Please check the examination details below before entering your candidate information

Candidate surname StemUp Model Solution | Other names

Centre Number | Candidate Number

Pearson Edexcel Level 3 GCE

Time 1 hour 30 minutes | **Paper reference** **9FM0/4B**

Further Mathematics

Advanced

PAPER 4B: Further Statistics 2

You must have:
Mathematical Formulae and Statistical Tables (Green), calculator

Total Marks

Candidates may use any calculator permitted by Pearson regulations. Calculators must not have the facility for symbolic algebra manipulation, differentiation and integration, or have retrievable mathematical formulae stored in them.

Instructions

- Use **black** ink or ball-point pen.
- If pencil is used for diagrams/sketches/graphs it must be dark (HB or B).
- **Fill in the boxes** at the top of this page with your name, centre number and candidate number.
- Answer **all** questions and ensure that your answers to parts of questions are clearly labelled.
- Answer the questions in the spaces provided – *there may be more space than you need.*
- You should show sufficient working to make your methods clear. Answers without working may not gain full credit.
- Values from statistical tables should be quoted in full. If a calculator is used instead of the tables the value should be given to an equivalent degree of accuracy.
- Inexact answers should be given to three significant figures unless otherwise stated.

Information

- A booklet 'Mathematical Formulae and Statistical Tables' is provided.
- There are 8 questions in this question paper. The total mark for this paper is 75.
- The marks for **each** question are shown in brackets – *use this as a guide as to how much time to spend on each question.*

Advice

- Read each question carefully before you start to answer it.
- Try to answer every question.
- Check your answers if you have time at the end.

Turn over ▶

P72096A

©2022 Pearson Education Ltd.

Q:1/1/1/

P 7 2 0 9 6 A 0 1 2 4

Pearson

1. Kwame is investigating a possible relationship between average March temperature, t °C, and tea yield, y kg/hectare, for tea grown in a particular location.
He uses 30 years of past data to produce the following summary statistics for a linear regression model, with tea yield as the dependent variable.

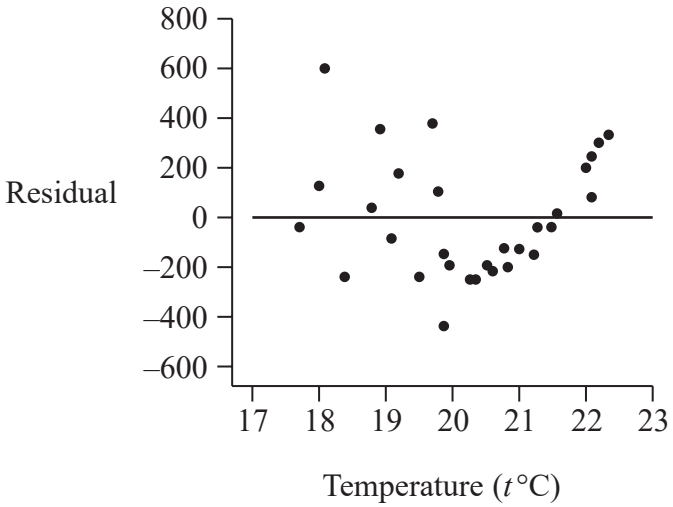
Residual Sum of Squares (RSS) = 1 666 567 $\quad S_{tt} = 52.0 \quad S_{yy} = 1\,774\,155$

least squares regression line: gradient = 45.5 $\quad y$-intercept = 2080

(a) Use the regression model to predict the tea yield for an average March temperature of 20 °C

(1)

He also produces the following residual plot for the data.



(b) Explain what you understand by the term residual.

(1)

(c) Calculate the product moment correlation coefficient between t and y

(2)

(d) Explain why the linear model may not be a good fit for the data

(i) with reference to your answer to part (c)

(ii) with reference to the residual plot.

(2)

Question 1 continues on page 4

(a) $y = 2080 + 45.5 \times 20 = 2990$ kg/hectare

(b) A residual is the difference between the observed value of y and the value of y predicted by the regression line, i.e. $y - \hat{y}$

DO NOT WRITE IN THIS AREA
DO NOT WRITE IN THIS AREA
DO NOT WRITE IN THIS AREA

Question 1 continued

(c) $RSS = S_{yy}(1 - r^2)$

$1666567 = 1774155(1 - r^2)$

$r^2 = 1 - \frac{1666567}{1774155} = 0.0606418...$

$r = \sqrt{0.0606418...} = 0.246255...$ (positive since gradient > 0)

$r = 0.246$

(d) (i) $r = 0.246$ is close to 0, so there is only weak linear correlation between t and y, so a linear model may not fit well.

