Maths with Matt

Physics and maths resources

Edexcel A Level Further Mathematics

9FM0/4B - Further Statistics 2

Monday 26 June 2023

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 2023 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 2023 paper for study/reference. Pearson retains copyright in the original examination material.

Contents

Question 1 2

Question 2 3

Question 3 5

Question 4 6

Question 5 7

Question 6 8

Question 7 9

Question 8 10

www.mathswithmatt.co.uk

Question 1

Question 1

For 20 children, the summary statistics for 100 m time x seconds and 500 m time y seconds are

$$S_{xx} = 314.55, \quad S_{yy} = 9026, \quad S_{xy} = 1610, \quad \bar{x} = 19.65, \quad \bar{y} = 108.$$

(a) Calculate the product moment correlation coefficient. (b) Show that the regression line of y on x is $y = 5.12x + 7.42$, with coefficients to 3 s.f. A child with $x = 20$ has $y = 104$. (c) Find the residual. (d) In increasing order of x, the residual signs are $+,+,+,+,-,-,+,-,-,-,-,-,-,-,+,+,+,+,+$. Explain what this says about predictions for the fastest and slowest children.

Model solution

(a) The product moment correlation coefficient measures the strength of the linear relationship, so use the three supplied corrected sums of squares and products.

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{1610}{\sqrt{314.55 \times 9026}} = 0.9555\ldots$$

$$r = 0.956 (3\text{s.f.}).$$

(2)

(b) For the regression of y on x, the gradient is S_{xy}/S_{xx}. The line must pass through $(\bar{x}, \bar{y})$, which gives its intercept.

$$b = \frac{1610}{314.55} = 5.1184\ldots, \qquad a = \bar{y} - b\bar{x} = 108 - (5.1184\ldots)(19.65) = 7.423\ldots$$

$$y = 5.12x + 7.42.$$

$$y = 5.12x + 7.42.$$

(3)

(c) A residual is observed value minus the value predicted by the regression line.

$$e = 104 - \{5.12(20) + 7.42\} = 104 - 109.82 = -5.82.$$

$$\text{Theresidualis} - 5.82\text{seconds}(\text{approximately} - 5.8\text{seconds}).$$

(1)

(d) A positive residual means actual y is greater than predicted y. The first and last entries, representing the fastest and slowest children over 100 m, are positive.

The model underestimates the 500 m times of both the fastest and the slowest children.

(1)

Question 2

Question 2

Green-apple weights have known standard deviation 3.5 g. A random sample of 80 has mean 128 g. (a) Find a 98% confidence interval for the population mean. Camilo claims that the green population mean exceeds the red population mean by more than 10 g. A sample of n red apples has mean 117 g and known standard deviation 4 g. A 5% test is used. (b) State hypotheses. (c) Find the smallest n that rejects H_0. (d) Explain the relevance of the CLT. (e) State the conclusion when $n = 85$.

Model solution

(a) The population standard deviation is known, so a normal critical value is used. A two-sided 98% interval leaves 1% in each tail, giving $z = 2.3263$.

$$128 \pm 2.3263\left(\frac{3.5}{\sqrt{80}}\right) = 128 \pm 0.91036\ldots$$

$$(127.09,\ 128.91)\text{g}.$$

(3)

(b) Let μ_G *and* μ_R *be the population means. The claim "more than 10 g greater" determines a right-tailed alternative.*

$$H_0 : \mu_G - \mu_R = 10, \qquad H_1 : \mu_G - \mu_R > 10.$$

$$H_0 : \mu_G - \mu_R = 10, H_1 : \mu_G - \mu_R > 10.$$

(1)

(c) The difference of independent sample means is approximately normal. Under H_0 *its mean is 10 and its variance is the sum of the two sampling variances.*

$$\overline{G} - \overline{R} \sim N\left(10, \frac{3.5^2}{80} + \frac{4^2}{n}\right).$$

The observed difference is $128 - 117 = 11$. *For a 5% upper-tail test, reject when the standardised value exceeds* 1.6449.

$$\frac{11 - 10}{\sqrt{12.25/80 + 16/n}} > 1.6449$$

$$\frac{1}{1.6449^2} > \frac{12.25}{80} + \frac{16}{n} \quad \Longrightarrow \quad n > 73.9\ldots$$

$$\text{Thesmallestintegersamplesizeis} n = 74.$$

(6)

(d) Both samples are large. Therefore, even without knowing the distributions of individual apple weights, the CLT makes the sample means, and hence their difference, approximately normally distributed.

The normal approximations used in parts (a) and (c) are justified by the Central Limit Theorem.

(1)

(e) Since $85 > 73.9\ldots$*, the observed difference is significant at the 5% level.*

$\text{Reject} H_0; \text{there is sufficient evidence that green apples have a mean weight more th}$

(1)

www.mathswithmatt.co.uk

Question 3

Question 3

Amounts dispensed by machines A and B are normally distributed. For A, $n = 9$, $\sum x = 2268$ and $\sum x^2 = 571700$. (a) Find a 95% confidence interval for the variance for A. For B, $n = 11$ and sample variance 12.7 ml^2. (b) At the 10% level, test whether the variances differ; state hypotheses and critical value.

Model solution

(a) Normality means that a confidence interval for a variance is based on the chi-squared distribution. First find the unbiased sample variance.

$$\bar{x} = \frac{2268}{9} = 252, \qquad s_A^2 = \frac{571700 - 9(252)^2}{9-1} = 20.5.$$

With 8 *degrees of freedom,* $\chi^2_{8,0.025} = 17.535$ *and* $\chi^2_{8,0.975} = 2.180$. *Thus*

$$\frac{8(20.5)}{17.535} < \sigma_A^2 < \frac{8(20.5)}{2.180}$$

$$9.35 < \sigma_A^2 < 75.23 \ ml^2.$$

(4)

(b) This is a two-tailed comparison of two normal-population variances, so use an F *statistic, putting the larger sample variance in the numerator.*

$$H_0 : \sigma_A^2 = \sigma_B^2, \qquad H_1 : \sigma_A^2 \neq \sigma_B^2,$$

$$F = \frac{s_A^2}{s_B^2} = \frac{20.5}{12.7} = 1.614\ldots$$

At the 10% two-tailed level, the upper critical value is $F_{8,10;0.95} = 3.07$ *(approximately). Since* $1.614 < 3.07$, *it is not in the rejection region.*

Do not reject H_0; there is insufficient evidence that the machine variances are diff

(4)

Question 4

Question 4

Egg weights $E \sim N(60, 3^2)$, empty small-box weights $S \sim N(24, 1.8^2)$ and empty large-box weights $L \sim N(40, 2.1^2)$. A small box contains 6 randomly selected eggs and a large box contains 12. (a) Find P(small box weighs < 387). (b) Find the probability that a randomly selected large box weighs more than twice a randomly selected small box.

Model solution

(a) The total is a sum of independent normal variables, so it is normal: means add and variances add. Let W be the small-box total.

$$W = E_1 + \cdots + E_6 + S \sim N\left(6(60) + 24,\ 6(3^2) + 1.8^2\right) = N(384, 57.24).$$

$$P(W < 387) = P\left(Z < \frac{387 - 384}{\sqrt{57.24}}\right) = 0.6541\ldots$$

$$P(W < 387) = 0.654.$$

(3)

(b) Let X be the large-box total. Compare X with $2W$ by forming one normal difference.

$$X \sim N\left(12(60) + 40,\ 12(3^2) + 2.1^2\right) = N(760, 112.41),$$

$$2W \sim N(2 \times 384,\ 2^2 \times 57.24) = N(768, 228.96).$$

$$X - 2W \sim N(760 - 768,\ 112.41 + 228.96) = N(-8, 341.37).$$

$$P(X > 2W) = P(X - 2W > 0) = P\left(Z > \frac{0 - (-8)}{\sqrt{341.37}}\right) = 0.3325\ldots$$

$$0.333.$$

(5)

Question 5

Question 5

Eight people have percentages recalled before training $24, 33, 33, 39, 30, 38, 32, 34$ and after training $28, 30, 37, 41, 32, 44, 35, 34$. (a) State why a paired t-test is suitable. (b) State an assumption needed. (c) At 5%, test for an increase after training, stating hypotheses, test statistic and critical value.

Model solution

(a) Each before score is paired with an after score from the same person, so the two samples are not independent.

