Please check the examination details below before entering your candidate information

Candidate surname StemUp Model Solution | Other names

Pearson Edexcel
International Advanced Level

Centre Number | Candidate Number

Tuesday 16 June 2020

Morning (Time: 1 hour 30 minutes) | Paper Reference **WST03/01**

Mathematics

International Advanced Subsidiary/Advanced Level
Statistics S3

You must have:
Mathematical Formulae and Statistical Tables (Blue), 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 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.
- If you change your mind about an answer, cross it out and put your new answer and any working underneath.

Turn over ▶

P64769A

©2020 Pearson Education Ltd.

1/1/1

P 6 4 7 6 9 A 0 1 2 4

Pearson

Leave blank

1. The random variable X has the discrete uniform distribution

$$P(X = x) = \frac{1}{\alpha} \quad \text{for } x = 1,\ 2,\ ...,\ \alpha$$

The mean of a random sample of size n, taken from this distribution, is denoted by $\bar{X}$

(a) Show that $2\bar{X}$ is a biased estimator of α

(2)

A random sample of 6 observations of X is taken and the results are given below.

8 7 3 7 2 9

(b) Use the sample mean to estimate the value of α

(2)

(a) $E(X) = \frac{\alpha + 1}{2}$ (mean of discrete uniform on 1, ..., α)

$E(2\bar{X}) = 2E(\bar{X}) = 2E(X) = 2 \times \frac{\alpha + 1}{2} = \alpha + 1$

$\alpha + 1 \neq \alpha$, so $2\bar{X}$ is a biased estimator of α

(b) $\bar{x} = \frac{8 + 7 + 3 + 7 + 2 + 9}{6} = \frac{36}{6} = 6$

Unbiased estimator $2\bar{X} - 1$: $\hat{\alpha} = 2 \times 6 - 1 = 11$

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

Leave blank

Question 1 continued

Q1

(Total 4 marks)

Leave blank

2. A university awards its graduates a degree in one of three categories, Distinction, Merit or Pass.

Table 1 shows information about a random sample of 200 graduates from three departments, Arts, Humanities and Sciences.

	Arts	Humanities	Sciences	Total
Distinction	22	32	38	92
Merit	15	30	13	58
Pass	18	15	17	50
Total	55	77	68	200

Table 1

Xiu wants to carry out a test of independence between the category of degree and the department.

Table 2 shows some of the values of $\frac{(O-E)^2}{E}$ for this test.

	Arts	Humanities	Sciences	Total
Distinction	0.43	0.33	1.44	2.20
Merit	0.06	2.63	2.29	4.98
Pass	1.31	0.94	0	2.25
				9.43

Table 2

(a) Complete Table 2

(4)

(b) Hence, complete Xiu's hypothesis test using a 5% level of significance. You should state the hypotheses, the degrees of freedom and the critical value used for this test.

(5)

(a) Expected frequencies for Pass: E = (row total × column total) / 200

Arts: $\frac{50 \times 55}{200} = 13.75$, Humanities: $\frac{50 \times 77}{200} = 19.25$

Sciences: $\frac{50 \times 68}{200} = 17$

$\frac{(18-13.75)^2}{13.75} = 1.31$, $\frac{(15-19.25)^2}{19.25} = 0.94$, $\frac{(17-17)^2}{17} = 0$

Pass total = 1.31 + 0.94 + 0 = 2.25

$\chi^2 = 2.20 + 4.98 + 2.25 = 9.43$

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

Leave blank

Question 2 continued

(b) H_0: degree category is independent of department (no association)

H_1: degree category is not independent of department (association)

$\nu = (3 - 1)(3 - 1) = 4$

Critical value $\chi^2_4(5\%) = 9.488$

$9.43 < 9.488$, so not significant: do not reject H_0

There is not enough evidence to show that the degree category is not independent of the department (no evidence of association).

Leave blank

Question 2 continued

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

Leave blank

Question 2 continued

Q2

(Total 9 marks)

Leave blank

3. Each of 7 athletes competed in a 200 metre race and a 400 metre race.

The table shows the time, in seconds, taken by each athlete to complete the 200 metre race.

