Maths with Matt

Physics and maths resources

Edexcel A Level Further Mathematics

9FM0/4B - Further Statistics 2

Tuesday 28 June 2022

Questions and Fully Worked Model Solutions

This booklet places the **question and the solution together**. The mathematics is set out using proper fractions, powers, roots and exact notation, with extra written explanation so that a student can see *why* each step is being taken.

Solutions have been checked against the Pearson Edexcel Summer 2022 mark scheme. The aim is a clear **full-mark model response**, not merely a list of final answers.

Question wording is reproduced from the user-provided 2022 paper for study/reference. Pearson retains copyright in the original examination material.

Contents

Question 1 2

Question 2 4

Question 3 5

Question 4 6

Question 5 7

Question 6 8

Question 7 10

Question 8 11

www.mathswithmatt.co.uk

Question 1

Question 1

Kwame fits the least-squares regression line $y = 45.5t + 2080$ to 30 years of tea-yield data. The summary statistics are RSS $= 1666567$, $S_{tt} = 52.0$ and $S_{yy} = 1774155$. Answer parts (a)–(e) concerning prediction, residuals and the suitability of the model.

Model solution

(a) A prediction from a regression line is found by substituting the given explanatory-variable value into the line.

$$\hat{y} = 45.5(20) + 2080 = 2990.$$

The predicted tea yield is 2990 *kg/hectare.*

(1)

(b) The dependent variable is the tea yield, so its residual measures the vertical error made by the fitted line.

A residual is the observed tea yield minus the tea yield predicted by the regression model.

(1)

(c) RSS is the unexplained part of the variation in y. Thus $r^2 = 1 - \text{RSS}/S_{yy}$; the positive gradient tells us to take the positive square root.

$$r^2 = 1 - \frac{1666567}{1774155} = 0.06053\ldots, \qquad r = \sqrt{0.06053\ldots} = 0.246\ldots$$

$$r \approx 0.246.$$

(2)

(d)(i) A useful linear model needs a reasonably strong linear association. Here r is close to zero, so the linear association is weak.

The small value of r indicates weak linear correlation, so a linear model may not fi

(1)

(d)(ii) For an appropriate linear model, residuals should be randomly scattered about zero with broadly constant spread. In the plot, particularly above $20°C$, they show a pattern rather than random scatter about zero.

The residual plot suggests a systematic pattern (and non-constant behaviour), rather than random scatter about zero.

(1)

(e) RSS is measured in the square of the response variable's units. The two response variables, rainfall and yield, have different units, so their RSS values are not directly comparable.

No. The RSS values have different units (mm^2 and $(kg/hectare)^2$), so their sizes can

(1)

www.mathswithmatt.co.uk

Question 2

Question 2

Independent samples give $n_Y = 120$, $\overline{Y} = 57.2$, $\sigma_Y = 1.2$ for yellow tennis balls and $n_W = 140$, $\overline{W} = 56.9$, $\sigma_W = 0.9$ for white tennis balls, with weights in grams. Find a confidence interval and test Jamie's claim that the population mean yellow-ball weight is larger.

Model solution

(a) The population standard deviation is known, so a Normal z-interval for the population mean is appropriate. For 95% confidence, $z = 1.96$.

$$57.2 \pm 1.96\frac{1.2}{\sqrt{120}} = 57.2 \pm 0.21466\ldots$$

A 95% confidence interval is $(56.985\ldots, 57.415\ldots)$, i.e. $(57.0, 57.4)$ g.

(3)

(b)(i) Under the null hypothesis the two population means are equal, so the mean of $\overline{Y} - \overline{W}$ is zero. Independent sample means have variances σ_Y^2/n_Y and σ_W^2/n_W, which add for a difference.

$$\overline{Y} - \overline{W} \sim N\left(0, \frac{1.2^2}{120} + \frac{0.9^2}{140}\right) = N\left(0, \frac{249}{14000}\right).$$

$\overline{Y} - \overline{W} \sim N(0, 249/14000)$ under H_0.