(ii) The residuals are not randomly scattered about 0: for $t > 20$ they follow a pattern, increasing from negative to positive (a curve), and the spread of the residuals is not the same either side of $t = 20$.
This suggests a non-linear relationship may be more appropriate.

DO NOT WRITE IN THIS AREA
DO NOT WRITE IN THIS AREA
DO NOT WRITE IN THIS AREA

Question 1 continued

Kwame also collects data on total March rainfall, w mm, for each of these 30 years.

For a linear regression model of w on t the following summary statistic is found.

Residual Sum of Squares (RSS) = 86 754

Kwame concludes that since this model has a smaller RSS, there must be a stronger linear relationship between w and t than between y and t (where RSS = 1 666 567)

(e) State, giving a reason, whether or not you agree with the reasoning that led to Kwame's conclusion.

(1)

(e) I do not agree. The RSS values cannot be compared because they are measured in different units (kg/hectare squared and mm squared) and depend on the size of the variables, so a smaller RSS for w does not mean a stronger linear relationship. Kwame's conclusion is not supported.

DO NOT WRITE IN THIS AREA
DO NOT WRITE IN THIS AREA
DO NOT WRITE IN THIS AREA

Question 1 continued

(Total for Question 1 is 7 marks)

DO NOT WRITE IN THIS AREA
DO NOT WRITE IN THIS AREA
DO NOT WRITE IN THIS AREA

2. A factory produces yellow tennis balls and white tennis balls. Independent samples, one of yellow tennis balls and one of white tennis balls, are taken. The table shows information about the weights of the yellow tennis balls, Y grams, and the weights of the white tennis balls, W grams.

	Sample size	Mean weight of random sample (grams)	Known population standard deviation of weights (grams)
Yellow tennis balls	120	57.2	1.2
White tennis balls	140	56.9	0.9

(a) Find a 95% confidence interval for the mean weight of **yellow** tennis balls.

(3)

Jamie claims that the mean weight of the population of yellow tennis balls is greater than the mean weight of the population of white tennis balls. A test of Jamie's claim is carried out.

(b) (i) Specify the approximate distribution of $\bar{Y} - \bar{W}$ under the null hypothesis of the test.

(3)

(ii) Explain the relevance of the large sample sizes to your answer to part (i).

(1)

(c) Complete the hypothesis test using a 5% level of significance.
You should state your hypotheses and the value of your test statistic clearly.

(5)

(a) 95% confidence interval: $z = 1.96$

$57.2 \pm 1.96 \times \frac{1.2}{\sqrt{120}}$

$= 57.2 \pm 0.214707...$

$= (56.985..., 57.414...)$

95% CI: (57.0, 57.4) grams

(b)(i) Under H_0: $\bar{Y} - \bar{W} \sim N(0, \frac{1.2^2}{120} + \frac{0.9^2}{140})$

$\bar{Y} - \bar{W} \sim N(0, 0.017785...)$ i.e. $N(0, \frac{249}{14000})$

(ii) The samples are large so, by the Central Limit Theorem, $\bar{Y}$ and $\bar{W}$ are approximately normally distributed, so we do not need to know the distributions of Y and W.

DO NOT WRITE IN THIS AREA
DO NOT WRITE IN THIS AREA
DO NOT WRITE IN THIS AREA

Question 2 continued

(c) H_0: $\mu_Y = \mu_W$ H_1: $\mu_Y > \mu_W$

where μ_Y and μ_W are the population mean weights of yellow and white balls

$$z = \frac{57.2 - 56.9}{\sqrt{0.017785...}} = 2.24949...$$

Test statistic $z = 2.25$

One-tailed 5% critical value: $z = 1.6449$

$2.25 > 1.6449$, so the result is significant: reject H_0

There is significant evidence to support Jamie's claim that the mean weight of the population of yellow tennis balls is greater than the mean weight of the population of white tennis balls.

DO NOT WRITE IN THIS AREA
DO NOT WRITE IN THIS AREA
DO NOT WRITE IN THIS AREA

Question 2 continued

DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA

Question 2 continued

(Total for Question 2 is 12 marks)

DO NOT WRITE IN THIS AREA
DO NOT WRITE IN THIS AREA
DO NOT WRITE IN THIS AREA

3. The random variable $X \sim N(5, 0.4^2)$ and the random variable $Y \sim N(8, 0.1^2)$

X and Y are independent random variables.

A random sample of a independent observations is taken from the distribution of X and one observation is taken from the distribution of Y

The random variable $W = X_1 + X_2 + X_3 + \ldots + X_a + bY$ and has the distribution $N(169, 2^2)$

Find the value of a and the value of b

(6)

$E(W) = aE(X) + bE(Y)$: $5a + 8b = 169$

$Var(W) = aVar(X) + b^2Var(Y)$: $0.4^2a + 0.1^2b^2 = 2^2$

$0.16a + 0.01b^2 = 4$

From the first equation: $a = 33.8 - 1.6b$

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

$5.408 - 0.256b + 0.01b^2 = 4$

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

$$b = \frac{0.256 \pm \sqrt{(0.256^2 - 4(0.01)(1.408))}}{2(0.01)} = \frac{0.256 \pm 0.096}{0.02}$$

$b = 17.6$ or $b = 8$

$b = 17.6$ gives $a = 33.8 - 28.16 = 5.64$, which is not possible since a must be a positive integer (a number of observations), so reject

$b = 8$ gives $a = 33.8 - 12.8 = 21$

$a = 21$, $b = 8$

DO NOT WRITE IN THIS AREA
DO NOT WRITE IN THIS AREA
DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA

Question 3 continued

(Total for Question 3 is 6 marks)

4. A doctor believes that a four-week exercise programme can reduce the resting heart rate of her patients. She takes a random sample of 7 patients and records their resting heart rate before the exercise programme and again after the exercise programme.

Patient	*A*	*B*	*C*	*D*	*E*	*F*	*G*
Resting heart rate before	65	68	77	79	80	88	92
Resting heart rate after	63	65	73	76	80	84	80

(a) Using a 5% level of significance, carry out an appropriate test of the doctor's belief. You should state your hypotheses, test statistic and critical value.

(7)

(b) State the assumption made about the resting heart rates that was required to carry out the test.

(1)

(a) Paired t-test. d = before − after: 2, 3, 4, 3, 0, 4, 12

$\Sigma d = 28, \quad \Sigma d^2 = 198, \quad \bar{d} = \frac{28}{7} = 4$

$s_d^2 = \frac{1}{6}(198 - 7 \times 4^2) = \frac{86}{6} = 14.333..., \quad s_d = 3.7859...$

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

(μ_d = mean reduction in resting heart rate)

$t = \frac{4}{3.7859... \div \sqrt{7}} = 2.79534...$ test statistic $t = 2.80$

Critical value: t_6(5%, one tail) = 1.943

2.80 > 1.943, so the result is significant: reject H_0

There is sufficient evidence to support the doctor's belief that the exercise programme reduces resting heart rate.

(b) The differences in resting heart rates (before − after) must be normally distributed.

DO NOT WRITE IN THIS AREA
DO NOT WRITE IN THIS AREA
DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA

Question 4 continued

(Total for Question 4 is 8 marks)

5. The concentration of an air pollutant is measured in micrograms/m^3

Samples of air were taken at two different sites and the concentration of this particular air pollutant was recorded.

For Site A the summary statistics are shown below.

	number of samples	s_A^2
Site *A*	13	6.39

For Site B there were 9 samples of air taken.

A test of the hypothesis $H_0 : \sigma_A^2 = \sigma_B^2$ against the hypothesis $H_1 : \sigma_A^2 \neq \sigma_B^2$ is carried out using a 2% level of significance.

(a) State a necessary assumption required to carry out the test.

(1)

Given that the assumption in part (a) holds,

(b) find the set of values of s_B^2 that would lead to the null hypothesis being rejected,

(4)

(c) find a 99% confidence interval for the variance of the concentration of the air pollutant at **Site *A***.

(3)

(a) The concentration of the air pollutant at each site follows a normal distribution.