Athlete	*A*	*B*	*C*	*D*	*E*	*F*	*G*
200 metre race (seconds)	23.4	23.1	22.9	23.7	27.6	24.4	24.1

The finishing order in the 400 metre race is shown below, with athlete *A* finishing in the fastest time.

A *B* *G* *C* *D* *F* *E*

(a) Calculate the Spearman's rank correlation coefficient between the finishing order in the 200 metre race and the finishing order in the 400 metre race.

(5)

(b) Stating your hypotheses clearly, test whether or not there is evidence of a positive correlation between the finishing order in the 200 metre race and the finishing order in the 400 metre race. Use a 5% level of significance.

(4)

The 7 athletes also competed in a long jump competition with the following results.

Athlete	*A*	*B*	*C*	*D*	*E*	*F*	*G*
Long jump (metres)	6.50	6.47	6.12	6.12	6.48	6.38	6.47

Yuliya wants to calculate the Spearman's rank correlation coefficient between the finishing order in the 200 metre race and the finishing order in the long jump for these athletes.

(c) Without carrying out any further calculations, explain how Yuliya should do this.

(2)

(a) Rank 1 = fastest in each race

Athlete	A	B	C	D	E	F	G
200 m rank	3	2	1	4	7	6	5
400 m rank	1	2	4	5	7	6	3
d	2	0	−3	−1	0	0	2
d^2	4	0	9	1	0	0	4

$\Sigma d^2 = 4 + 0 + 9 + 1 + 0 + 0 + 4 = 18$

$$r_s = 1 - \frac{6\Sigma d^2}{n(n^2-1)} = 1 - \frac{6 \times 18}{7 \times 48} = 1 - \frac{108}{336} = \frac{19}{28}$$

$r_s = 0.67857\ldots = 0.679$

(b) H_0: $\rho = 0$ H_1: $\rho > 0$

Critical value ($n = 7$, 5% one-tail) = 0.7143

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

Leave blank

Question 3 continued

0.679 < 0.7143, so not significant: insufficient evidence to reject H_0
There is insufficient evidence to suggest a positive correlation between the finishing order in the 200 metre race and the finishing order in the 400 metre race.

(c) There are tied lengths in the long jump (C and D both 6.12 m, B and G both 6.47 m), so she should give tied athletes the average of the ranks they share (e.g. 6.5 for C and D, 3.5 for B and G).
She should then calculate the product moment correlation coefficient of the ranks: $r_s = \frac{S_{xy}}{\sqrt{S_{xx}\, S_{yy}}}$ (not the $1 - 6\Sigma d^2/n(n^2 - 1)$ formula)

Leave blank

Question 3 continued

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

Leave blank

Question 3 continued

(Total 11 marks)

Q3

Leave blank

4. Luka wants to carry out a survey of students at his school.

He obtains a list of all 280 students.

(a) Explain how he can use this list to select a systematic sample of 40 students.

(2)

Luka is trying to make his own random number table. He generates 400 digits to put in his table. Figure 1 shows the frequency of each digit in his table.

Digit generated	0	1	2	3	4	5	6	7	8	9
Frequency	36	42	33	41	44	43	48	38	32	43

Figure 1

A test is carried out at the 10% level of significance to see if the digits Luka generates follow a uniform distribution.

For this test $\sum \frac{(O-E)^2}{E} = 5.9$

(b) Determine the conclusion of this test.

(3)

The digits generated by Luka are taken two at a time to form two-digit numbers.

Figure 2 shows the frequency of two-digit numbers in his table.

Two-digit numbers generated	00–19	20–39	40–59	60–79	80–99
Frequency	31	49	30	42	48

Figure 2

(c) Test, at the 10% level of significance, whether the two-digit numbers generated by Luka follow a uniform distribution. You should state the hypotheses, the degrees of freedom and the critical value used for this test.

(8)

There are 70 students in Year 12 at his school.

(d) State, giving a reason, the advice you would give to Luka regarding the use of his table of numbers for generating a simple random sample of 10 of the Year 12 students.

(2)

(a) 280 ÷ 40 = 7. Number the students 1 to 280 as listed. Use a random number to select a student from 1 to 7 as the first student, then take every 7th student after that on the list (e.g. 3, 10, 17, ...).

