Please check the examination details below before entering your candidate information

Candidate surname	Other names
MODEL SOLUTIONS	

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 ▶

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.

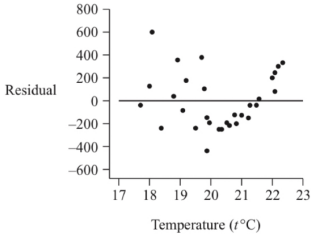
Residual Sum of Squares (RSS) = 1 666 567 $S_{tt} = 52.0$ $S_{yy} = 1\,774\,155$

least squares regression line: gradient = 45.5 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

We are given the gradient and y-intercept.

$y = mt + c$

$y = 45.5t + 2080$

Question 1 continued

We are given that the temperature is 20°C

$y = 45.5(20) + 2080$

$\Rightarrow y = 2990\,kg/hectare$ (1)

b) A residual is the difference between the observed value and the predicted value. (1)

c) In the formula booklet,

Residual Sum of Squares (RSS) $= S_{yy}(1 - r^2)$

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

$\Rightarrow r = 0.246$ (1)

d) i) 0.246 is close to 0, so there is weak correlation. (1)

ii) After $t = 20$, we can see that the plot is definitely not random, so not randomly scattered about 0. (1)

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) Kwame's conclusion cannot be supported using RSS because the two values of RSS have different values. (1)

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) Recall that for the Normal confidence interval, we use

$$\bar{x} \pm z \cdot \frac{\sigma}{\sqrt{n}}$$

$$\left(57.2 + 1.96 \left(\frac{1.2}{\sqrt{120}}\right), 57.2 - 1.96 \left(\frac{1.2}{\sqrt{120}}\right)\right) \text{①}$$

$$= (56.985, 57.414) \text{①}$$

We use 1.96, from either the table or calculator. Remember that this is two-tailed

Question 2 continued

b) Recall that for two independent samples X, Y,

$$\text{Var}(aX + bY) = a^2\text{Var}(X) + b^2\text{Var}(Y)$$

We apply the CLT, so we know that the variance is $\frac{\sigma^2}{n}$.

$$\text{Var}(\bar{Y} - \bar{W}) = \frac{\sigma_Y^2}{n_Y} + \frac{\sigma_W^2}{n_W}$$

$$= \frac{1.2^2}{120} + \frac{0.9^2}{140} \quad (1)$$

so $\bar{Y} - \bar{W} \sim N\left(0, \frac{1.2^2}{120} + \frac{0.9^2}{140}\right)$ (1) (1)

ii) We do not need to know the distribution of Y and W and can assume $\bar{Y}$ and $\bar{W}$ are normally distributed by the CLT. (1)

c) $H_0: \mu_Y = \mu_W$

$H_1: \mu_Y > \mu_W$ (1)

Note that $\mu_X - \mu_Y = 0$, so our test statistic is

$$z = \frac{57. - 56.9}{\sqrt{\frac{1.2^2}{120} + \frac{0.9^2}{140}}} = 2.249 \quad (1) \quad (1)$$

From the tables we see that the

Question 2 continued

Critical value at the 5% level is 1.6449 (1)

$2.249 > 1.6449$

Hence, we reject H_0 as there is significant evidence to support Jamie's claim that the mean weight of the population of yellow balls is greater than the mean weight of the population of white tennis balls. (1)

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)

$E[W] = \sum_{i=1}^{a} E[X_i] + E[bY]$

$= aE[X] + bE[Y]$

① $169 = 5a + 8b$ (1)

$Var(W) = Var(X_1 + \ldots + X_a) + Var(bY)$

$= \sum_{i=1}^{a} Var(X_i) + Var(bY)$ as $Var(X+Y) = Var(X) + Var(Y)$ if X and Y are independent

$= aVar(X) + b^2 Var(Y)$

① $4 = 0.16a + 0.01b^2$

$\Rightarrow 25 - 0.0625b^2 = a$ (2)

So we sub (2) into (1)

$\Rightarrow 169 = 5(25 - 0.0625b^2) + 8b$

$\Rightarrow 169 = 125 - 0.3125b^2 + 8b$

$\Rightarrow 0 = 0.3125b^2 - 8b + 44$ ①

Using a calculator, or the quadratic formula,

$b = 8$ or $b = 17.6$

Question 3 continued

If we first sub $b = 17.6$ in,

$169 = 5a + 8(17.6)$

This gives a non-integer value of a, so this is not a solution. (1)

Now, sub in $b = 8$.

$169 = 5a + 8(8)$

$\Rightarrow a = 21$ (1)

(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) $H_0: \mu_d = 0$ (1)

$H_1: \mu_d > 0$

We calculate the difference of the table

$\underline{d} = (2, 3, 4, 3, 0, 4, 12)$ (1)

As the variance is unknown, we use a t-test, so our test statistic is

$\frac{\bar{X} - \mu}{s/\sqrt{n}}$ and so we need to calculate s.

$s = \frac{1}{n-1}\left(\Sigma d^2 - n\bar{d}^2\right)$, or the formula in the formula booklet.