(b) $F_{12,8}(1\%) = 5.67$ $F_{8,12}(1\%) = 4.50$

Reject H_0 if $\frac{6.39}{s_B^2} > 5.67$ or $\frac{s_B^2}{6.39} > 4.50$

$s_B^2 < \frac{6.39}{5.67} = 1.12698...$ or $s_B^2 > 4.50 \times 6.39 = 28.755$

Set of values: $0 < s_B^2 < 1.13$ or $s_B^2 > 28.8$

(c) $\chi^2_{12}(0.995) = 3.074$ $\chi^2_{12}(0.005) = 28.300$

$\frac{12 \times 6.39}{28.300} < \sigma_A^2 < \frac{12 \times 6.39}{3.074}$

$2.70954... < \sigma_A^2 < 24.9446...$

99% CI for σ_A^2: (2.71, 24.9)

DO NOT WRITE IN THIS AREA
DO NOT WRITE IN THIS AREA
DO NOT WRITE IN THIS AREA

Question 5 continued

(Total for Question 5 is 8 marks)

DO NOT WRITE IN THIS AREA
DO NOT WRITE IN THIS AREA
DO NOT WRITE IN THIS AREA

6. Korhan and Louise challenge each other to find an estimator for the mean, μ, of the continuous random variable X which has variance σ^2

$X_1, X_2, X_3, \ldots, X_n$ are n independent observations taken from X

Korhan's estimator is given by

$$K = \frac{2}{n(n+1)} \sum_{r=1}^{n} rX_r$$

Louise's estimator is given by

$$L = \frac{X_1 + X_2}{3} + \frac{X_3 + X_4 + \ldots + X_n}{3(n-2)}$$

(a) Show that K and L are both unbiased estimators of μ

(5)

(b) (i) Find $\text{Var}(K)$

(ii) Find $\text{Var}(L)$

(7)

The winner of the challenge is the person who finds the better estimator.

(c) Determine the winner of the challenge for large values of n.
Give reasons for your answer.

(3)

(a) $E(K) = \frac{2}{n(n+1)} \sum_{r=1}^{n} r\,E(X_r)$ (independent, $E(X_r) = \mu$)

$= \frac{2E(X)}{n(n+1)} + \frac{4E(X)}{n(n+1)} + \frac{6E(X)}{n(n+1)} + \ldots + \frac{2nE(X)}{n(n+1)}$

$= \frac{2\mu}{n(n+1)} \sum_{r=1}^{n} r = \frac{2\mu}{n(n+1)} \times \frac{n(n+1)}{2} = \mu$

so $E(K) = \mu$ and K is an unbiased estimator of μ

$E(L) = \frac{E(X_1) + E(X_2)}{3} + \frac{E(X_3) + \ldots + E(X_n)}{3(n-2)}$

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

so $E(L) = \mu$ and L is an unbiased estimator of μ

DO NOT WRITE IN THIS AREA
DO NOT WRITE IN THIS AREA
DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA
DO NOT WRITE IN THIS AREA
DO NOT WRITE IN THIS AREA

Question 6 continued

(b)(i) $\text{Var}(K) = \frac{2^2}{n^2(n+1)^2} \sum_{r=1}^{n} r^2 \text{Var}(X_r)$ (independent)

$= \frac{2^2}{n^2(n+1)^2} \text{Var}(X) + \frac{4^2}{n^2(n+1)^2} \text{Var}(X) + \ldots + \frac{(2n)^2}{n^2(n+1)^2} \text{Var}(X)$

$= \frac{4\sigma^2}{n^2(n+1)^2} \sum_{r=1}^{n} r^2$

$= \frac{4\sigma^2}{n^2(n+1)^2} \times \frac{n(n+1)(2n+1)}{6}$

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

(ii) $\text{Var}(L) = \text{Var}\left(\frac{X_1 + X_2}{3}\right) + \text{Var}\left(\frac{X_3 + X_4 + \ldots + X_n}{3(n-2)}\right)$

$= \frac{2}{9} \text{Var}(X) + \frac{(n-2)}{9(n-2)^2} \text{Var}(X)$

$= \frac{2\sigma^2}{9} + \frac{\sigma^2}{9(n-2)} = \frac{2(n-2)+1}{9(n-2)} \sigma^2$

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

Question 6 continued

(c) For large n:

$$\text{Var}(K) = \frac{2(2n+1)}{3n(n+1)}\,\sigma^2 \approx \frac{4n}{3n^2}\,\sigma^2 = \frac{4}{3n}\,\sigma^2 \to 0$$

$$\text{Var}(L) = \frac{2n-3}{9(n-2)}\,\sigma^2 \to \frac{2}{9}\,\sigma^2$$

Both K and L are unbiased, so the better estimator is the one with the smaller variance. For large n, $\text{Var}(K) \to 0 < \frac{2}{9}\sigma^2$, so $\text{Var}(K) < \text{Var}(L)$.

Therefore K is the better estimator and Korhan wins the challenge.