A paired t-test is appropriate because the before and after scores are matched for each person

(1)

(b) The test is performed on the within-person differences.

Assume that the differences in scores are normally distributed.

(1)

(c) Define $d = \text{after} - \text{before}$*, so a positive mean difference represents the claimed improvement.*

$$d: \quad 4,\ -3,\ 4,\ 2,\ 2,\ 6,\ 3,\ 0.$$

$$\bar{d} = \frac{18}{8} = 2.25, \qquad s_d^2 = 7.642857\ldots, \qquad s_d = 2.76457\ldots$$

$$H_0 : \mu_d = 0, \qquad H_1 : \mu_d > 0.$$

Under H_0*, the paired-test statistic has a* t *distribution with* $8 - 1 = 7$ *degrees of freedom.*

$$t = \frac{\bar{d} - 0}{s_d/\sqrt{8}} = \frac{2.25}{2.76457\ldots/\sqrt{8}} = 2.3019\ldots$$

For a 5% upper-tailed test, $t_{7,0.95} = 1.895$*. Since* $2.302 > 1.895$*, reject* H_0*.*

There is sufficient evidence that the training increases the percentage of words recalled.

(7)

Question 6

Question 6

The continuous random variable X has cdf

$$F(x) = \begin{cases} 0, & x < 0, \\ k(x - ax^2), & 0 \leqslant x \leqslant 4, \\ 1, & x > 4, \end{cases}$$

where $a, k > 0$ and $P(X < 2) = \frac{2}{3}$. (a) Find the exact median. (b) Find the pdf. (c) Hence state the mode, with a reason.

Model solution

(a) Use the two given cdf facts: $F(2) = 2/3$, *and because the cdf reaches 1 at the upper end of its support,* $F(4) = 1$. *These determine the two unknown constants.*

$$k(2 - 4a) = \frac{2}{3}, \qquad k(4 - 16a) = 1.$$

$$\frac{2 - 4a}{4 - 16a} = \frac{2}{3} \quad \Longrightarrow \quad 3(2 - 4a) = 2(4 - 16a) \quad \Longrightarrow \quad a = \frac{1}{10},$$

$$k\left(4 - 16 \times \frac{1}{10}\right) = 1 \quad \Longrightarrow \quad k = \frac{5}{12}.$$

The median m *satisfies* $F(m) = 1/2$.

$$\frac{5}{12}\left(m - \frac{m^2}{10}\right) = \frac{1}{2} \quad \Longrightarrow \quad m^2 - 10m + 12 = 0$$

$$m = 5 \pm \sqrt{13}.$$

Only $5 - \sqrt{13}$ *lies in the support* $[0, 4]$.

$$\text{The median is } 5 - \sqrt{13}.$$

(6)

(b) The pdf is the derivative of the cdf on its continuous range.

$$f(x) = \begin{cases} \frac{5}{12}\left(1 - \frac{x}{5}\right), & 0 \leqslant x \leqslant 4, \\ 0, & \textit{otherwise}. \end{cases}$$

$$f(x) = \tfrac{5}{12}(1 - x/5) \text{ for } 0 \leqslant x \leqslant 4, \text{ and } 0 \text{ otherwise.}$$

(2)

(c) On $0 \leqslant x \leqslant 4$, *the pdf is a line with negative gradient, so its greatest value occurs at the left endpoint.*

$$\text{The mode is } 0.$$

(2)

Question 7

Question 7

$R \sim U[2,10]$. A sphere has radius R cm, surface area $S = 4\pi R^2$ cm^2 and volume $V = \frac{4}{3}\pi R^3$ cm^3. (a) Write $f_R(r)$. (b) Show that $E(S) = \frac{496\pi}{3}$. (c) Find $E(V)$ by algebraic integration.

Model solution

(a) A continuous uniform distribution has constant density $1/(10-2)$ over its interval.

$$f_R(r) = \begin{cases} \frac{1}{8}, & 2 \leqslant r \leqslant 10, \\ 0, & \textit{otherwise}. \end{cases}$$

$$f_R(r) = \tfrac{1}{8} \text{for} 2 \leqslant r \leqslant 10, \text{and} 0 \text{otherwise}.$$

(1)

(b) Since area depends on R^2, calculate $E(R^2)$. Using $E(R^2) = \text{Var}(R) + [E(R)]^2$ is convenient.

$$E(R) = \int_2^{10} r\left(\frac{1}{8}\right)\,dr = 6, \qquad \text{Var}(R) = \frac{(10-2)^2}{12} = \frac{16}{3}.$$

$$E(S) = 4\pi E(R^2) = 4\pi\left\{\frac{16}{3} + 6^2\right\} = \frac{496\pi}{3}.$$

$$E(S) = \tfrac{496\pi}{3} \textit{ cm}^2.$$

(4)

(c) Here the required power is R^3, so integrate $\frac{4}{3}\pi r^3$ against the uniform density.

$$E(V) = \int_2^{10} \frac{4}{3}\pi r^3\left(\frac{1}{8}\right)\,dr = \frac{\pi}{6}\left[\frac{r^4}{4}\right]_2^{10} = \frac{\pi}{24}(10^4 - 2^4) = 416\pi.$$

$$E(V) = 416\pi \textit{ cm}^3.$$

(4)

Question 8

Question 8

Three independent random samples of size n are taken from a large bag with yellow proportion p, giving Y_1, Y_2, Y_3. The estimators are

$$\widehat{p}_1 = \frac{Y_1 + 3Y_2 - 2Y_3}{2n}, \qquad \widehat{p}_2 = \frac{2Y_1 + 3Y_2 + Y_3}{6n}.$$

(a) Show both are unbiased. (b) Find $\mathrm{Var}(\widehat{p}_1)$. Given $\mathrm{Var}(\widehat{p}_2) = \frac{7p(1-p)}{18n}$, (c) identify the better estimator. (d) For $\widehat{p}_3 = \frac{Y_1 + aY_2 + 3Y_3}{bn}$, where a, b are positive integers, find a, b so that it is unbiased and better than both previous estimators.

Model solution

(a) Each count has distribution $Y_i \sim B(n, p)$*, so* $E(Y_i) = np$*. Apply linearity of expectation.*

$$E(\widehat{p}_1) = \frac{np + 3np - 2np}{2n} = p, \qquad E(\widehat{p}_2) = \frac{2np + 3np + np}{6n} = p.$$

Both $\widehat{p}_1$ and $\widehat{p}_2$ are unbiased estimators of p.

(3)

(b) The samples are independent, so variances of the weighted counts add. Also $\mathrm{Var}(Y_i) = np(1-p)$.

$$\mathrm{Var}(\widehat{p}_1) = \frac{np(1-p) + 9np(1-p) + 4np(1-p)}{(2n)^2} = \frac{7p(1-p)}{2n}.$$

$$\mathrm{Var}(\widehat{p}_1) = \frac{7p(1-p)}{2n}.$$

(2)

(c) Both estimators are unbiased, so the one with the smaller variance is more precise.

$$\frac{7p(1-p)}{18n} < \frac{7p(1-p)}{2n}.$$

$\widehat{p}_2$ is the better estimator because it has the smaller variance.

(2)

(d) First impose unbiasedness. Then compare the variance with that of $\widehat{p}_2$*, the better of the two existing estimators.*

$$E(\widehat{p}_3) = \frac{np + anp + 3np}{bn} = \frac{a+4}{b}p.$$

For $E(\widehat{p}_3) = p$, $\qquad b = a + 4.$

$$\mathrm{Var}(\widehat{p}_3) = \frac{(1+a^2+9)np(1-p)}{b^2n^2} = \frac{(a^2+10)p(1-p)}{b^2n}.$$

For $\widehat{p}_3$ to improve on $\widehat{p}_2$, cancel the common positive factor $p(1-p)/n$ and require

$$\frac{a^2+10}{b^2} < \frac{7}{18}.$$

$$18(a^2+10) < 7(a+4)^2 \quad \Longrightarrow \quad 11a^2 - 56a + 68 < 0.$$

$$11a^2 - 56a + 68 = (11a - 34)(a - 2),$$

$$2 < a < \frac{34}{11},$$

so the only positive integer in this interval is $a = 3$. Hence $b = a + 4 = 7$.

$$a = 3,\ b = 7.$$

(5)

www.mathswithmatt.co.uk