(Learning outcomes concerning R progiciency are assessed via coursework in this module)

MA 26620 : Applied Statistics

2023

SECTION A

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)

Q1. S

a) Continuous, interval. 2

b) Discrete, ordinal. 2

c) (Effectively) continuous, ratio. 2

d) Discrete, nominal. 2

(8)$_{Q1}$

Q2. S

Let X_i denote the weight of apple-packet i, $i = 1, \ldots, 20$.

Assume $X_1, X_2, \ldots, X_{20} \sim N(\mu, \sigma^2)$ independently. 3

Then $T = \dfrac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{[19]}$.

For a 95% confidence interval, note that $t_{0.025[19]} = 2.0930$. 3

It follows that the 95% C.I. for μ is given by $\bar{X} \pm t_{0.025[19]} \times \dfrac{S}{\sqrt{n}}$

$= 700 \pm 2.0930 \times \dfrac{30}{\sqrt{20}}$

$= 700 \pm 14.040$

$= (685.960,\ 714.040)$ (in g). 3

(9)$_{Q2}$

Q3. S

If the die is fair, then $P(\text{rolling } i) = \frac{1}{6}$ for all $i = 1, \ldots, 6$.

	1	2	3	4	5	6
Observed (O_i)	21	18	10	15	17	9
Expected (E_i)	15	15	15	15	15	15
$\dfrac{(O_i - E_i)^2}{E_i}$	$\frac{36}{15}$	$\frac{9}{15}$	$\frac{25}{15}$	0	$\frac{4}{15}$	$\frac{36}{15}$

Summing the bottom row, $X^2_{obs} = \frac{110}{15} = 7.3\dot{3}$. 6 Moreover, $P(X^2_{(5)} > 7.3\dot{3})$ lies between 0.1 and 0.2 (from tables). Test is not significant at the 10% level; we have insufficient evidence to reject the null hypothesis that the die is fair. 3

(9)$_{Q3}$

Q4. S

a) $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{121}{175} = 0.691.$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 8.0 - \frac{121}{175} \times 20 = -5.829.$

Thus the regression line is $y = 0.691x - 5.829.$ 6

b) $R^2 = \frac{S_{xy}^2}{S_{xx} S_{yy}} = \frac{121^2}{175 \times 106} = 0.789.$

78.9% of the variation in sales is accounted for by its linear relationship with temperature (a strong linear relationship). 2

c) $y_* = 0.691 \times 22 - 5.829$

$= 9.383,$

so £9383. 2

(10) Q4

Q5. S

Let $N(t)$ denote the number of accidents in t months.

Assuming accidents occur randomly at a constant rate λ per month, with events independent of each other, then $N(t) \sim Po(\lambda t)$. 3

We test $H_0: \lambda = 4$ vs $H_1: \lambda > 4.$

Under H_0, $N(t) \sim Po(4t)$, so $N(6) \sim Po(24)$. 3

We calculate $p^+ = P(N(6) \geq 37) = P(Po(24) \geq 37) = 0.0082.$ (from tables)

The test is significant at the 1% level; strong evidence against H_0 in favour of H_1. The rate has significantly increased. 4

(10) Q5

Q6. a) S

Source	SS	DF	MS	F-ratio	P
Between composts	300	4	75	3	0.0281
Within composts	1125	45	25		
Total (corr)	1425	49			

7 (one for each blue quantity)

b) $m = \frac{45}{5} + 1 = 10.$ 2

c) $Y_{ij} = \mu_i + \varepsilon_{ij}, \quad i = 1, \ldots, 5, \quad j = 1, \ldots, 10.$

$\varepsilon_{ij} \sim N(0, \sigma^2)$ independently.

Test $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ vs H_1: not all group means are equal. 3

Since the p-value is less than 0.05, the test is significant at the 5% level; moderate evidence against H_0 in favour of H_1. Not all group means are equal. 2

(14) Q6

Q7. Letting X denote response time in minutes, assume $X \sim N(\mu, \sigma^2)$ S independently for each breakdown. 2

We test $H_0: \mu = 50$ vs $H_1: \mu > 50$. 2

Under H_0, we have that $T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{[24]}$, 2 and

$$T_{obs} = \frac{70 - 60}{(19/5)} = 2 \cdot 632$$

Thus $p_0^+ = P(t_{[24]} > 2 \cdot 632)$, which (from tables) is a little less than 0·01. 2

The test is significant at the 1% level. Strong evidence to reject H_0 in favour of H_1. Response time is slower than claimed. 2

10 Q7

SECTION A
70

SECTION B

Q8. a) $\text{Var}(\bar{X}-\bar{Y}) = \text{Var}(\bar{X}) + \text{Var}(\bar{Y}) - \overbrace{2\text{Cov}(\bar{X},\bar{Y})}^{=0 \text{ by assumption of independence.}}$