DO NOT WRITE IN THIS AREA
DO NOT WRITE IN THIS AREA
DO NOT WRITE IN THIS AREA

Question 6 continued

(Total for Question 6 is 15 marks)

DO NOT WRITE IN THIS AREA
DO NOT WRITE IN THIS AREA
DO NOT WRITE IN THIS AREA

7. A rectangle is to have an area of 40 cm^2

The length of the rectangle, L cm, follows a continuous uniform distribution over the interval [4, 10]

Find the expected value of the perimeter of the rectangle.
Use algebraic integration, rather than your calculator, to evaluate any definite integrals.

(7)

Width $= \frac{40}{L}$, so perimeter $P = 2(L + \frac{40}{L}) = 2L + \frac{80}{L}$

$L \sim U[4, 10]$, pdf of L: $f(x) = \frac{1}{6}$ for $4 \leq x \leq 10$

$E(P) = \int_4^{10} \frac{1}{6}(2x + \frac{80}{x})\,dx$

$= \frac{1}{6}\int_4^{10} (2x + \frac{80}{x})\,dx$

$= [\frac{1}{6}x^2 + \frac{40}{3}\ln x]_4^{10}$

$= (\frac{100}{6} + \frac{40}{3}\ln 10) - (\frac{16}{6} + \frac{40}{3}\ln 4)$

$= 14 + \frac{40}{3}\ln 2.5$

$= 26.2172...$

Expected perimeter = 26.2 cm

DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA

Question 7 continued

(Total for Question 7 is 7 marks)

DO NOT WRITE IN THIS AREA
DO NOT WRITE IN THIS AREA
DO NOT WRITE IN THIS AREA

8. The continuous random variable X has cumulative distribution function given by

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

(a) Find the exact value of the median of X

(2)

(b) Find $P(X < 1.6 \,|\, X > 1.2)$

(3)

The random variable $Y = \dfrac{1}{X}$

(c) Specify fully the cumulative distribution function of Y

(4)

(d) Hence or otherwise find the mode of Y

(3)

(a) $F(m) = 0.5$: $1.5m - 0.25m^2 - 1.25 = 0.5$

$0.25m^2 - 1.5m + 1.75 = 0 \Rightarrow m^2 - 6m + 7 = 0$

$m = 3 \pm \sqrt{2}$; reject $3 + \sqrt{2}$ since it is not in $1 \leq x \leq 3$

median $m = 3 - \sqrt{2}$

(b) $P(X < 1.6 \mid X > 1.2) = \dfrac{P(1.2 < X < 1.6)}{P(X > 1.2)} = \dfrac{F(1.6) - F(1.2)}{1 - F(1.2)}$

$F(1.6) = 2.4 - 0.64 - 1.25 = 0.51$, $F(1.2) = 1.8 - 0.36 - 1.25 = 0.19$

$P(X < 1.6 \mid X > 1.2) = \dfrac{0.51 - 0.19}{1 - 0.19} = \dfrac{0.32}{0.81} = \dfrac{32}{81}$

(c) $F_Y(y) = P(Y \leq y) = P(\dfrac{1}{X} \leq y) = P(X \geq \dfrac{1}{y}) = 1 - F(\dfrac{1}{y})$

$= 1 - (1.5(\dfrac{1}{y}) - 0.25(\dfrac{1}{y})^2 - 1.25) = 2.25 - \dfrac{1.5}{y} + \dfrac{0.25}{y^2}$

$x = 1 \Rightarrow y = 1$, $x = 3 \Rightarrow y = \dfrac{1}{3}$

DO NOT WRITE IN THIS AREA
DO NOT WRITE IN THIS AREA
DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA
DO NOT WRITE IN THIS AREA
DO NOT WRITE IN THIS AREA

Question 8 continued

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

(d) $f(y) = F_Y'(y) = 1.5y^{-2} - 0.5y^{-3}$ for $\frac{1}{3} \leq y \leq 1$

$f'(y) = -3y^{-3} + 1.5y^{-4}$

$f'(y) = 0$: $1.5y^{-4} = 3y^{-3} \Rightarrow \frac{1.5}{y^4} = \frac{3}{y^3} \Rightarrow y = 0.5$

$f''(y) = 9y^{-4} - 6y^{-5}$; $f''(0.5) = 144 - 192 = -48 < 0$, so a maximum and 0.5 lies in $\frac{1}{3} \leq y \leq 1$

Mode of $Y = 0.5$

Question 8 continued

(Total for Question 8 is 12 marks)

TOTAL FOR PAPER IS 75 MARKS

DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA