Please check the examination details below before entering your candidate information

Candidate surname StemUp Model Solution | Other names

Pearson Edexcel Level 3 GCE

Centre Number | Candidate Number

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 algebraic 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 7 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.
- Good luck with your examination.

Turn over ▶

P66803A

©2021 Pearson Education Ltd.

1/1/1/1/1/

P 6 6 8 0 3 A 0 1 2 8

Pearson

1. Anisa is investigating the relationship between marks on a History test and marks on a Geography test. She collects information from 7 students. She wants to calculate the Spearman's rank correlation coefficient for the 7 students so she ranks their performance on each test.

Student	History mark	Geography mark	History rank	Geography rank
A	76	58	1	3
B	70	60	2	2
C	64	57	*s*	*t*
D	64	63	*s*	1
E	64	57	*s*	*t*
F	59	50	6	7
G	55	52	7	6

(a) Write down the value of s and the value of t

(2)

The full product moment correlation coefficient (pmcc) formula is used with the ranks to calculate the Spearman's rank correlation coefficient instead of $r_s = 1 - \dfrac{6\Sigma d^2}{n(n^2-1)}$ and the value obtained is 0.7106 to 4 significant figures.

(b) Explain why the full pmcc formula is used to carry out the calculation.

(1)

(c) Stating your hypotheses clearly, test whether or not there is evidence to suggest that the higher a student ranks in the History test, the higher the student ranks in the Geography test. Use a 5% level of significance.

(4)

(a) $s = 4$ $t = 4.5$

(b) Because there are tied ranks.

(c) H_0: $\rho_s = 0$ H_1: $\rho_s > 0$

$n = 7$, 5% one-tailed critical value = 0.7143

$r_s = 0.7106 < 0.7143$, so r_s does not lie in the critical region:
do not reject H_0
There is insufficient evidence to suggest that the higher a student ranks in the History test, the higher the student ranks in the Geography test.

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 company produces two colours of candles, blue and white. The standard deviation of the burning times of the blue candles is 2.6 minutes and the standard deviation of the burning times of the white candles is 2.4 minutes.

Nissim claims that the mean burning time of blue candles is more than 5 minutes greater than the mean burning time of white candles.

A random sample of 90 blue candles is found to have a mean burning time of 39.5 minutes. A random sample of 80 white candles is found to have a mean burning time of 33.7 minutes.

(a) Stating your hypotheses clearly, use a suitable test to assess Nissim's belief. Use a 1% level of significance.

(6)

(b) Explain how the hypothesis test in part (a) would be carried out differently if the variances of the burning times of candles were unknown.

(1)

The burning times for the candles may not follow a normal distribution.

(c) Describe the effect this would have on the calculations in the hypothesis test in part (a). Give a reason for your answer.

(2)

(a) H_0: $\mu_{blue} = \mu_{white} + 5$ $\quad H_1$: $\mu_{blue} > \mu_{white} + 5$

μ_{blue}, μ_{white} = mean burning times of blue and white candles

standard error $= \sqrt{\left(\frac{2.6^2}{90} + \frac{2.4^2}{80}\right)} = \sqrt{0.147111\ldots} = 0.383550\ldots$

$z = \frac{39.5 - 33.7 - 5}{0.383550\ldots} = 2.085773\ldots \quad z = 2.09$

1% one-tailed critical value: $z = 2.3263$

$2.09 < 2.3263$, so the result is not significant: do not reject H_0

There is insufficient evidence to support Nissim's claim that the mean burning time of blue candles is more than 5 minutes greater than the mean burning time of white candles.

(b) The samples are large, so the sample variances s^2 would be used as approximations for the population variances σ^2 (then the same z test).

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

Question 2 continued

(c) No effect: the calculations in part (a) can still be carried out.

Reason: the sample sizes (90 and 80) are large, so by the Central Limit Theorem the sample means are approximately normally distributed.

