

Please check the examination details below before entering your candidate information

Candidate surname				Other names			
Pearson Edexcel				Centre Number		Candidate Number	
Level 3 GCE				<input type="text"/> <input type="text"/> <input type="text"/> <input type="text"/> <input type="text"/> <input type="text"/>		<input type="text"/> <input type="text"/> <input type="text"/> <input type="text"/> <input type="text"/> <input type="text"/>	
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 $r_s = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$

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) $S = \frac{3+4+5}{3} = 4$ ① We find the mean of the missing ranks.

$t = \frac{4+5}{2} = 4.5$ ①

b) Because some of the ranks are tied. ①



Question 1 continued

$$c) H_0: \rho = 0$$

$$H_1: \rho > 0 \quad \text{— Testing positive correlation.}$$

We are given that $n=7$ and have a one-tailed test. Using the Spearman's Coefficient table in the formula book, we see that

$$CV = 0.7143$$

$$0.7143 > 0.7106 = r_s$$

Hence, r_s is not in the critical region. ①

So there is insufficient evidence to suggest that the higher the rank in the History test, the higher the rank in the Geography test. ①

(Total for Question 1 is 7 marks)



P 6 6 8 0 3 A 0 3 2 8

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)

$$a) H_0: \mu_B = \mu_W + 5$$

$$H_1: \mu_B > \mu_W + 5$$

From the formula book, we see that

$$Z = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}$$

Subbing in values, we see that

$$Z = \frac{39.5 - 33.7 - (\mu_B - \mu_W)}{\sqrt{\frac{2.6^2}{90} + \frac{2.4^2}{80}}}$$



Question 2 continued

From H_0 , we can use $\mu_B - \mu_W = 5$

$$z = \frac{39.5 - 33.7 - 5}{\sqrt{\frac{2.6^2}{90} + \frac{2.4^2}{80}}} = 2.085773$$

Using the Normal table in the formula book, we look at $p=0.01$ as we have a 1% level of significance.

$$CV = 2.3263$$

$$2.3263 > 2.085773$$

Do not reject H_0 as there is insufficient evidence to support Nissim's claim.

b) We would use a t-test

c) As we have a large sample size, the sample means will be approximately normally distributed by the Central Limit Theorem.

So there will be no effect as the calculations in part (a) can still be carried out.



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

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

- (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) $Pr([X < 5] \cup [X > 8])$ $F(x) = Pr(X \leq x)$

① $= F(5) + (1 - F(8))$ We can ignore the = in \leq as our data is continuous.

$= (1.25 - \frac{2.5}{5}) + 1 - (1.25 - \frac{2.5}{8})$

$= \frac{13}{16}$ ①

b) The median of X is at $F(x) = 0.5$

$$1.25 - \frac{2.5}{m} = 0.5$$
 ①

Rearranging yields $m = \frac{10}{3}$ ①



Question 3 continued

c) Recall that $E[g(x)] = \int_a^b g(x)f(x) dx$
 where $f(x)$ is the PDF.

We also recall that $f(x) = \frac{d}{dx} (F(x))$

$$f(x) = \frac{d}{dx} (F(x)) = \frac{d}{dx} (1.25 - 2.5x^{-1})$$

$$= 2.5x^{-2} \quad (1)$$

$$E[x^2] = \int_2^{10} x^2 \times 2.5x^{-2} dx$$

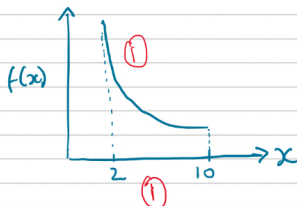
$$= \int_2^{10} 2.5 dx \quad (1)$$

$$= [2.5x]_2^{10}$$

$$= 2.5(10) - 2.5(2)$$

$$= 20 \quad (1)$$

d) i) $f(x) = \begin{cases} 2.5x^{-2} & 2 \leq x \leq 10 \\ 0 & \text{otherwise} \end{cases}$



Plot $f(x) = \frac{2.5}{x^2}$
 for $2 \leq x \leq 10$



Question 3 continued

ii) From the graph, we can see that for smaller values of x , the larger the PDF

Positive Skew. ①

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

$$S_x = 8\,820\,655$$

$$S_t = 444.7$$

$$\sum x = 28\,130$$

$$\sum t = 94.62$$

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) The higher the elevation, the lower the temperature.

(1)

b) From the formula book, we see that for least squares regression line of y on x is $y = a + bx$,

The regression coefficient of y on x is $b = \frac{S_{xy}}{S_{xx}}$ (1)



Question 4 continued

We also see that

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \quad (2)$$

from the formula book, where r is the product moment correlation coefficient.

