(Learning outcomes concerning R proficiency are tested via coursework in this module)

MA26620 : May 2022 Solutions

S: similar to a type previously seen (NB: levels of similarity vary)
B: bookwork
U: unseen

(Not much is unseen, since this is a methods course)

S 1. a) ONE-WAY ANOVA

Source	SS	DF	MS	F-ratio
Between species ✓	18 ✓	3 ✓	6 ✓	3
Within species	40 ✓	20 ✓	2 ✓	
Total (corr.)	58	23 ✓		

(1 mark per filled cell +2 if all correct)

Working: $3 = \dfrac{SS_{BETWEEN}/3}{SS_{WITHIN}/20}$, and $SS_{BETWEEN} + SS_{WITHIN} = 58$, implying

$$\begin{pmatrix} 20 & -9 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} SS_{BETWEEN} \\ SS_{WITHIN} \end{pmatrix} = \begin{pmatrix} 0 \\ 58 \end{pmatrix} \Rightarrow \begin{pmatrix} SS_{BETWEEN} \\ SS_{WITHIN} \end{pmatrix} = \frac{1}{29}\begin{pmatrix} 1 & 9 \\ -1 & 20 \end{pmatrix}\begin{pmatrix} 0 \\ 58 \end{pmatrix} = \begin{pmatrix} 18 \\ 40 \end{pmatrix}$$

b) H_0 : mean distance travelled is the same for all species.
H_1 : mean distance travelled is not the same for all species. 2

Yes, we would believe that the distance travelled is the same for all species, since the p-value of 0.05486 (from R) is not significant at the 5% level. Insufficient evidence to reject H_0. 3

(Equivalently, stats tables give $P(F_{3,20} > 3.0984) = 0.05$, and since $3 < 3.0984$, the test is not significant at the 5% level)

15 Q1

(Continues overleaf)

5 2. We conduct a chi-squared test:

χ^2 TEST

	A	B	C	D	E
Expected frequency (E_i)	$\frac{16}{31} \times 200$	$\frac{8}{31} \times 200$	$\frac{4}{31} \times 200$	$\frac{2}{31} \times 200$	$\frac{1}{31} \times 200$
Observed frequency (O_i)	127	42	20	8	3
$\frac{(O_i - E_i)^2}{E_i}$	5.4755	1.7904	1.3065	1.8632	1.8466

Thus $\chi^2_{obs} = \sum_{A,B,C,D,E} \frac{(O_i - E_i)^2}{E_i} = 12.2822.$ 5

Assuming observations are independent, ✓ χ^2 is approximately distributed as $\chi^2_{(4)}$, ✓ and $P(\chi^2_{(4)} > 12.2822) = 0.01537 < 0.05$, ✓

So the test is significant at the 5% level. ✓

(Equivalently, stats tables give that $0.01 < p_0 < 0.02$ – this is equally acceptable)

The model is not a good fit for the data. ✓ 5

10 Q2

(Continues overleaf)

S 3. ONE-SAMPLE T

Let X_i denote the weight in grams of the i-th Gibert's potoroo, $i = 1, \dots, 7$, and assume $X_1, \dots, X_7 \sim N(\mu, \sigma^2)$ independently. 3

We test $H_0 : \mu = 1050$ vs $H_1 : \mu < 1050$. 2

Under H_0, $T = \dfrac{\bar{X} - \mu}{(S/\sqrt{7})} \sim t_{[6]}$, and $T_{obs} = \dfrac{970 - 1050}{\sqrt{\frac{23000}{7}}} = -1.3956$.

Now, $p_0^- = P(T < -1.3956) = P(t_{[6]} < -1.3956) = P(t_{[6]} > 1.3956)$
$= 0.106$. 3

Since $p_0^- < 0.1$, the test is not significant (or between 10% and 20% from tables) at even the 10% level.

Insufficient evidence to reject H_0; the marsupials on the reduced-variety diet are not significantly lighter. 2

10 Q3

S 4. REGRESSION

a) $\bar{x} = 120.83\dot{3}$, $\bar{y} = 27$, $\sum_{i=1}^{6} x_i^2 = 88{,}775$, $\sum_{i=1}^{6} x_i = 725$, $\sum_{i=1}^{6} x_i y_i = 19{,}825$, $\sum_{i=1}^{6} y_i = 162$.

Therefore $S_{xx} = \sum_{i=1}^{6} x_i^2 - \frac{1}{6}\left(\sum_{i=1}^{6} x_i\right)^2 = 88{,}775 - \frac{1}{6}725^2 = 1{,}170.83\dot{3}$,

and $S_{xy} = \sum_{i=1}^{6} x_i y_i - \frac{1}{6}\sum_{i=1}^{6} x_i \sum_{i=1}^{6} y_i = 19{,}825 - \frac{1}{6}(725)(162) = 250$.
4

Thus $\hat{\beta}_1 = \dfrac{S_{xy}}{S_{xx}} = \dfrac{250}{1170.83\dot{3}} \approx 0.2135$, 2

and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 27 - \frac{250}{117.83\dot{3}} \times 120.83\dot{3} = 1.1993$. 2

The regression line's equation is therefore $y = 0.2135x + 1.1993$. 1

b) $R^2 = \dfrac{S_{xy}^2}{S_{xx}S_{yy}} = \dfrac{250^2}{1{,}170.83\dot{3} \times 60} = 0.8897$.

89% of the variability in y is accounted for by its linear relationship with x; a strong linear relationship. 3

c) (i) $0.2135 \times 125 + 1.1993 = 27.9$ mm
(ii) $0.2135 \times 200 + 1.1993 = 43.9$ mm although this is extrapolation 3

d) $20\hat{\beta}_1 = 4.27$mm. 2

17 Q4

S 5. Assume gull attacks are independent and occur at a constant rate λ per hour. Then $N(t)$, the number of attacks in t hours, is distributed as $Po(\lambda t)$. 3

POISSON HYPOTHESIS TEST

We test $H_0 : \lambda = 3$ vs $H_1 : \lambda > 3$. 2

Under H_0, $N(10) \sim Po(30)$, so $p_0^+ = P(N(10) \geq 42) = P(Po(30) \geq 42)$
$= 0.0221$. 2

This is significant at the 5% level; moderate evidence against H_0 in favour of H_1. The rate has increased. 2

9 Q5

S 6. Let m denote the number of ewes in the sample that are expecting twins. Then the total number of lambs expected is $126 = 2m + (70 - m) = m + 70 \Rightarrow m = 56$.

BINOMIAL CONFIDENCE INTERVAL

Thus $\hat{p}$, the unbiased estimate of the proportion of ewes expecting twins, is $\hat{p} = \frac{56}{70}$. 3

Assuming the probability of a ewe expecting twins has fixed probability, p say, for all ewes, independently of each other, then the number of ewes from a flock of size n is distributed as $Bin(n, p)$.

Thus $ESE(\hat{p}) = \sqrt{\hat{p}(1-\hat{p})/n} = \sqrt{\frac{56}{70} \cdot \frac{14}{70} \cdot \frac{1}{70}} = 0.047809$, 3

and since n is large, the 90% confidence interval is given by $\hat{p} \pm z_{0.05} \times ESE(\hat{p})$

$$= \frac{56}{70} \pm 1.6449 \times 0.047809$$

$$= 0.8 \pm 0.0786$$

$$= (0.721, 0.879).$$ 3

9 Q6

SECTION A

70

5 7. Let X_i denote the maximum speeds of the red kangaroos $(i=1,\dots,22)$.

TWO-SAMPLE T-TEST Let Y_i denote the maximum speeds of the eastern grey kangaroos $(i=1,\dots,17)$.

2

Assuming all observations are independent and $X_i \sim N(\mu_1, \sigma^2)$, $Y_i \sim N(\mu_2, \sigma^2)$

then $\mathrm{Var}(\bar{X}-\bar{Y}) = \mathrm{Var}(\bar{X}) + \mathrm{Var}(\bar{Y})$ (by independence)

$$= \frac{\sigma^2}{22} + \frac{\sigma^2}{17}$$

$$= \frac{39\sigma^2}{374}.$$

2

Thus $ESE(\bar{X}-\bar{Y}) = \sqrt{\frac{39S^2}{374}}$, where S^2 is the pooled sample variance.

Let us compute S^2:

$$S^2 = \frac{21\times 7^2 + 16\times 6^2}{22+17-2} = \frac{1605}{37} = 43.378.$$

Thus $ESE(\bar{X}-\bar{Y}) = 2.1268.$

We aim to test $H_0: \mu_1-\mu_2=0$ vs $H_1: \mu_1-\mu_2>0$

The T-statistic, $T = \frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{ESE(\bar{X}-\bar{Y})} \sim t_{[37]}$, and

under H_0, $T_{obs} = \frac{5-0}{2.1268} = 2.3509.$

4

Now, $p_0^+ = P(T > 2.3509) = P(t_{[37]} > 2.3509)$

$= 0.01208.$

(or between 1% and 2.5% from tables)

The test is significant at the 2% level, fairly strong evidence to reject H_0 in favour of H_1. The red kangaroos are faster.