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

Leave blank

Question 4 continued

(b) $\nu = 10 - 1 = 9$, critical value $\chi^2_9(10\%) = 14.684$

$5.9 < 14.684$, so not significant: do not reject H_0

No evidence (at the 10% level) that the digits generated do not follow a uniform distribution.

(c) H_0: a uniform distribution is a good fit for the two-digit numbers

H_1: a uniform distribution is not a good fit for the two-digit numbers

Total = 200, 5 classes: $E = 200 \div 5 = 40$ for each class

Numbers	00–19	20–39	40–59	60–79	80–99	Total
O	31	49	30	42	48	200
E	40	40	40	40	40	200
$(O-E)^2/E$	2.025	2.025	2.5	0.1	1.6	8.25

$$\chi^2 = \Sigma\frac{(O-E)^2}{E} = 2.025 + 2.025 + 2.5 + 0.1 + 1.6 = 8.25$$

$\nu = 5 - 1 = 4$, critical value $\chi^2_4(10\%) = 7.779$

$8.25 > 7.779$, so significant: reject H_0

There is evidence to suggest that the two-digit numbers generated by Luka do not follow a uniform distribution.

(d) To select a simple random sample of 10 from the 70 students (numbered 00 to 69) he needs random two-digit numbers. His table is not suitable for this, since (c) shows its two-digit numbers are not uniformly distributed (it may only be suitable for single random digits), so he should not use it and should use a proper random number table instead.

Leave blank

Question 4 continued

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

Leave blank

Question 4 continued

Q4

(Total 15 marks)

Leave blank

5. A greengrocer is investigating the weights of two types of orange, type *A* and type *B*. She believes that on average type *A* oranges weigh greater than 5 grams more than type *B* oranges. She collects a random sample of 40 type *A* oranges and 32 type *B* oranges and records the weight, *x* grams, of each orange.

The table shows a summary of her data.

	n	$\bar{x}$	$\sum x^2$
Type *A* oranges	40	140.4	790 258
Type *B* oranges	32	134.7	581 430

(a) Calculate unbiased estimates for the variance of the weights of the population of type *A* oranges and the variance of the weights of the population of type *B* oranges.

(3)

(b) Test, at the 5% level of significance, the greengrocer's belief. You should state the hypotheses and the critical value used for this test.

(7)

(c) Explain how you have used the fact that the sample sizes are large in your answer to part (b).

(2)

(a) $s_A^2 = \frac{790258 - 40 \times 140.4^2}{39} = \frac{1771.6}{39} = 45.4256... = 45.4$

$s_B^2 = \frac{581430 - 32 \times 134.7^2}{31} = \frac{819.12}{31} = 26.4232... = 26.4$

(b) $H_0: \mu_A - \mu_B = 5$ $\quad H_1: \mu_A - \mu_B > 5$

s.e. $= \sqrt{\left(\frac{45.4256...}{40} + \frac{26.4232...}{32}\right)} = \sqrt{(1.96136...)} = 1.40048...$

$z = \frac{(140.4 - 134.7) - 5}{1.40048...} = \frac{0.7}{1.40048...} = 0.49983... = 0.50$

5% one-tail critical value: $z = 1.6449$

$0.50 < 1.6449$, so not significant: do not reject H_0

Insufficient evidence to support the greengrocer's belief that type A oranges weigh on average more than 5 grams more than type B oranges.

(c) The sample sizes are large, so by the Central Limit Theorem the sample means $\bar{X}_A$ and $\bar{X}_B$ are both approximately normally distributed, and the unbiased estimates can be used for the population variances: $s_A^2 \approx \sigma_A^2$ and $s_B^2 \approx \sigma_B^2$

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

Leave blank

Question 5 continued

Leave blank

Question 5 continued

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

Leave blank

Question 5 continued

Q5

(Total 12 marks)

Leave blank

6. The number of toasters sold by a shop each week may be modelled by a Poisson distribution with mean 4

A random sample of 35 weeks is taken and the mean number of toasters sold per week is found.

(a) Write down the approximate distribution for the mean number of toasters sold per week from a random sample of 35 weeks.

(2)

The number of kettles sold by the shop each week may be modelled by a Poisson distribution with mean λ

A random sample of 40 weeks is taken and the mean number of kettles sold per week is found. The width of the 99% confidence interval for λ is 2.6

(b) Find an estimate for λ

(4)

A second, independent random sample of 40 weeks is taken and a second 99% confidence interval for λ is found.

(c) Find the probability that only one of these two confidence intervals contains λ

(2)

(a) Mean toasters per week $\bar{X}_T \approx N\left(4, \frac{4}{35}\right)$ i.e. $N(4, 0.114)$

(b) $\bar{X}_K \approx N\left(\lambda, \frac{\lambda}{40}\right)$, 99% CI: $\bar{x} \pm 2.5758\sqrt{(\lambda/40)}$

Width $= 2 \times 2.5758 \times \sqrt{\left(\frac{\lambda}{40}\right)} = 2.6$

$\sqrt{\left(\frac{\lambda}{40}\right)} = \frac{2.6}{5.1516} = 0.50469... \Rightarrow \lambda = 40 \times 0.25472... = 10.188...$

$\lambda = 10.2$

(c) P(an interval contains λ) = 0.99

P(only one contains λ) $= 2 \times 0.99 \times 0.01 = 0.0198$

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

Leave blank

Question 6 continued

Q6

(Total 8 marks)

Leave blank

7. A company makes cricket balls and tennis balls.

The weights of cricket balls, C grams, follow a normal distribution

$$C \sim \mathrm{N}(160, 1.25^2)$$

Three cricket balls are selected at random.

(a) Find the probability that their total weight is more than 475.8 grams.

(4)

The weights of tennis balls, T grams, follow a normal distribution

$$T \sim \mathrm{N}(60, 2^2)$$

Five tennis balls and two cricket balls are selected at random.

(b) Find the probability that the total weight of the five tennis balls and the two cricket balls is more than 625 grams.

(4)

A random sample of n tennis balls $T_1, T_2, T_3, \ldots, T_n$ is taken.

The random variable $Y = (n-1)T_1 - \sum_{r=2}^{n} T_r$

Given that $\mathrm{P}(Y > 40) = 0.0838$ correct to 4 decimal places,

(c) find n.

(8)

(a) $C_1 + C_2 + C_3 \sim N(3 \times 160, 3 \times 1.25^2) = N(480, 4.6875)$

$P(C_1 + C_2 + C_3 > 475.8) = P\left(Z > \frac{475.8 - 480}{\sqrt{4.6875}}\right) = P(Z > -1.9399\ldots)$

$= 0.97380\ldots = 0.974$

(b) $W = T_1 + \ldots + T_5 + C_1 + C_2$

$W \sim N(5 \times 60 + 2 \times 160, 5 \times 2^2 + 2 \times 1.25^2) = N(620, 23.125)$

$P(W > 625) = P\left(Z > \frac{625 - 620}{\sqrt{23.125}}\right) = P(Z > 1.03975\ldots)$

$= 1 - 0.85077\ldots = 0.14923\ldots = 0.149$

(c) $E(Y) = (n-1) \times 60 - (n-1) \times 60 = 0$

$\mathrm{Var}(Y) = (n-1)^2 \times 4 + (n-1) \times 4 = 4(n-1)(n-1+1) = 4n^2 - 4n$

$Y \sim N(0, 4n^2 - 4n)$

$P(Y > 40) = 0.0838 \Rightarrow P(Z < z) = 0.9162 \Rightarrow z = 1.38$

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

Leave blank

Question 7 continued

$$\frac{40 - 0}{\sqrt{(4n^2 - 4n)}} = 1.38 \Rightarrow 4n^2 - 4n = \left(\frac{40}{1.38}\right)^2 = 840.159...$$

$$4n^2 - 4n - 840.159... = 0$$

$$n = \frac{4 \pm \sqrt{16 + 16 \times 840.159...}}{8} = 15.0 \text{ or } -14.0$$

n must be positive, so reject -14: $n = 15$

Leave blank

Question 7 continued

Q7

(Total 16 marks)

TOTAL FOR PAPER: 75 MARKS

END

DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA