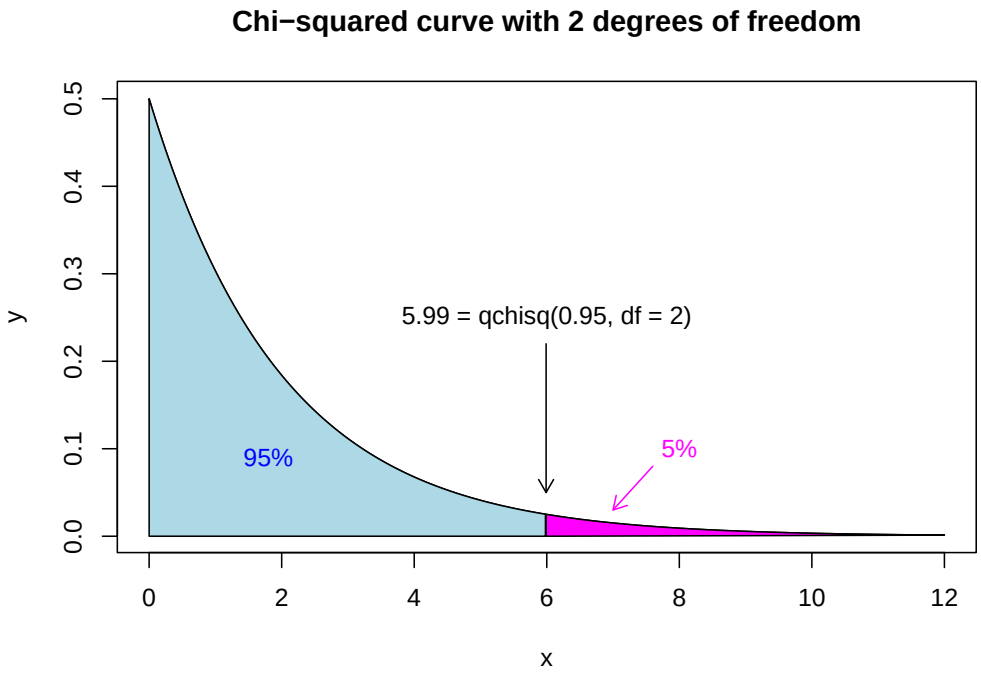
```
qchisq.values=outer(c(.9, .95, .98, .99), c(1, 2, 3, 4, 5, 6), qchisq)
rownames(qchisq.values)=c("90%", "95%", "98%", "99%")
colnames(qchisq.values)=c("1 df", "2 df", "3 df", "4 df", "5 df", "6 df")
qchisq.values
```

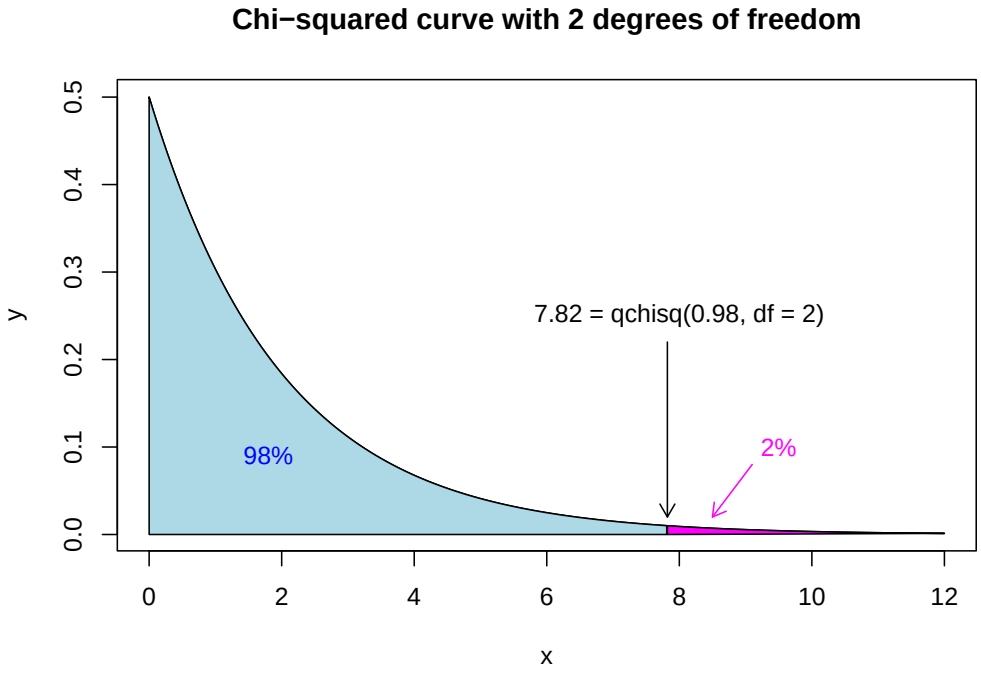
```
##         1 df     2 df      3 df      4 df      5 df     6 df
## 90% 2.705543 4.605170  6.251389  7.779440  9.236357 10.64464
## 95% 3.841459 5.991465  7.814728  9.487729 11.070498 12.59159
## 98% 5.411894 7.824046  9.837409 11.667843 13.388223 15.03321
## 99% 6.634897 9.210340 11.344867 13.276704 15.086272 16.81189
```

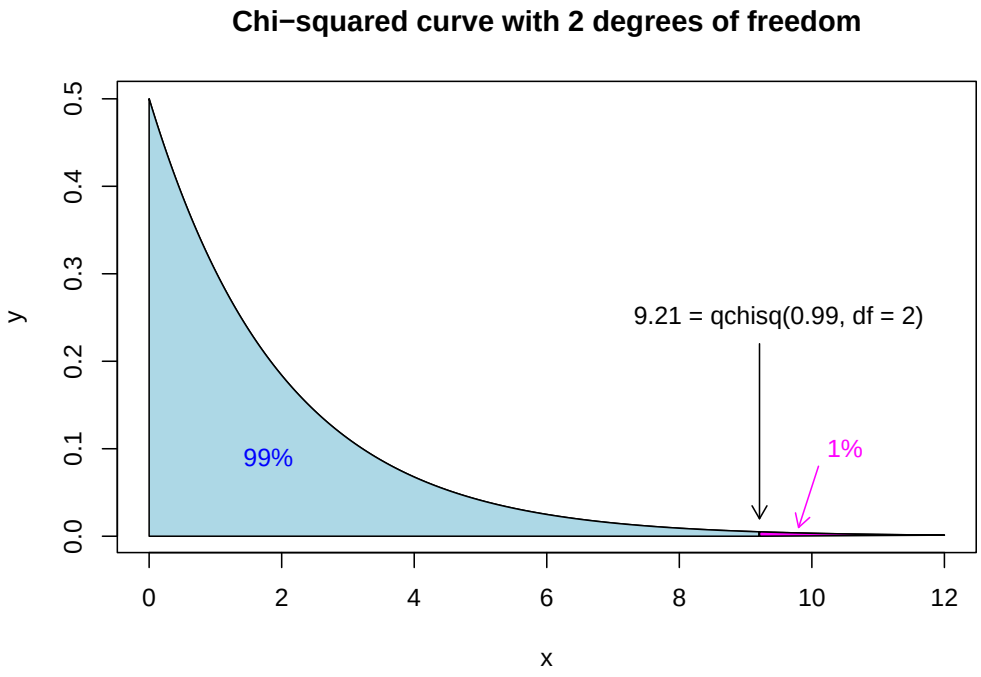
Solution: The appropriate values are in the second column of the matrix `qchisq.values`, since the correct degrees of freedom is 2. The critical value (to two decimal places) at the

- 10% level is 4.61
- 5% level is 5.99
- 2% level is 7.82
- 1% level is 9.21




Chi−squared curve with 2 degrees of freedom
5.99 = qchisq(0.95, df = 2)
95%
5%
y
x
0.0
0.1
0.2
0.3
0.4
0.5
0
2
4
6
8
10
12


Chi−squared curve with 2 degrees of freedom
7.82 = qchisq(0.98, df = 2)
98%
2%
y
x
0.0
0.1
0.2
0.3
0.4
0.5
0
2
4
6
8
10
12



(f) Using the R output below determine the value of Pearson's statistic for testing this null hypothesis.

```
nonpure =c(7, 4, 9)
pure=c(16, 0, 9)
Of = rbind(nonpure, pure)
rs = apply(Of, 1, sum)
rs
```

```
## nonpure    pure
##      20      25
```

```
cs = apply(Of, 2, sum)
cs
```

```
## [1] 23  4 18
```

```
Ef = outer(rs,cs)/sum(Of)
Ef
```

```
##             [,1]     [,2] [,3]
## nonpure 10.22222 1.777778    8
## pure    12.77778 2.222222   10
```

```
SR = (Of-Ef)/sqrt(Ef)
SR
```

```
##                [,1]       [,2]       [,3]
## nonpure -1.0078197  1.666667  0.3535534
## pure     0.9014213 -1.490712 -0.3162278
```

```
SR^2
```

```
##              [,1]     [,2]  [,3]
## nonpure 1.0157005 2.777778 0.125
```

```
## pure    0.8125604 2.222222 0.100
```

Solution: The statistic is the sum of the entries in this last matrix, which is (after some rounding)

$$1.016 + 2.778 + 0.125 + 0.813 + 2.222 + 0.1 = 7.054\,.$$

(g) Using the chi-squared approximation at which level(s) is the result significant?

Solution: The result is significant at the 5% level (thus also 10%) but **not** at the 2% (nor the 1%) level.

(h) Is there anything to suggest that the chi-squared approximation should not be trusted?

Solution: Yes, since in the matrix of expected frequencies (called `Ef`) there are two which are less than 5. In such a situation it is advisable to also use simulation.