3

11 Q7

U 8. In general, the confidence interval will be of the form

(Choosing sample size not covered in lectures or previously seen questions)

CHOOSING SAMPLE SIZE

$$\bar{X} \pm t_{\alpha[n-1]} \frac{S}{\sqrt{n}}.$$

This interval has width $2t_{\alpha[n-1]} \frac{S}{\sqrt{n}}$. 3

We must therefore choose n such that $2t_{\alpha[n-1]} \frac{S}{\sqrt{n}} \leqslant 5$

$$\Rightarrow t_{\alpha[n-1]} \frac{15}{\sqrt{n}} \leqslant \frac{5}{2}$$

$$\Rightarrow 15\, t_{\alpha[n-1]} \leqslant \frac{5}{2}\sqrt{n}$$

$$\Rightarrow 225\left(t_{\alpha[n-1]}\right)^2 \leqslant \left(\frac{5}{2}\right)^2 n$$

$$\Rightarrow n \geqslant 36\left(t_{\alpha[n-1]}\right)^2,$$

where $\alpha = 0.025$. 3

Such a value can be found via trial-and-error, or via an iterative scheme (which could be quickly implemented in R for instance). It's perhaps quicker to instead examine the relevant column in statistical tables.

In any case, the smallest n for which the inequality holds is $n = 141$. 6

12 Q8

Example of R implementation using only syntax seen in practicals:

```
for (n in 2:100){
    if (36 * (qt (0.025, df = n-1, lower.tail = FALSE))^2 <= n){
        cat (n); break;
    }
}
```

(Continues overleaf)

S 9.

BINOMIAL HYPOTHESIS TEST WITH APPROXIMATION

Denoting the number of nests containing a cuckoo chick by R, and assuming nests are independent with constant probability p of having a cuckoo chick, $R \sim Bin(350, p)$. 2

We test $H_0 : p = 0.06$ vs $H_1 : p < 0.06$. 2

Under H_0, $p_0^- = P(R \leqslant 10) = P(Bin(350, 0.06) \leqslant 10)$.

Since this isn't in the tables, we will make an approximation; since p is small ($p < 0.1$), we approximate $Bin(350, 0.06)$ as $Po(350 \times 0.06) = Po(21)$, so

$$\begin{aligned} p_0^- \doteqdot P(Po(21) \leqslant 10) &= 1 - P(Po(21) \geqslant 11) \\ &= 1 - 0.9937 \\ &= 0.0063. \end{aligned}$$

4

The test is significant at the 1% level; strong evidence against H_0 in favour of H_1. The proportion of cuckoo-containing nests has declined.

2

10 Q9

(Continues overleaf)

10. 5

(although candidates haven't seen many examples of using contrasts in (c))

CONTRASTS & ANOVA

a) 7, informed by the Df column (since $k=4$ groups, and the error Df is $k(m-1)$, where m is the number of observations per group. Thus $24 = 4(m-1) \Rightarrow m=7$). 2

b) Yes, p-value 0.0228 is significant at the 5% level. 2

c) We construct three orthogonal contrasts as follows:

$$\lambda_{\text{terriers}} = \mu_{ST} - \mu_{WHWT}$$

$$\lambda_{\text{collies}} = \mu_{BC} - \mu_{RC}$$

$$\lambda_{T\text{ vs }C} = \tfrac{1}{2}(\mu_{ST} + \mu_{WHWT}) - \tfrac{1}{2}(\mu_{BC} + \mu_{RC})$$ 3

These are estimated as:

$$\hat{\lambda}_{\text{terriers}} = 56.14286 - 53.71429 = 2.42857;$$

$$\hat{\lambda}_{\text{collies}} = 48 - 42.71429 = 5.28571;$$

$$\hat{\lambda}_{T\text{ vs }C} = \tfrac{1}{2}(56.14286 + 53.71429 - 48 - 42.71429) = 9.57143.$$ 3

The sums of squares associated with these contrasts are:

$$L(\hat{\lambda}_{\text{terriers}}) = \frac{7 \times 2.42857^2}{2} = 20.64283,$$

$$L(\hat{\lambda}_{\text{collies}}) = \frac{7 \times 5.28571^2}{2} = 97.78556,$$

$$L(\hat{\lambda}_{T\text{ vs }C}) = 7 \times 9.57143^2 = 641.2859.$$ 3

Consequently:

Source	DF	SS	MS	F	P
Between terriers	1	20.64283	20.64283	0.311	0.58
Between collies	1	97.78556	97.78556	1.474	0.24
Terriers vs collies	1	641.2859	641.2859	9.670	0.0048
Error	24	1591.7	66.32		

Equivalently, from table given, since $9.670 > 4.260$, only the final p-value reaches significance at 5% level.

The first two p-values are not significant; the third is significant at the 0.5% level, hence (i) no, (ii) no, (iii) yes. 4

17 Q10