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 9 marks)

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

$$F(x) = \begin{cases} 0 & x < 2 \\ 1.25 - \dfrac{2.5}{x} & 2 \leqslant x \leqslant 10 \\ 1 & x > 10 \end{cases}$$

(a) Find $P(\{X < 5\} \cup \{X > 8\})$

(2)

(b) Find the median of X.

(2)

(c) Find $E(X^2)$

(3)

(d) (i) Sketch the probability density function of X.

(ii) Describe the skewness of the distribution of X.

(3)

(a) $P(\{X < 5\} \cup \{X > 8\}) = F(5) + (1 - F(8))$

$= (1.25 - \frac{2.5}{5}) + (1 - (1.25 - \frac{2.5}{8})) = \frac{3}{4} + (1 - \frac{15}{16})$

$= \frac{13}{16}$

(b) $F(m) = 0.5$: $1.25 - \frac{2.5}{m} = 0.5$

$\frac{2.5}{m} = 0.75 \quad \Rightarrow \quad m = \frac{10}{3}$

median $= \frac{10}{3}$

(c) $f(x) = \frac{d}{dx}(F(x)) = 2.5x^{-2}$ for $2 \leq x \leq 10$

$E(X^2) = \int_2^{10} x^2 \times 2.5x^{-2}\, dx = \int_2^{10} 2.5\, dx = [2.5x]_2^{10}$

DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA

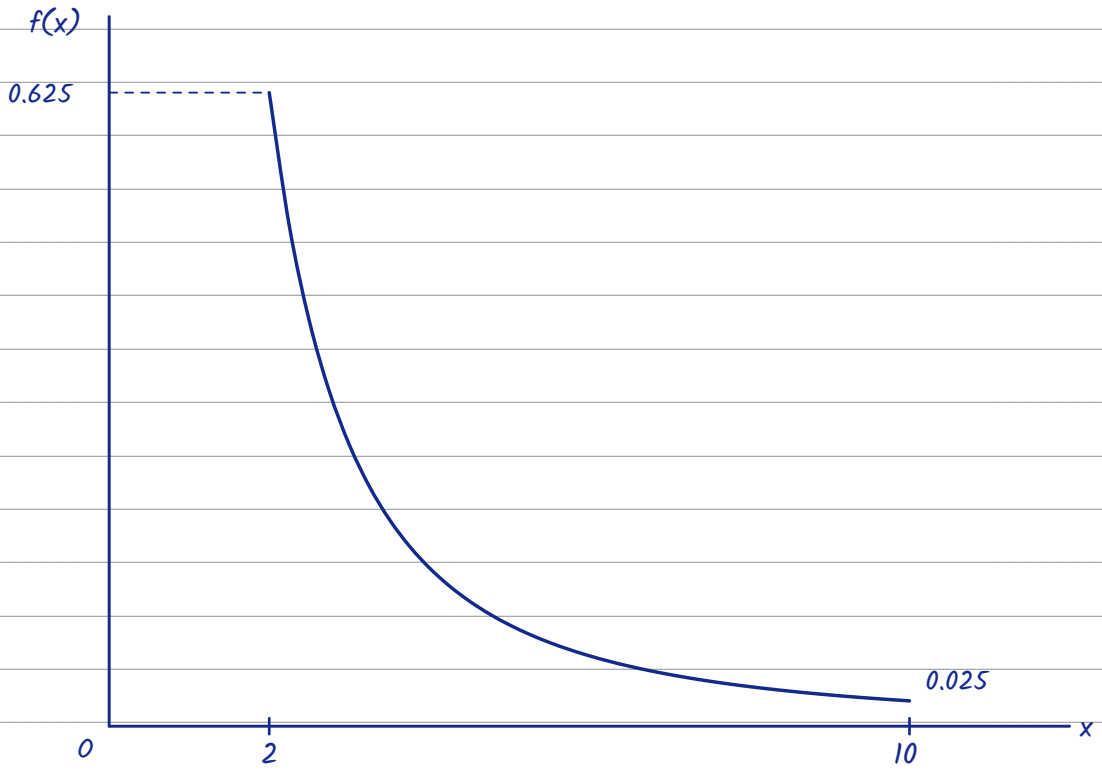
DO NOT WRITE IN THIS AREA

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

Question 3 continued

$= 25 - 5 = 20$

(d)(i) $f(x) = 2.5x^{-2}$ for $2 \leqslant x \leqslant 10$ (0 otherwise): $f(2) = 0.625$, $f(10) = 0.025$



(d)(ii) The distribution has positive skew.