Using these formulas yields

$$-0.959 = \frac{S_{xt}}{\sqrt{8820655 \times 444.7}} \quad (2)$$

$$\Rightarrow S_{xt} = -60062.38727 \quad (1)$$

$$b = \frac{S_{xt}}{S_{xx}} = \frac{-60062.38727}{8820655} \quad (1)$$

$$= -0.00681 \quad (1)$$

Recall that $a = \bar{t} - (b \times \bar{x})$, $\bar{t} = \frac{\sum t}{n}$, $\bar{x} = \frac{\sum x}{n}$

$$a = \frac{94.62}{20} - (-0.00681 \times \frac{28130}{20})$$

$$= 14.3 \quad (1)$$

$$t = a + bx$$

$$t = 14.3 - 0.00681x \quad (1)$$

c) x is in metres so $x = 1000w$

$$t = 14.3 - 6.81w \quad (1)$$



Question 4 continued

d) From the formula book,

$$RSS = s_{yy} - \frac{(s_{xy})^2}{s_{xx}} = s_{yy}(1-r^2)$$

$$RSS = 444.7 - \frac{(-60062.387)^2}{8820655}$$

$$= 35.71$$

$$= 35.7 \quad (1 \text{ d.p.}) \textcircled{1}$$

Alternatively,

$$RSS = 447(1 - (-0.959)^2)$$

$$= 35.71$$

$$= 35.7 \quad (1 \text{ d.p.})$$

e) i) Subbing the given value yields

$$t = 14.3 - (0.00681 \times 1100)$$

$$= 6.809$$

We square the difference of this and the mean.

$$(1.4 - 6.809)^2 = 29.257 \textcircled{1}$$

So the weather station's contribution would be



Question 4 continued

$$\frac{29.257}{35.7} = 81.95\% \quad (1)$$

ii) The point is likely an outlier (1)

This means that most of the contribution comes from this one data point, which seems very unlikely.

(Total for Question 4 is 10 marks)



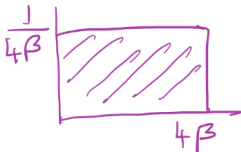
5. The continuous random variable X is uniformly distributed over the interval $[0, 4\beta]$, where β is an unknown constant.

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) Recall from the formula book that for the Uniform distribution on $[a, b]$

$$E[X] = \frac{1}{2}(a+b), \text{ Var}(X) = \frac{1}{12}(b-a)^2$$

$$\text{So for } X \sim U(0, 4\beta), \quad E[X_1 + X_2] = E[X_1] + E[X_2]$$

$$E[X] = 2\beta, \quad \text{Var}(X) = \frac{4\beta^2}{3} \quad E[aX] = aE[X]$$

$$\begin{aligned} E[A] &= E\left[\frac{X_1 + X_2}{2}\right] = \frac{1}{2} E[X_1 + X_2] \\ &= \frac{1}{2} E[X_1] + \frac{1}{2} E[X_2] \\ &= \frac{1}{2} (2\beta) + \frac{1}{2} (2\beta) \\ &= 2\beta \end{aligned}$$



Question 5 continued

$$\begin{aligned} E[\beta] &= E\left[\frac{x_1 + 2x_2 + 3x_3}{8}\right] \\ &= \frac{1}{8} E[x_1 + 2x_2 + 3x_3] \\ &= \frac{1}{8} E[x_1] + \frac{1}{4} E[x_2] + \frac{3}{8} E[x_3] \\ &= \frac{1}{4}\beta + \frac{1}{2}\beta + \frac{3}{4}\beta \\ &= \frac{3}{2}\beta \end{aligned}$$

$$\begin{aligned} E[c] &= E\left[\frac{x_1 + 2x_2 - x_3}{8}\right] \\ &= \frac{1}{8} E[x_1 + 2x_2 - x_3] \\ &= \frac{1}{8} E[x_1] + \frac{1}{4} E[x_2] - \frac{1}{8} E[x_3] \\ &= \frac{1}{4}\beta + \frac{1}{2}\beta - \frac{1}{4}\beta \\ &= \frac{1}{2}\beta \quad \textcircled{1} \end{aligned}$$

Recall $b(\hat{\beta}) = E[\hat{\beta}] - \beta$

Bias of $A = \textcircled{1} 2\beta - \beta = \beta \quad \textcircled{1}$

Bias of $B = 1.5\beta - \beta = 0.5\beta \quad \textcircled{1}$

Bias of $C = 0.5\beta - \beta = -0.5\beta \quad \textcircled{1}$



Question 5 continued

b) Recall $\text{Var}(aX) = a^2 \text{Var}(X)$, $\text{Var}(X_1 + X_2) = \text{Var}(X_1) + \text{Var}(X_2)$
 (if X_1 and X_2 are independent)

$$\text{Var}(B) = \text{Var}\left(\frac{X_1 + 2X_2 + 3X_3}{8}\right)$$

$$= \frac{1}{64} \text{Var}(X_1) + \frac{1}{64} \text{Var}(2X_2) + \frac{1}{64} \text{Var}(3X_3)$$

$$= \frac{1}{64} \text{Var}(X_1) + \frac{1}{16} \text{Var}(X_2) + \frac{9}{64} \text{Var}(X_3)$$

$$= \frac{1}{64} \times \frac{4}{3} \beta^2 + \frac{1}{16} \times \frac{4}{3} \beta^2 + \frac{9}{64} \times \frac{4}{3} \beta^2$$

$$= \frac{7}{24} \beta^2 \quad \textcircled{1}$$

$$\text{Var}(C) = \text{Var}\left(\frac{X_1 + 2X_2 - X_3}{8}\right)$$

$$= \frac{1}{64} \text{Var}(X_1 + 2X_2 - X_3)$$

$$= \frac{1}{64} \text{Var}(X_1) + \frac{1}{64} \text{Var}(2X_2) + \frac{1}{64} \text{Var}(-X_3)$$

$$= \frac{1}{64} \text{Var}(X_1) + \frac{1}{16} \text{Var}(X_2) + \frac{1}{64} \text{Var}(X_3)$$

$$= \frac{1}{64} \times \frac{4}{3} \beta^2 + \frac{1}{16} \times \frac{4}{3} \beta^2 + \frac{1}{64} \times \frac{4}{3} \beta^2 \quad \textcircled{1}$$

$$= \frac{1}{8} \beta^2 \quad \textcircled{1}$$



Question 5 continued

The better estimator would have the smallest bias and the least variance.

As B and C have equal bias, we choose the estimator with smaller variance.

As $\beta^2 > 0$,

$$\text{Var}(B) = \frac{7}{24} \beta^2 > \frac{1}{8} \beta^2 = \text{Var}(C)$$

so C is the better estimator. ①

c) We want $E[\hat{\beta}] - \beta = 0$

$$\frac{X_1}{2}$$

①

This works as $E\left[\frac{X_1}{2}\right] = \beta$ and $\beta - \beta = 0$.

(Total for Question 5 is 10 marks)



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	21032
Meadow Brown	6	294	14426

- (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)

$$a) H_0: \sigma_R^2 = \sigma_{MB}^2$$

$$H_1: \sigma_R^2 \neq \sigma_{MB}^2 \quad (1)$$

$$\text{Recall that } s^2 = \frac{1}{n-1} (\sum x^2 - n\bar{x}^2)$$

$$s_R^2 = \frac{1}{7} (21032 - 8 \left(\frac{410}{8} \right)^2) \quad (1)$$

$$= 2.7857 \quad (1)$$

$$s_{MB}^2 = \frac{1}{5} (14426 - 6 \left(\frac{294}{6} \right)^2)$$

$$= 4 \quad (1)$$



Question 6 continued

Because the test is two tailed, we use 0.01 probability.

Looking at the table in the formula book, we choose $v_1 = 6 - 1 = 5$ and $v_2 = 8 - 1 = 7$.

$$CV = 7.46 \quad (1)$$

Because we used v_1 from the Meadow Brown data, our test statistic is

$$\frac{4}{2.7857} = 1.436 \quad (1)$$

$$1.436 < 7.46 \quad \text{so,}$$

Do not reject H_0 , as there is insufficient evidence to suggest that the variances of the wingspans are different. (1)

b) Recall that $\frac{(n-1)s^2}{\chi_{n-1}(\frac{\alpha}{2})} = \text{Lower bound.}$

$$\frac{(6-1) \times 4^2}{\chi_{6-1}(\alpha/2)} = 1.194 \quad (1)$$

$$\Rightarrow \chi(\alpha/2) = 16.75$$

We compare this with the formula book and see that (1)

$$\alpha/2 = 0.005 \Rightarrow \alpha = 0.01 \Rightarrow k = 99. \quad (1)$$



Question 6 continued

c) Recall that the confidence interval is
 $(\bar{x} - \bar{y}) \pm t_{n_x+n_y-2} s_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}$

$$\text{where } s_p^2 = \frac{(n_x-1)S_x^2 + (n_y-1)S_y^2}{n_x + n_y - 2}$$

$$s_p^2 = \frac{(8-1)(2.7857) + (6-1)(4)}{6 + 8 - 2}$$

$$= 3.2917$$

$$\Rightarrow s_p = 1.8143 \quad (1)$$

In the formula book, we find use 12 degrees of freedom and use $p = 0.025$.

$$t_{12}(0.025) = 2.179 \quad (1)$$

Using the formula,

$$(51.25 - 49) \pm (2.179)(1.8143) \sqrt{\frac{1}{8} + \frac{1}{6}} \quad (1)$$

$$= (0.115, 4.39) \quad (1)$$



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

$$A \sim N(160, 12^2) \quad R \sim N(140, 10^2)$$

- (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) $(A+R) \sim N(160+140, 12^2+10^2)$ (1)

$(A+R) \sim N(300, 244)$ (1)

ii) $(A_1+A_2) \sim N(2 \times 160, 12^2 \times 2)$

$(A_1+A_2) \sim N(320, 288)$ (1)

b) Here, there are two cases. 2 apples or 1 of each

$P(AA) = \frac{4}{5} \times \frac{3}{4} = \frac{3}{5}$ Let $X = A_1 + A_2$

$P(X > 310) = 1 - P(X \leq 310)$



Question 7 continued

Type this into your calculator.

$$= 1 - 0.2778$$

$$= 0.722$$

$$p(A0) = 1 - 3/5 = 2/5 \text{ (1)}$$

$$\text{Let } Y = A + R$$

$$P(Y > 310) = 1 - P(Y < 310)$$

$$\text{(Using calculator)} = 1 - 0.73897$$

$$= 0.261$$

$$3/5 \text{ (1)} (0.722) + 2/5 \text{ (1)} (0.261) = 0.5377$$

$$c) W = A_1 + \dots + A_m - nR$$

$$E[W] = E[A_1 + \dots + A_m - nR]$$

$$= mE[A] - nE[R]$$

$$= 160m - 140n \text{ (1)}$$

$$\text{Var}(W) = \text{Var}(A_1 + \dots + A_m - nR)$$

$$= m\text{Var}(A) + n^2\text{Var}(R)$$

$$= 144m + 100n^2 \text{ (1)}$$

$$\text{So, } W \sim N(160m + 140n, 144m + 100n^2)$$



Question 7 continued

Recall that the confidence interval has formula

$$\mu \pm z \times \frac{\sigma}{\sqrt{n}} \quad (\mu - z \frac{\sigma}{\sqrt{n}}, \mu + z \frac{\sigma}{\sqrt{n}})$$

So the width is $\frac{2z\sigma}{\sqrt{n}}$

as $n=1$, we have $2z\sigma$ (1)

$$\text{Also, } (\mu - z \frac{\sigma}{\sqrt{n}}) + (\mu + z \frac{\sigma}{\sqrt{n}}) = 2\mu \quad (2)$$

$$\begin{aligned} \text{By (2), we have } \mu &= \frac{1100.08 + 1444.92}{2} \\ &= 1300 \quad (1) \end{aligned}$$

$$\text{So } 1300 = 160m - 140n$$

By (1), we have

$$1444.92 - 1100.08 = 2 \times z \times \sigma \quad (1)$$

$z = 1.96$ using the calculator

$$(1) \quad 102 = \sigma = \sqrt{144m + 100n^2}$$

We now have a pair of simultaneous equations

$$1300 = 160m - 140n, \quad 102 = \sqrt{144m + 100n^2}$$



Question 7 continued

$$102 = \sqrt{144m + 100n^2}$$

$$\Rightarrow 10404 = 144m + 100n^2$$

$$\Rightarrow 10404 - 100n^2 = 144m$$

$$\Rightarrow \frac{10404 - 100n^2}{144} = m$$

Subbing into the other equation yields

$$1300 = 160 \left(\frac{10404 - 100n^2}{144} \right) - 140n$$

$$\Rightarrow 1000n^2 + 1260n - 92340 \quad (1)$$

$$\Rightarrow n = 9 \quad (1) \text{ or } n = -10.26 \quad \text{Using calculator.}$$

Reject the negative as $n > 0$

Sub this into the other equation

$$\frac{10404 - 100(9)^2}{144} = m$$

$$\Rightarrow m = 16 \quad (1) \text{ and } n = 9$$