B

$$= \text{Var}\left(\frac{1}{40}\sum_{i=1}^{40} X_i\right) + \text{Var}\left(\frac{1}{30}\sum_{i=1}^{30} Y_i\right) - 0$$

$$= \frac{1}{40^2}\sum_{i=1}^{40} \text{Var}\, X_i + \frac{1}{30^2}\sum_{i=1}^{30} \text{Var}\, Y_i \quad \text{(again by independence)}$$

$$= \frac{1}{40^2}(40\sigma_x^2) + \frac{1}{30^2}(30\sigma_y^2)$$

$$= \frac{\sigma_x^2}{40} + \frac{\sigma_y^2}{30}. \qquad 3$$

Replacing population parameters with sample estimates and taking square roots:

$$ESE(\bar{X}-\bar{Y}) = \sqrt{\frac{S_x^2}{40} + \frac{S_y^2}{30}} = \sqrt{\frac{2^2}{40} + \frac{6^2}{30}}$$

$$= 1.140175. \qquad 2$$

S b) We test $H_0: \mu_y - \mu_x = 5$ vs $H_1: \mu_y - \mu_x > 5$. 2

Under H_0, $T = \dfrac{\bar{y}-\bar{x}-5}{ESE(\bar{y}-\bar{x})} \sim t_{[\nu]}$, where

$$\nu = \frac{(f_x + f_y)^2}{\frac{1}{39}f_x^2 + \frac{1}{29}f_y^2},$$ where $f_x = \frac{4}{40} = 0.1$ and $f_y = \frac{36}{30} = 1.2$.

Thus $\nu = 33.86$, which we round down to 33. 4

Note that $T_{obs} = \dfrac{3}{1.140175} = 2.631174.$ 2

Finally, $p_0^+ = P(t_{[33]} > 2.631174)$, which lies between 0.5% and 1% from tables.

Strong evidence against H_0 in favour of H_1; the sports scientists' claim is supported by the evidence. 3

16 Q8

Q9. S

Denoting the number of failing components by R, and assuming failures are independent with constant probability of failure p, $R \sim Bin(450, p)$. 2

We test $H_0: p = 0.04$ vs $H_1: p > 0.04$. 2

Under H_0, $p_0^+ = P(R \geq 25)$.

Since this is not in the tables, and $p < 0.1$ is small with $np > 5$, a Poisson approximation is appropriate:

$$P(R \geq 23) \doteq P(Po(450 \times 0.04) \geq 25)$$
$$= P(Po(18) \geq 25)$$
$$= 0.0683 \quad \text{(from tables)} \quad 4$$

This exceeds 0.05; the test is not significant at the 5% level. We do not reject H_0. 2

10 Q9

Q10. S

a) (i) $\lambda_1 = \mu_B - \mu_A$,

(ii) $\lambda_2 = \mu_D - \mu_C$,

(iii) $\lambda_3 = \frac{1}{2}(\mu_A + \mu_B) - \frac{1}{2}(\mu_C + \mu_D)$. 3 (any multiples of these also fine)

The vectors of coefficients are $\underline{c}_1 = (-1, 1, 0, 0)$,
$\underline{c}_2 = (0, 0, -1, 1)$,
$\underline{c}_3 = (\frac{1}{2}, \frac{1}{2}, -\frac{1}{2}, -\frac{1}{2})$.

Since $\underline{c}_1 \cdot \underline{c}_2 = 0$, $\underline{c}_2 \cdot \underline{c}_3 = 0$, and $\underline{c}_1 \cdot \underline{c}_3 = 0$, the three contrasts are mutually orthogonal. 2

We estimate them as $\hat{\lambda}_1 = -73 + 73.25 = 0.250$,
$\hat{\lambda}_2 = 62.25 - 58.375 = 3.875$,
$\hat{\lambda}_3 = \frac{1}{2}(73 + 73.25) - \frac{1}{2}(58.375 + 62.250) = 12.8125$.

3

b) $L(\hat{\lambda}_1) = \frac{8 \times 0.250^2}{2} = 0.250$,

$L(\hat{\lambda}_2) = \frac{8 \times 3.875^2}{2} = 60.0625$,

$L(\hat{\lambda}_3) = \frac{8 \times 12.8125^2}{1} = 1313.281$.

The sum of these (1374 to the nearest integer) is the between groups sum of squares. 4

Q10. (cont.) c)

Source	SS	DF	MS	F	p-value
λ_1	0·25	1	0·25	0·00255	0·95 – 1·00
λ_2	60·0625	1	60·0625	0·6123	0·40 – 0·45
λ_3	1313·281	1	1313·281	13·387	0·001 – 0·005
Residuals	2747	28	98·1		

(1 mark) (1 mark) (1 mark) (3 marks) (3 marks)

Only the p-value corresponding to λ_3 is significant. We deduce:

- there is no significant difference between mean scores in the two "outstanding" schools.
- there is no significant difference between mean scores in the two "good" schools.
- there is a strongly significant (at the 0·5% level) difference between the two "good" schools' and the two "outstanding" schools' scores.

3

24 Q10

SECTION B

50