$\bar{d} = \frac{28}{7} = 4$

$\Sigma d^2 - n\bar{d} = (2-4)^2 + (3-4)^2 + (4-4)^2 + (3-4)^2 + \ldots = 86$

Question 4 continued

$s^2 = \frac{86}{6} = 14.3$

$\Rightarrow s = 3.7859$ (1) Keep this number exact by using the ANS button in the calculator.

Our test statistic $t = \frac{\pm 4 - 0}{\frac{3.7859}{\sqrt{7}}}$ (1) $= \pm 2.795$ (1)

As we have a sample size of 7, we compare our test statistic at the 5% significance level with $7 - 1 = 6$ degrees of freedom.

From the table, the critical value is ± 1.943 (1)

2.975 71.943

So we reject H_0 as there is sufficient evidence to support the doctor's belief that the resting heart rate has reduced. (1)

b) The difference in resting heart rates must be normally distributed for the test to be valid. (1)

(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 air pollutant for each site follows a normal distribution. (1)

b) From the formula book, we know that

$\frac{S_x^2/\sigma_x^2}{S_y^2/\sigma_y^2} \sim F_{n_x-1, n_y-1}$ We use 0.01 as our value of a 2% significant level.

Using the table,

$F_{12,8}(0.01) = 5.67$, $F_{8,12}(0.01) = 4.50$ (1)

$\frac{6.39}{S_B^2} > 5.67$ (1) , $\frac{S_B^2}{6.39} > 4.50$ (1)

Here, we use greater than, as we are

Question 5 continued

finding the set of values that would lead to the null hypothesis being rejected.

$1.127 > s_B^2$ or $s_B^2 > 28.755$ (1)

c) Recall the formula of a confidence interval for the variance is

$$\frac{(n-1)s^2}{\chi^2_{n-1}}$$

From the tables, our two values of χ^2 are 3.074 and 28.300. (1)

$$\frac{(13-1)(6.34)}{3.074} = 24.9447$$

$$\frac{(13-1)(6.34)}{28.3} = 2.7095 \quad (1)$$

$2.7095 < \sigma^2 < 24.9447$ (1)

(Total for Question 5 is 8 marks)

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 Var(K)

(ii) Find 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) Estimators L and K are unbiased if

$E[L] = E[X]$ and $E[K] = E[X]$ respectfully.

$$K = \frac{2}{n(n+1)} [X_1 + 2X_2 + \ldots + nX_n]$$

$$E[K] = E\left[\frac{2}{n(n+1)} (X_1 + 2X_2 + \ldots + nX_n)\right] \quad (1)$$

$$= \frac{2}{n(n+1)} E[X_1 + 2X_2 + \ldots + nX_n]$$

$$= \frac{2}{n(n+1)} \left(E[X_1] + E[2X_2] + \ldots + E[nX_n]\right)$$

Question 6 continued

$$= \frac{2}{n(n+1)}\left(E[X_1] + 2E[X_2] + \ldots + nE[X_n]\right)$$

$$= \frac{2}{n(n+1)}\left(\mu + 2\mu + \ldots + n\mu\right)$$

$$= \frac{2\mu}{n(n+1)}\left(1 + 2 + \ldots + n\right)$$

$$= \frac{2\mu}{n(n+1)}\left(\frac{1}{2}n(1+n)\right)$$ (1) Using the arithmetic sequence formula.

$= \mu = E[X]$. Hence K is unbiased (1)

$$E[L] = E\left[\frac{X_1 + X_2}{3} + \frac{X_3 + X_4 + \ldots + X_n}{3(n-2)}\right]$$ (1)

$$= \frac{1}{3}E[X_1 + X_2] + \frac{1}{3(n-2)}E[X_3 + X_4 + \ldots + X_n]$$

$$= \frac{1}{3}\left(E[X_1] + E[X_2]\right) + \frac{1}{3(n-2)}\left(E[X_3] + E[X_4] + \ldots + E[X_n]\right)$$

$$= \frac{1}{3}(2\mu) + \frac{1}{3(n-2)}\left[(n-2)\mu\right]$$

$$= \frac{2}{3}\mu + \frac{1}{3}\mu = \mu = E[X]$$ (1)

Hence, both L and K are unbiased.

b) i) $Var(K) = \left(\frac{2}{n(n+1)}\right)^2 Var(X_1 + 2X_2 + \ldots + nX_n)$ (1)

$$= \frac{4}{n^2(n+1)^2}\left[Var(X_1) + Var(2X_2) + \ldots + Var(nX_n)\right]$$

Question 6 continued

$$= \frac{4}{n^2(n+1)^2}\left[1^2\text{Var}(X_1) + 2^2\text{Var}(X_2) + \ldots + n^2\text{Var}(X_n)\right] \quad (1)$$

$$= \frac{4}{n^2(n+1)^2}\left[1^2\sigma^2 + 2^2\sigma^2 + \ldots + n^2\sigma^2\right]$$

$$= \frac{4\sigma^2}{n^2(n+1)^2}\left[1^2 + 2^2 + \ldots + n^2\right]$$

$$= \frac{4\sigma^2}{n^2(n+1)^2}\left[\frac{1}{6}n(n+1)(2n+1)\right] \quad (1)$$

Using the summation formula in the formula booklet.

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

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

$$= \frac{1}{9}\text{Var}(X_1 + X_2) + \frac{1}{9(n-2)^2}\text{Var}(X_3 + X_4 + \ldots + X_n) \quad (1)$$

$$= \frac{1}{9}\left[\text{Var}(X_1) + \text{Var}(X_2)\right] + \frac{1}{9(n-2)^2}\left[\text{Var}(X_3) + \ldots + \text{Var}(X_n)\right]$$

$$= \frac{1}{9}(2\sigma^2) + \frac{1}{9(n-2)^2}\left[(n-2)\sigma^2\right] \quad (1)$$

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

$$\Rightarrow \text{Var}(L) = \frac{2n-3}{9(n-2)}\sigma^2 \quad (1)$$

DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA

Question 6 continued

c) $\text{Var}(K) = \frac{2\sigma^2(2n+1)}{3n(n+1)}$, $\text{Var}(L) = \frac{2n-3}{9(n-2)}\sigma^2$

As $n \to \infty$, $\text{Var}(K) \to 0$

As $n \to \infty$, $\text{Var}(L) \to \frac{2}{9}\sigma^2$ (1)

The better estimator is the one with the smaller variance. (1)

Hence, K is the better estimator and so Korhun wins the challenge. (1)

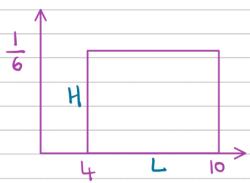
(Total for Question 6 is 15 marks)

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)



$LH = 40 \Rightarrow H = \frac{40}{L}$

Perimeter $= 2L + 2H$

$= 2L + \frac{80}{L}$ (1)

To find the expectation of this function, we use the formula

$E[g(x)] = \int g(x) f(x)\, dx$. Here $f(L) = 1/6$ (1)

$E\left[2L + \frac{80}{L}\right] = \frac{1}{3}\int_4^{10} L + 40L^{-1}\, dL$ (1)

$= \frac{1}{3}\left[\frac{L^2}{2} + 40\ln(L)\right]_4^{10}$ (1) (1)

$= \frac{1}{3}\left[\frac{10^2}{2} + 40\ln(10) - \frac{4^2}{2} - 40\ln(4)\right]$ (1)

$= 26.217$ (1)

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 $P(X < 1.6 \mid X > 1.2)$

(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) Recall that for the median, we want to find the value of x such that $F(x) = 1/2$.

$1.5m - 0.25m^2 - 1.25 = 0$ (1)

Using a calculator, this gives us

$m = 3 \pm \sqrt{2}$

We reject $3 + \sqrt{2} > 3$

so $m = 3 - \sqrt{2}$ (1)

b) Recall that

$$Pr(B|A) = \frac{Pr(A \cap B)}{Pr(A)}$$

Question 8 continued

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

As $F(x)=\Pr(X\leq x)$ $= \dfrac{F(1.6)-F(1.2)}{1-F(1.2)}$ (1)

$\Rightarrow 1-F(x)=1-\Pr(X>x)$

$= \dfrac{32}{81}$ (1)

By subbing in $x=1.6$ and $x=1.2$.

c) $F(y) = \Pr(Y\leq y)$ (1)

$= \Pr\left(\frac{1}{X}\leq y\right)$

$= \Pr\left(\frac{1}{y}\leq x\right)$

$= 1-\Pr\left(X\leq \frac{1}{y}\right)$

$= 1-F\left(\frac{1}{y}\right)$ (1)

So $F\left(\frac{1}{y}\right) = \begin{cases} 0 & 1<y \\ 1.5\left(\frac{1}{y}\right)-0.25\left(\frac{1}{y}\right)^2-1.25 & 1\geq y\geq \frac{1}{3} \\ 1 & \frac{1}{3}>y \end{cases}$

Question 8 continued

$$1 - F\left(\frac{1}{y}\right) = \begin{cases} 1 & y > 1 \\ 1 - 1.5\left(\frac{1}{y}\right) + 0.25\left(\frac{1}{y}\right)^2 + 1.25 & \frac{1}{3} \le y \le 1 \\ 0 & y < \frac{1}{3} \end{cases}$$ (1)

$$\text{So } F(y) = \begin{cases} 0 & y < \frac{1}{3} \\ 0.25\left(\frac{1}{y}\right)^2 - 1.5\left(\frac{1}{y}\right) + 2.25 & \frac{1}{3} \le y \le 1 \\ 1 & y > 1 \end{cases}$$ (1)

d) The mode is the peak on the pdf graph.

We differentiate the cdf to find the pdf.

$f(y) = -0.5y^{-3} + 1.5y^{-2}$ (1)

$\Rightarrow f'(y) = 1.5y^{-4} - 3y^{-3} = 0$ (1) We are looking for a maximum point.

$\Rightarrow y^{-1} = 2 \Rightarrow y = 1/2$

$f''(1/2) = -48 < 0$, so we have a maximum.

So the mode of y is $1/2$. (1)

(Total for Question 8 is 12 marks)

TOTAL FOR PAPER IS 75 MARKS