The University of Sydney
School of Mathematics and Statistics

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.

(b) State the appropriate null and alternative hypotheses.

(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".

(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.

(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.

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?

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

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

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

(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
```

(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
```

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

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