S: Seen similar
B: Bookwork
(as a methods course, nothing is truly of an unseen type, though of course questions become more demanding in Section B)

MA26620 : Applied Statistics

2024

Q1. a) Continuous, interval 2

S b) Discrete, ordinal 2

CLASSIFICATION c) (Effectively) continuous, ratio 2

d) Discrete, nominal 2

8 Q1

Q2. Let $N(t)$ denote the number of software freezes in t hours. 2

S Assume that freezes occur randomly at a constant rate, λ per hour say.

POISSON HYPOTHESIS TEST Then $N(t) \sim Po(\lambda t)$. 2

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

Under H_0, $N(8) \sim Po(12)$. Thus $p_0^+ = P(N(8) \geq 16) = P(Po(12) \geq 16) = 0.1556$ 2 from tables.

The test is not significant at the 10% level, so we do not reject H_0; the rate of freezing has not significantly increased. 2

10 Q2

Q3. a) 9 seeds (since $\overbrace{\text{(number of groups)}}^{4} \times \text{(number of observations per group)} - 1 = 35$) 2

S

ONE-WAY ANOVA b)

Source	SS	DF	MS	F	P
Between composts	45	3	15	5	0.00589
Within composts	96	32	3		
Total (corr.)	141	35			

7 (one for each highlighted cell)

c) $Y_{ij} \sim \mu_i + \varepsilon_{ij}$, $\varepsilon_{ij} \sim N(0, \sigma^2)$, ε_{ij} uncorrelated for $i = 1, \dots, 4$, $j = 1, \dots, 9$. 3

We test $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$ vs H_1: not all means are equal.

The test is significant at the 1% level; strong evidence against H_0 in favour of H_1. 2

14 Q3

Q4. Assume the outcome of opening each door is independent, with constant probability p of

S the figurine being a locomotive. Then, denoting the number of locomotives by L,

BINOMIAL HYPOTHESIS TEST $L \sim Bin(50, p)$. 3

We test $H_0: p = \frac{2}{3}$ vs $H_1: p < \frac{2}{3}$. Under H_0, $L \sim Bin(50, \frac{2}{3})$. 3

$p_0^- = P(Bin(50, \frac{2}{3}) \leq 25) = P(Bin(50, \frac{1}{3}) \geq 25) = 0.0108$.

The test is very nearly significant at the 1% level; strong evidence against H_0 in favor of H_1.

The rate of locomotives is lower than claimed. 4

10 Q4

Q5. Let X_i denote the boot-up time of phone i in seconds. 2

S

T CONFIDENCE INTERVAL

Assume $X_i \sim N(\mu, \sigma^2)$ for $i = 1, \ldots, 15$ independently. 2

Then the 95% confidence interval is given by

$$\bar{X} \pm t_{0.025[14]} \frac{S}{\sqrt{n}} = 18 \pm 2.1448 \times \frac{4}{\sqrt{15}} = 18 \pm 2.21514$$

2

$$= (15.78,\ 20.22) \text{ seconds.}$$

2

8 Q5

Q6. a) $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{32}{30} = \frac{16}{15} = 1.0\dot{6}.$

S

REGRESSION

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{2850}{50} - \frac{16}{15}\,\frac{1500}{50} = 25$$

$$\therefore\ y = \frac{16}{15}x + 25.$$

6

b) $R = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} = \frac{32}{\sqrt{30}\sqrt{50}} = 0.826 \Rightarrow R^2 = \frac{32^2}{1500} = \frac{256}{375} = 0.683.$

68.3% of the variance in crime rate is accounted for by its linear relationship with population density. 3

c) $\frac{16}{15} \times 25 + 25 = 51.\dot{6}$, so the predicted crime rate is 51.7 crimes per 100,000 people. 2

d) An increase in x of 3 increases y by $3\hat{\beta}_1 = 3 \times \frac{16}{15} = \frac{16}{5} = 3.2$. Thus the crime rate increases by 3.2 crimes per 100,000 people. 2

13 Q6

Q7. Note that $\sum_{i=1}^{k}\sum_{j=1}^{m}(Y_{ij} - Y_{\cdot\cdot})^2 = \sum_{i=1}^{k}\sum_{j=1}^{m}(Y_{ij} - Y_{i\cdot} + Y_{i\cdot} - Y_{\cdot\cdot})^2$ 3