(3)

(b)(ii) Although the population distributions are not stated, both samples are large. The central limit theorem therefore makes their sample means, and hence their difference, approximately Normal.

The large sample sizes justify the approximate Normal model by the central limit theorem.

(1)

(c) Jamie's wording "greater than" gives a one-tailed test. Use the known-standard-deviation test statistic for the difference of independent sample means.

$$H_0 : \mu_Y = \mu_W, \qquad H_1 : \mu_Y > \mu_W,$$

$$z = \frac{57.2 - 56.9}{\sqrt{1.2^2/120 + 0.9^2/140}} = 2.249\ldots$$

At the 5% upper-tail level, the critical value is 1.6449. Since $2.249\ldots > 1.6449$, the statistic is in the critical region.

Reject H_0. There is significant evidence at the 5% level that the mean weight of yello

(5)

Question 3

Question 3

$X \sim N(5, 0.4^2)$ and $Y \sim N(8, 0.1^2)$ are independent. For independent $X_1, \ldots, X_a$ from X, $W = X_1 + \cdots + X_a + bY \sim N(169, 2^2)$. Find a and b.

Model solution

Independence means means and variances add. Equating the given mean and variance of W supplies two equations for a and b.

$$E(W) = 5a + 8b = 169, \qquad Var(W) = 0.16a + 0.01b^2 = 4.$$

Eliminate a using $a = (169 - 8b)/5 = 33.8 - 1.6b$.

$$0.16(33.8 - 1.6b) + 0.01b^2 = 4$$

$$0.01b^2 - 0.256b + 1.408 = 0.$$

$$b = 8 \text{ or } 17.6.$$

The corresponding a-values are 21 and 5.64. Since a is a sample size, it must be a positive integer.

$$a = 21 \text{ and } b = 8.$$

(6)

Question 4

Question 4

For seven patients, resting heart rates before the exercise programme are $65, 68, 77, 79, 80, 88, 92$, and after are $63, 65, 73, 76, 80, 84, 80$. Test at 5% whether the programme reduces resting heart rate, and state the required assumption.

Model solution

(a) Each after reading is paired with the same patient's before reading, so analyse the seven differences rather than treating the samples as independent. Let $d = \text{before} - \text{after}$, *so a reduction corresponds to* $\mu_d > 0$.

$$d: \quad 2,\ 3,\ 4,\ 3,\ 0,\ 4,\ 12, \qquad \bar{d} = \frac{28}{7} = 4.$$

$$s_d^2 = \frac{\sum d^2 - 7\bar{d}^2}{7-1} = \frac{198 - 7(4)^2}{6} = 14.333\ldots, \qquad s_d = 3.7859\ldots$$

$$H_0 : \mu_d = 0, \qquad H_1 : \mu_d > 0, \qquad t = \frac{\bar{d}}{s_d/\sqrt{7}} = \frac{4}{3.7859\ldots/\sqrt{7}} = 2.795\ldots$$

With 6 *degrees of freedom, the 5% one-tailed critical value is* 1.943. *The observed* t*-value exceeds this.*

RejectH$_0$.Thereissufficientevidenceatthe5%levelthattheexerciseprogrammere

(7)

(b) The paired t*-test is applied to the differences, so it is their distribution that needs the Normal assumption.*

The population of before-minus-after resting-heart-rate differences is Normally distributed.

(1)

Question 5

Question 5

At Site A, $n_A = 13$ and $s_A^2 = 6.39$; at Site B, $n_B = 9$. A two-sided 2% test compares the two population variances. State an assumption, determine values of s_B^2 that reject H_0, and find a 99% confidence interval for σ_A^2.

Model solution

(a) Both the F variance test and the chi-squared variance interval rely on normal sampling distributions.

For each site, pollutant concentration must follow a Normal distribution.