(the pdf decreases from $x = 2$; mode 2 < median $\frac{10}{3}$ < mean)

Question 3 continued

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 10 marks)

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

4. A researcher is investigating the relationship between elevation, x metres, and annual mean temperature, t °C.

From a random sample of 20 weather stations in Switzerland, the following results were obtained

$$S_{xx} = 8\,820\,655 \qquad S_{tt} = 444.7 \qquad \sum x = 28\,130 \qquad \sum t = 94.62$$

The product moment correlation coefficient for these data is found to be −0.959

(a) Interpret the value of this correlation coefficient.

(1)

(b) Show that the equation of the regression line of t on x can be written as

$$t = 14.3 - 0.00681x$$

(4)

The random variable W represents the elevations of the weather stations in kilometres.

(c) Write down the equation of the regression line of t on w for these 20 weather stations in the form $t = a + bw$

(1)

(d) Show that the residual sum of squares (RSS) for the model for t and x is 35.7 correct to one decimal place.

(1)

One of the weather stations in the sample had a recorded elevation of 1100 metres and an annual mean temperature of 1.4 °C

(e) (i) Calculate this weather station's contribution to the residual sum of squares. Give your answer as a percentage.

(2)

(ii) Comment on the data for this weather station in light of your answer to part (e)(i).

(1)

(a) There is strong negative correlation: as elevation increases, annual mean temperature decreases.

(b) $S_{xt} = r\sqrt{(S_{xx}S_{tt})} = -0.959 \times \sqrt{(8\,820\,655 \times 444.7)}$
$= -60\,062.38727...$

$b = \frac{S_{xt}}{S_{xx}} = \frac{-60\,062.387...}{8\,820\,655} = -0.006809288...$

DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA

Question 4 continued

$a = \bar{t} - b\bar{x} = \frac{94.62}{20} - (-0.006809288\ldots) \times \frac{28\,130}{20}$

$= 4.731 + 9.577264\ldots = 14.30826\ldots$

$t = 14.3 - 0.00681x$ (to 3 s.f.)

(c) $x = 1000w$: $t = 14.3 - 6.81w$

(d) $RSS = S_{tt}(1 - r^2) = 444.7(1 - (-0.959)^2) = 444.7 \times 0.080319$

$= 35.717859\ldots = 35.7$ (1 d.p.)

(e)(i) Predicted $t = 14.3 - 0.00681 \times 1100 = 6.809$

residual $= 1.4 - 6.809 = -5.409$

$(\text{residual})^2 = (-5.409)^2 = 29.257281$

$\text{contribution} = \frac{29.257281}{35.7} \times 100\% = 81.953\ldots\% = 82\%$ (2 s.f.)

(e)(ii) This one station contributes about 82% of the RSS (the majority), so this point is possibly an outlier and should be investigated.

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

Question 4 continued

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 10 marks)

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

5. The continuous random variable X is uniformly distributed over the interval $[0, 4\beta]$, where β is an unknown constant.

Three independent observations, X_1, X_2 and X_3, are taken of X and the following estimators for β are proposed

$$A = \frac{X_1 + X_2}{2}$$

$$B = \frac{X_1 + 2X_2 + 3X_3}{8}$$

$$C = \frac{X_1 + 2X_2 - X_3}{8}$$

(a) Calculate the bias of A, the bias of B and the bias of C

(5)

(b) By calculating the variances, explain which of B or C is the better estimator for β

(4)

(c) Find an unbiased estimator for β

(1)

(a) $X \sim U[0, 4\beta]$: $E(X) = 2\beta$, $Var(X) = \frac{(4\beta)^2}{12} = \frac{4}{3}\beta^2$

$E(A) = \frac{1}{2}[E(X_1) + E(X_2)] = \frac{1}{2}(2\beta + 2\beta) = 2\beta$

$E(B) = \frac{1}{8}[E(X) + 2E(X) + 3E(X)] = \frac{6 \times 2\beta}{8} = 1.5\beta$

$E(C) = \frac{1}{8}[E(X) + 2E(X) - E(X)] = \frac{2 \times 2\beta}{8} = 0.5\beta$

$Bias(A) = E(A) - \beta = 2\beta - \beta = \beta$

$Bias(B) = E(B) - \beta = 1.5\beta - \beta = 0.5\beta$

$Bias(C) = E(C) - \beta = 0.5\beta - \beta = -0.5\beta$

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 5 continued

(b) B and C have biases of equal size (0.5β), so the better estimator is the one with the smaller variance. $(X_1, X_2, X_3$ independent)

$$\text{Var}(B) = \frac{1}{64}[\text{Var}(X) + 4\text{Var}(X) + 9\text{Var}(X)] = \frac{14}{64}\text{Var}(X) = \frac{7}{32} \times \frac{4}{3}\beta^2 = \frac{7}{24}\beta^2$$

$$\text{Var}(C) = \frac{1}{64}[\text{Var}(X) + 4\text{Var}(X) + \text{Var}(X)] = \frac{6}{64}\text{Var}(X) = \frac{3}{32} \times \frac{4}{3}\beta^2 = \frac{1}{8}\beta^2$$

Since both have the same size of bias and $\text{Var}(C) = \frac{1}{8}\beta^2 < \text{Var}(B) = \frac{7}{24}\beta^2$,

C is the better estimator.

(c) $E(X_1 + X_2 + X_3) = 6\beta$, so an unbiased estimator is $\frac{X_1 + X_2 + X_3}{6}$

Question 5 continued

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 10 marks)

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

6. Elsa is collecting information on the wingspan of two different species of butterfly, Ringlet and Meadow Brown. She takes a random sample of each type of butterfly. The wingspans, w cm, are summarised in the table below. The wingspans of Ringlet and Meadow Brown butterflies each follow normal distributions.

	Number of butterflies	$\sum w$	$\sum w^2$
Ringlet	8	410	21 032
Meadow Brown	6	294	14 426

(a) Test, at the 2% level of significance, whether or not there is evidence that the variance of the wingspans of Ringlet butterflies is different from the variance of the wingspans of Meadow Brown butterflies. You should state your hypotheses clearly.

(7)

The k% confidence interval for the variance of the wingspans of Meadow Brown butterflies is (1.194, 48.54)

(b) Find the value of k

(3)

(c) Calculate a 95% confidence interval for the difference between the mean wingspan of the Ringlet butterfly and the mean wingspan of the Meadow Brown butterfly.

(5)

(a) $H_0: \sigma_R^2 = \sigma_M^2$ $H_1: \sigma_R^2 \neq \sigma_M^2$

$s_R^2 = \frac{1}{7}\left(21\,032 - 8 \times \left(\frac{410}{8}\right)^2\right) = \frac{19.5}{7} = 2.785714...$

$s_M^2 = \frac{1}{5}\left(14\,426 - 6 \times \left(\frac{294}{6}\right)^2\right) = \frac{20}{5} = 4$

$F = \frac{s_M^2}{s_R^2} = \frac{4}{2.785714...} = 1.435897... \quad (= 1.44)$

Critical value (2%, two-tailed): $F_{5,7} = 7.46$

$1.44 < 7.46$, so not significant: do not reject H_0

There is insufficient evidence to suggest that the variances of the wingspans of Ringlet and Meadow Brown butterflies are different.

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) $\frac{5s_M^2}{\sigma^2} \sim \chi^2_5$: $\frac{5 \times 4}{1.194} = 16.75$ and $\frac{5 \times 4}{48.54} = 0.412$

$\chi^2_5(0.005) = 16.750$ so the upper tail probability is 0.005

$\chi^2_5(0.995) = 0.412$ so the lower tail probability is 0.005

$1 - 2 \times 0.005 = 0.99$, so $k = 99$

(c) $s_p^2 = \frac{7 \times 2.785714... + 5 \times 4}{8 + 6 - 2} = \frac{39.5}{12} = 3.291666...$

$\bar{x}_R = \frac{410}{8} = 51.25$, $\bar{x}_M = \frac{294}{6} = 49$, $t_{12}(2.5\%) = 2.179$

$(51.25 - 49) \pm 2.179 \times \sqrt{3.291666...} \times \sqrt{\left(\frac{1}{8} + \frac{1}{6}\right)}$

$= 2.25 \pm 2.135052...$ $= (0.114947..., 4.385052...)$

95% confidence interval: (0.115, 4.39) cm

Question 6 continued

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. The weights of a particular type of apple, A grams, and a particular type of orange, R grams, each follow independent normal distributions.

$$A \sim \mathrm{N}(160, 12^2) \qquad R \sim \mathrm{N}(140, 10^2)$$

(a) Find the distribution of

(i) $A + R$

(ii) the total weight of 2 randomly selected apples.

(3)

A box contains 4 apples and 1 orange only. Jesse selects 2 pieces of fruit at random from the box.

(b) Find the probability that the total weight of the 2 pieces of fruit exceeds 310 grams.

(3)

From a large number of apples and oranges, Celeste selects m apples and 1 orange at random. The random variable W is given by

$$W = \left(\sum_{i=1}^{m} A_i\right) - n \times R$$

where n is a positive integer.

Given that the middle 95% of the distribution of W lies between 1100.08 and 1499.92 grams,

(c) find the value of m and the value of n

(8)

(a)(i) $A + R \sim N(160 + 140, 12^2 + 10^2)$, i.e. $A + R \sim N(300, 244)$

(ii) $A_1 + A_2 \sim N(2 \times 160, 2 \times 12^2)$, i.e. $A_1 + A_2 \sim N(320, 288)$

(b) $P(\text{both apples}) = \frac{4}{5} \times \frac{3}{4} = \frac{3}{5}$ $P(\text{one apple, one orange}) = \frac{2}{5}$

$P(A_1 + A_2 > 310) = P\left(Z > \frac{310 - 320}{\sqrt{288}}\right) = P(Z > -0.58926) = 0.722155...$

$P(A + R > 310) = P\left(Z > \frac{310 - 300}{\sqrt{244}}\right) = P(Z > 0.64018) = 0.261026...$

$P(\text{total} > 310) = \frac{3}{5} \times 0.722155... + \frac{2}{5} \times 0.261026...$

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

$= 0.537703... = 0.538$

(c) $W = A_1 + A_2 + ... + A_m - nR$

$E(W) = 160m - 140n, \quad Var(W) = 12^2m + 10^2n^2 = 144m + 100n^2$

$W \sim N(160m - 140n, \ 144m + 100n^2)$

Mean (by symmetry): $160m - 140n = \frac{1100.08 + 1499.92}{2} = 1300$

Middle 95%: $\mu \pm 1.96\sigma$, so $2 \times 1.96\sigma = 1499.92 - 1100.08 = 399.84$

$\sigma = 102 \Rightarrow 144m + 100n^2 = 102^2 = 10\,404$

From the mean: $m = \frac{1300 + 140n}{160}$

$144 \times \frac{1300 + 140n}{160} + 100n^2 = 10\,404$

$1170 + 126n + 100n^2 = 10\,404$

$100n^2 + 126n - 9234 = 0$

$n = \frac{-126 \pm \sqrt{(126^2 + 4 \times 100 \times 9234)}}{200} = \frac{-126 \pm 1926}{200}$

$n = 9$ or $n = -10.26$ (reject, n is a positive integer)

$m = \frac{1300 + 140 \times 9}{160} = \frac{2560}{160} = 16$

$m = 16, \quad n = 9$

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

Question 7 continued

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

Question 7 continued

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 14 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