B

SUM OF SQUARES DECOMPOSITION

$$= \sum_{i=1}^{k}\sum_{j=1}^{m}(Y_{ij} - Y_{i\cdot})^2 + \sum_{i=1}^{k}\sum_{j=1}^{m}(Y_{i\cdot} - Y_{\cdot\cdot})^2 + 2\sum_{i=1}^{k}(Y_{i\cdot} - Y_{\cdot\cdot})\sum_{j=1}^{m}(Y_{ij} - Y_{i\cdot})$$

2

$$= \sum_{i=1}^{k}\sum_{j=1}^{m}(Y_{ij} - Y_{i\cdot})^2 + m\sum_{i=1}^{k}(Y_{i\cdot} - Y_{\cdot\cdot})^2 + 2\sum_{i=1}^{k}(Y_{i\cdot} - Y_{\cdot\cdot})\left\{\sum_{j=1}^{m}(Y_{ij}) - m\frac{1}{m}\sum_{j=1}^{m}Y_{ij}\right\},$$

by definition of $Y_{i\cdot}$

and clearly the term in curly braces is zero, thus establishing the result.

2

7 Q7

SECTION A
70

Q8 S TWO-WAY ANOVA

a) $\sum_{i=1}^{2} \alpha_i = 0$, $\sum_{j=1}^{3} \beta_j = 0$, $\gamma_{i1} + \gamma_{i2} + \gamma_{i3} = 0 \quad \forall\ i \in \{1, 2\}$,
$\gamma_{1j} + \gamma_{2j} = 0 \quad \forall\ j \in \{1, 2, 3\}$. 2

b) $Y_{123} = 28$, $Y_{\cdot 2 \cdot} = 24.5$. 2

c) $\hat{\mu} = 23.167$.

$$\hat{\beta}_2 = Y_{\cdot 2 \cdot} - Y_{\cdot\cdot\cdot} = 24.5 - 23.167 = 1.333$$

$$\hat{\gamma}_{22} = Y_{22\cdot} - Y_{2\cdot\cdot} - Y_{\cdot 2 \cdot} + Y_{\cdot\cdot\cdot}$$
$$= 19 - 19.667 - 24.5 + 23.167$$
$$= -2.$$ 3

d) $A = 1$, $B = 2$, $C = 2$, $D = 12$, $E = 158.00$, $F = 6.170$. 3

e) $27.5 - 17.5 = 10$. This is an estimate of how many more "Cute Kittens" calendars are sold per week than "Cliff Richard" calendars. 2

Consider $Var(Y_{\cdot 1 \cdot} - Y_{\cdot 3 \cdot}) = \frac{\sigma^2}{6} + \frac{\sigma^2}{6}$ (by assumption of uncorrelated obs.)

$$\Rightarrow ESE(Y_{\cdot 1 \cdot} - Y_{\cdot 3 \cdot}) = \sqrt{\frac{\hat{\sigma}^2}{3}} = \sqrt{\frac{11.67}{3}} = 1.972.$$ 3

f) (i) No (0.000949)
(ii) No (0.000837)
(iii) No (0.14350) 3

18 Q8

Q9 S CHI-SQUARE TEST

Let Y denote the number of views per day. Assuming independent observations, we test H_0: Y has a Poisson distribution vs H_1: it does not.
We estimate the daily view rate as $\hat{\lambda} = \frac{\#\text{views}}{\#\text{days}} = \frac{108}{90} = 1.2$. 2
The expected cell values are therefore:

- $\hat{p}_0 = P(Y=0) = \frac{\hat{\lambda}^0 e^{-\hat{\lambda}}}{0!} = 0.301194 \Rightarrow \mathbb{E}[n_0] = 27.1075$
- $\hat{p}_1 = P(Y=1) = \frac{\hat{\lambda}^1 e^{-\hat{\lambda}}}{1!} = 0.361433 \Rightarrow \mathbb{E}[n_1] = 32.529$
- $\hat{p}_2 = P(Y=2) = \frac{\hat{\lambda}^2 e^{-\hat{\lambda}}}{2!} = 0.21686 \Rightarrow \mathbb{E}[n_2] = 19.5174$
- $\hat{p}_3 = P(Y=3) = \frac{\hat{\lambda}^3 e^{-\hat{\lambda}}}{3!} = 0.0867439 \Rightarrow \mathbb{E}[n_3] = 7.80695$
- $\hat{p}_4 = P(Y=4) = \frac{\hat{\lambda}^4 e^{-\hat{\lambda}}}{4!} = 0.0260232 \Rightarrow \mathbb{E}[n_4] = 2.34209$,