(1)

(b) Under $H_0 : \sigma_A^2 = \sigma_B^2$, $s_A^2/s_B^2 \sim F_{12,8}$. ***A two-sided 2% test has 1% in each tail.***

$$F_{12,8;0.01} = 5.67, \qquad F_{8,12;0.01} = 4.50.$$

Reject in either tail:

$$\frac{6.39}{s_B^2} > 5.67 \quad \textit{or} \quad \frac{s_B^2}{6.39} > 4.50.$$

$$s_B^2 < \frac{6.39}{5.67} = 1.12698\ldots \quad \textit{or} \quad s_B^2 > 4.50(6.39) = 28.755.$$

$$\text{Reject} H_0 \text{when} s_B^2 < 1.13 \text{or} s_B^2 > 28.8.$$

(4)

(c) For a Normal population, $(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}$. *Here there are 12 degrees of freedom; using the 0.5% and 99.5% chi-squared critical values gives a central 99% interval.*

$$\chi^2_{12,0.005} = 28.3, \qquad \chi^2_{12,0.995} = 3.074,$$

$$\frac{12(6.39)}{28.3} < \sigma_A^2 < \frac{12(6.39)}{3.074}.$$

$$2.71 < \sigma_A^2 < 24.9 (\text{micrograms/m}^3)^2.$$

(3)

Question 6

Question 6

For independent observations $X_1, \ldots, X_n$ from a distribution with mean μ and variance σ^2,

$$K = \frac{2}{n(n+1)} \sum_{r=1}^{n} rX_r, \qquad L = \frac{X_1 + X_2}{3} + \frac{X_3 + \cdots + X_n}{3(n-2)}.$$

Show both estimators are unbiased, find their variances, and decide which is better for large n.

Model solution

(a) Linearity of expectation and $E(X_r) = \mu$ allow each estimator to be checked against the definition of unbiasedness.

$$E(K) = \frac{2\mu}{n(n+1)} \sum_{r=1}^{n} r = \frac{2\mu}{n(n+1)} \cdot \frac{n(n+1)}{2} = \mu.$$

$$E(L) = \frac{2\mu}{3} + \frac{(n-2)\mu}{3(n-2)} = \frac{2\mu}{3} + \frac{\mu}{3} = \mu.$$

$$E(K) = E(L) = \mu, \text{so both K and L are unbiased.}$$

(5)

(b)(i) The observations are independent, so variances add; also $\mathrm{Var}(aX) = a^2 \mathrm{Var}(X)$.

$$\mathrm{Var}(K) = \frac{4}{n^2(n+1)^2} \sum_{r=1}^{n} r^2\sigma^2 = \frac{4\sigma^2}{n^2(n+1)^2} \cdot \frac{n(n+1)(2n+1)}{6} = \frac{2(2n+1)\sigma^2}{3n(n+1)}.$$

$$\mathrm{Var}(K) = \frac{2(2n+1)\sigma^2}{3n(n+1)}.$$

(4)

(b)(ii) The first two observations have coefficient $1/3$, while each remaining observation has coefficient $1/[3(n-2)]$.

$$\mathrm{Var}(L) = 2\left(\frac{1}{3}\right)^2 \sigma^2 + (n-2)\left(\frac{1}{3(n-2)}\right)^2 \sigma^2 = \frac{\sigma^2}{9(n-2)} + \frac{2\sigma^2}{9} = \frac{(2n-3)\sigma^2}{9(n-2)}.$$

$$\mathrm{Var}(L) = \frac{(2n-3)\sigma^2}{9(n-2)}.$$

(3)

(c) Since both estimators are unbiased, the one with smaller variance is better. Compare their limiting variances as the sample size becomes large.

$$\mathrm{Var}(K) \longrightarrow 0, \qquad \mathrm{Var}(L) \longrightarrow \frac{2\sigma^2}{9}.$$

Korhanwins : forlargen, Khasthesmallervariance.

(3)

www.mathswithmatt.co.uk

Question 7

Question 7

A rectangle has area 40 cm^2. Its length L has a continuous uniform distribution on $[4, 10]$. Find the expected perimeter using algebraic integration.

Model solution

The fixed area gives width $40/L$, so first express the perimeter in terms of the random length. A uniform distribution on an interval of width 6 has density $1/6$.

$$P = 2\left(L + \frac{40}{L}\right) = 2L + \frac{80}{L}, \qquad f_L(l) = \frac{1}{6} \quad (4 \leqslant l \leqslant 10).$$

$$E(P) = \frac{1}{6}\int_4^{10}\left(2l + \frac{80}{l}\right)\,dl = \frac{1}{6}\left[l^2 + 80\ln|l|\right]_4^{10}.$$

$$E(P) = \frac{1}{6}\{100 + 80\ln 10 - (16 + 80\ln 4)\} = 14 + \frac{40}{3}\ln(2.5) = 26.217\ldots$$

The expected perimeter is 26.2 *cm* (3 s.f.).

(7)

Question 8

Question 8

$$F_X(x) = \begin{cases} 0, & x < 1, \\ 1.5x - 0.25x^2 - 1.25, & 1 \leqslant x \leqslant 3, \\ 1, & x > 3. \end{cases}$$

Find the median, a conditional probability, the cdf of $Y = 1/X$, and the mode of Y.

Model solution

(a) A median m satisfies $F_X(m) = 0.5$. The middle branch applies because the support is from 1 to 3.

$$1.5m - 0.25m^2 - 1.25 = 0.5 \implies m^2 - 6m + 7 = 0 \implies m = 3 \pm \sqrt{2}.$$

Only $3 - \sqrt{2}$ lies in $[1, 3]$.

$$\text{The median is } 3 - \sqrt{2}.$$

(2)

(b) Conditional probability restricts the event $X > 1.2$, so divide the probability of the overlap by the probability of this condition.

$$\Pr(X < 1.6 \mid X > 1.2) = \frac{\Pr(1.2 < X < 1.6)}{\Pr(X > 1.2)} = \frac{F_X(1.6) - F_X(1.2)}{1 - F_X(1.2)}.$$

$$F_X(1.6) = 0.51, \qquad F_X(1.2) = 0.19, \qquad \frac{0.51 - 0.19}{1 - 0.19} = \frac{32}{81}.$$

$$\Pr(X < 1.6 \mid X > 1.2) = \frac{32}{81}.$$

(3)

(c) Since $X > 0$, taking reciprocals reverses the inequality. Thus $Y \leqslant y$ is equivalent to $X \geqslant 1/y$. The range $1 \leqslant X \leqslant 3$ becomes $1/3 \leqslant Y \leqslant 1$.

$$F_Y(y) = \Pr(Y \leqslant y) = 1 - F_X(1/y) = 1 - \left(\frac{1.5}{y} - \frac{0.25}{y^2} - 1.25\right) = 2.25 - \frac{1.5}{y} + \frac{0.25}{y^2}.$$

$$F_Y(y) = \begin{cases} 0, & y < \frac{1}{3}, \\ 2.25 - \frac{1.5}{y} + \frac{0.25}{y^2}, & \frac{1}{3} \leqslant y \leqslant 1, \\ 1, & y > 1. \end{cases}$$

(4)

(d) The mode maximises the density. Differentiate the cdf to obtain the density, then set its derivative to zero.

$$f_Y(y) = \frac{1.5}{y^2} - \frac{0.5}{y^3}, \qquad f'_Y(y) = -\frac{3}{y^3} + \frac{1.5}{y^4}.$$

$$-\frac{3}{y^3} + \frac{1.5}{y^4} = 0 \implies -3y + 1.5 = 0 \implies y = 0.5.$$

Also $f''_Y(0.5) < 0$*, so this stationary point is a maximum.*

The mode of Y is 0.5.

(3)

www.mathswithmatt.co.uk