though since $\mathbb{E}[n_4] < 5$, we instead of $\hat{p}_4$ define $\hat{p}_{4+} = P(Y \geq 4) = 1 - P(Y \leq 3)$
$= 0.033769$,

which again leads to an expectation below 5, so we define $\hat{p}_{3+} = P(Y \geq 3) = 0.120513$, with corresponding expectation $\mathbb{E}[n_{3+}] = 10.8462$. 6

The chi-squared statistic is therefore

$$\chi^2_{obs} = \frac{(30-\mathbb{E}[n_0])^2}{\mathbb{E}[n_0]} + \frac{(30-\mathbb{E}[n_1])^2}{\mathbb{E}[n_1]} + \frac{(20-\mathbb{E}[n_2])^2}{\mathbb{E}[n_2]} + \frac{(10-\mathbb{E}[n_{3+}])^2}{\mathbb{E}[n_{3+}]}$$

$$= 0.583211,$$ 2

with $\chi^2 \sim \chi^2_{(2)}$, and from tables, the probability of $\chi^2_{(2)}$ exceeding 0.583211 lies between 70% and 75%. We therefore do not reject H_0; the data is well-described by a Poisson distribution. 2

12 Q9

Q10. S CONTRASTS

a) (i) $\lambda_1 = \frac{1}{2}(\mu_{Kittens} + \mu_{cars}) - \mu_{cliff}$

(ii) $\lambda_2 = \mu_{Kittens} - \mu_{cars}$ 2

λ_1 has coefficient vector $(\frac{1}{2}, \frac{1}{2}, -1)$, while λ_2 has $(1, -1, 0)$. Since $(\frac{1}{2}, \frac{1}{2}, -1)\cdot(1,-1,0) = \frac{1}{2} - \frac{1}{2} - 0 = 0$, the contrasts are orthogonal. 2

Estimates: $\hat{\lambda}_1 = \frac{1}{2}(27.5 + 24.5) - 17.5 = 8.5$

$\hat{\lambda}_2 = 27.5 - 24.5 = 3$ 2

b) $$L(\hat{\lambda}_1) = \frac{m\hat{\lambda}_1^2}{\sum_{i=1}^{3} c_i^2} = \frac{6 \times 8.5^2}{3/2} = 289.$$

$$L(\hat{\lambda}_2) = \frac{m\hat{\lambda}_2^2}{\sum_{i=1}^{3} c_i^2} = \frac{6 \times 3^2}{2} = 27.$$ 2

c)

	DF	Sum Sq	Mean Sq	F value	Pr(>F)
λ_1	1	289	289	8.594	< 0.05
λ_2	1	27	27	0.803	> 0.05
Residuals	15	504.5	33.63		

(from $F_{1,15}$ tables... upper 5% point is 4.5431) 3

We conclude that λ_1 significantly differs from zero, while λ_2 does not. Thus sales of Cliff Richard calendars differ significantly from the other two, whereas there is no significant difference between sales of "Cute Kittens" and "Supercars". 2

13 Q10

Q11. Assume ad-clicks are independent with constant probability p of being clicked.

S

BINOMIAL HYPOTHESIS TEST WITH APPROXIMATION

Then the number of ads clicked, N, out of 356 is distributed as $Bin(356, p)$.

We test $H_0 : p = 0.12$ vs $H_1 : p < 0.12$. 3

Calculate $p_0 = P(N \leq 14) = P(Bin(356, 0.12) \leq 24)$

$$\doteqdot P(N(42.72, 37.5936) < 24.5) \quad \text{(Normal approx. with cont. corr.)}$$

$$= P\left(N(0,1) < \frac{24.5 - 42.72}{\sqrt{37.5936}} = -2.9716\right)$$

$$= P(N(0,1) > 2.9716) \quad \text{(by symmetry)}$$

$$\doteqdot 0.0015 \quad \text{from tables.}$$ 3

The test is significant at the 0.5% level; strong evidence to reject H_0 in favour of H_1. The advertisement is performing less strongly than claimed.

1

7 Q11

SECTION B
50

PAPER TOTAL : 120