Please check the examination details below before entering your candidate information

Candidate surname StemUp Model Solution | Other names

Centre Number | Candidate Number

Pearson Edexcel International Advanced Level

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 (Yellow), 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 the 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 ▶

P72469A

©2022 Pearson Education Ltd.

Q:1/1/

P 7 2 4 6 9 A 0 1 2 8

Pearson

1. The table below shows the number of televised tournaments won and the total number of tournaments won by the top 10 ranked darts players in 2020

Player's rank	Televised tournaments won	Total tournaments won
1	55	135
2	7	33
3	5	17
4	2	14
5	4	9
6	2	5
7	9	36
8	0	15
9	3	3
10	0	13

Michael did not want to calculate Spearman's rank correlation coefficient between player's rank and the rank in televised tournaments won because there would be tied ranks.

(a) Explain how Michael could have dealt with these tied ranks.

(1)

Given that the largest number of total tournaments won is ranked number 1

(b) calculate the value of Spearman's rank correlation coefficient between player's rank and the rank in total tournaments won.

(4)

(c) Stating your hypotheses and critical value clearly, test at the 5% level of significance, whether or not there is evidence of a positive correlation between player's rank and the rank in total tournaments won for these darts players.

(4)

Michael does not believe that there is a positive correlation between player's rank and the rank in total number of tournaments won.

(d) Find the largest level of significance, that is given in the tables provided, which could be used to support Michael's claim.
You must state your critical value.

(1)

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 1 continued

(a) Give each of the tied values the average (mean) of the ranks they would have occupied, e.g. the two 2s share ranks 6 and 7 so each gets rank 6.5

(b) Ranking total tournaments won (largest = 1):

Player's rank	1	2	3	4	5	6	7	8	9	10
Total rank	1	3	4	6	8	9	2	5	10	7
d	0	1	1	2	3	3	−5	−3	1	−3
d^2	0	1	1	4	9	9	25	9	1	9

$\Sigma d^2 = 0 + 1 + 1 + 4 + 9 + 9 + 25 + 9 + 1 + 9 = 68$

$$r_s = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)} = 1 - \frac{6 \times 68}{10(10^2 - 1)} = 1 - \frac{408}{990} = 0.58787\ldots$$

$r_s = 0.588$

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

$n = 10$, 5% one-tailed: critical value = 0.5636 (critical region $r_s \geq 0.5636$)

$0.588 > 0.5636$ so significant, lies in the critical region: reject H_0

There is sufficient evidence of a positive correlation between player's rank and the rank in total tournaments won.

(d) Need the significance level at which 0.588 is not significant:

$n = 10$, one-tailed 2.5%: critical value = 0.6485

$0.588 < 0.6485$ so not significant at 2.5%: largest level is 2.5%

Largest significance level = 2.5% (critical value 0.6485)

Question 1 continued

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 10 marks)

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

2. An experiment is conducted to compare the heat retention of two brands of flasks, brand A and brand B. Both brands of flask have a capacity of 750 ml.

In the experiment 750 ml of boiling water is poured into the flask, which is then sealed. Four hours later the temperature, in °C, of the water in the flask is recorded.

A random sample of 100 flasks from brand A gives the following summary statistics, where x is the temperature of the water in the flask after four hours.

$$\sum x = 7690 \qquad \sum (x - \bar{x})^2 = 669.24$$

(a) Find unbiased estimates for the mean and variance of the temperature of the water, after four hours, for brand A.

(3)

A random sample of 80 flasks from brand B gives the following results, where y is the temperature of the water in the flask after four hours.

$$\bar{y} = 75.9 \qquad s_y = 2.2$$

(b) Test, at the 1% significance level, whether there is a difference in the mean water temperature after four hours between brand A and brand B.
You should state your hypotheses, test statistic and critical value clearly.

(7)

(c) Explain why it is reasonable to assume that $\sigma^2 = s^2$ in this situation.

(1)

(a) $\bar{x} = \frac{7690}{100} = 76.9$

$s_x^2 = \frac{669.24}{99} = 6.76$

Mean = 76.9 °C, variance = 6.76

(b) H_0: $\mu_A = \mu_B$ H_1: $\mu_A \neq \mu_B$

$SE = \sqrt{6.76/100 + 2.2^2/80} = \sqrt{0.1281} = 0.35791...$

$z = \frac{76.9 - 75.9}{0.35791...} = 2.7939... = 2.79$

Two-tailed 1% test: critical value $z = 2.5758$

$2.79 > 2.5758$ so significant, in the critical region: reject H_0

There is sufficient evidence to suggest that the mean water temperature after 4 hours for brand A is different from that for brand B.

(c) Both samples are large (100 and 80), so the sample variances are close to the population variances and s^2 is a good estimate of σ^2.

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

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 11 marks)

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

3. The random variable X is normally distributed with unknown mean μ and known variance σ^2

A random sample of 25 observations of X produced a 95% confidence interval for μ of (26.624, 28.976)

(a) Find the mean of the sample.

(1)

(b) Show that the standard deviation is 3

(3)

The a % confidence interval using the 25 observations has a width of 2.1

(c) Calculate the value of a

(6)

(d) Find the smallest sample size, of observations from X, that would be required to obtain a 95% confidence interval of width at most 1.5

(4)

(a) $\bar{x} = \frac{26.624 + 28.976}{2} = 27.8$

(b) $95\% \Rightarrow z = 1.96$: $28.976 - 26.624 = 2 \times 1.96 \times \frac{\sigma}{\sqrt{25}}$

$2.352 = 3.92 \times \frac{\sigma}{5} \quad \Rightarrow \quad \sigma = \frac{2.352 \times 5}{3.92} = 3$

$\sigma = 3$ as required

(c) $2 \times z \times \frac{3}{\sqrt{25}} = 2.1 \quad \Rightarrow \quad 1.2z = 2.1 \quad \Rightarrow \quad z = 1.75$

$P(Z > 1.75) = 1 - 0.9599 = 0.0401$

Confidence level $= 100 \times (1 - 2 \times 0.0401) = 91.98$

$a = 92.0$ (91.98%)

(d) $2 \times 1.96 \times \frac{3}{\sqrt{n}} \leq 1.5$

$\sqrt{n} \geq \frac{6 \times 1.96}{1.5} = 7.84$

$n \geq 7.84^2 = 61.4656$

Smallest sample size $n = 62$

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

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

Question 3 continued

(Total for Question 3 is 14 marks)

4. Navtej travels to work by train. A train leaves the station every 7 minutes and Navtej's arrival at the station is independent of when the train is due to leave.

(a) Write down a suitable model for the distribution of the time, T minutes, that he has to wait for a train to leave.

(1)

(b) Find the mean and standard deviation of T

(3)

During a 10-week period, Navtej travels to work by train on 46 occasions.

(c) Estimate the probability that the mean length of time that he has to wait for a train to leave is between 3.4 and 3.6 minutes.

(5)

(d) State a necessary assumption for the calculation in part (c).

(1)

(a) $T \sim U[0, 7]$ (continuous uniform distribution on the interval $[0, 7]$)

(b) Mean $= \frac{0 + 7}{2} = 3.5$ minutes

Standard deviation $= \sqrt{(7 - 0)^2/12} = \frac{7}{\sqrt{12}} = 2.0207... = 2.02$ minutes

(c) By the Central Limit Theorem, $\bar{T} \approx N\left(3.5, \frac{49}{12 \times 46}\right) = N\left(3.5, \frac{49}{552}\right)$

Standard deviation of $\bar{T} = 7 \div \sqrt{552} = 0.29793...$

$P(3.4 < \bar{T} < 3.6) = P\left(\frac{3.4 - 3.5}{0.29793...} < Z < \frac{3.6 - 3.5}{0.29793...}\right)$

$= P(-0.3356... < Z < 0.3356...) = 2 \times 0.63143... - 1$

$= 0.26286... \approx 0.263$

(d) The 46 journeys form a random sample, i.e. the waiting times are independent (and $n = 46$ is large enough to use the Central Limit Theorem).

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. A random sample of two observations X_1 and X_2 is taken from a population with unknown mean μ and unknown variance σ^2

(a) Explain why $\dfrac{X_1 - \mu}{\sigma}$ is not a statistic.

(1)

(b) Explain what you understand by an unbiased estimator for μ

(1)

Two estimators for μ are U_1 and U_2 where

$$U_1 = 3X_1 - 2X_2 \qquad \text{and} \qquad U_2 = \frac{X_1 + 3X_2}{4}$$

(c) Show that both U_1 and U_2 are unbiased estimators for μ

(3)

The most efficient estimator among a group of unbiased estimators is the one with the smallest variance.

(d) By finding the variance of U_1 and the variance of U_2 state, giving a reason, the most efficient estimator for μ from these two estimators.

(4)

(a) It contains the unknown population parameters μ and σ, so it cannot be calculated from the sample values alone.

(b) An estimator for μ is unbiased if its expected value is equal to μ.

(c) $E(U_1) = 3E(X_1) - 2E(X_2) = 3\mu - 2\mu = \mu$ *so* U_1 *is unbiased*

$E(U_2) = \frac{1}{4}(E(X_1) + 3E(X_2)) = \frac{1}{4}(\mu + 3\mu) = \mu$ *so* U_2 *is unbiased*

(d) $Var(U_1) = 9Var(X_1) + 4Var(X_2) = 9\sigma^2 + 4\sigma^2 = 13\sigma^2$

$Var(U_2) = \frac{1}{16}Var(X_1) + \frac{9}{16}Var(X_2) = \frac{10}{16}\sigma^2 = \frac{5}{8}\sigma^2$

$Var(U_1) = 13\sigma^2 > Var(U_2) = \frac{5}{8}\sigma^2$

U_2 *is the most efficient estimator for μ as it has the smaller variance.*

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

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 9 marks)

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

6 A particular lift has a maximum load capacity of 700 kg.

The weights of men are normally distributed with mean 80 kg and standard deviation 10 kg.

The weights of women are normally distributed with mean 69 kg and standard deviation 5 kg.

You may assume that weights of people are independent.

(a) Find the probability that when 6 men and 3 women are in the lift, the load exceeds 700 kg.

(4)

A sign in the lift states: "Maximum number of people in the lift is c"

(b) Find the value of c such that the probability of the load exceeding 700 kg is less than 2.5% no matter the gender of the occupants.

(6)

(a) $M \sim N(80, 10^2)$, $W \sim N(69, 5^2)$

$X = M_1 + \ldots + M_6 + W_1 + W_2 + W_3$

$E(X) = 6 \times 80 + 3 \times 69 = 687$, $\text{Var}(X) = 6 \times 100 + 3 \times 25 = 675$

$X \sim N(687, 675)$

$$P(X > 700) = P\left(Z > \frac{700 - 687}{\sqrt{675}}\right) = P(Z > 0.50037\ldots)$$

$= 1 - 0.69153\ldots = 0.30846\ldots \approx 0.308$

(b) Men have the larger mean (and variance), so the worst case is when all the occupants are men. Let Y = total weight of x men in the lift:

$Y \sim N(80x, 100x)$

$$P(Y > 700) = P\left(Z > \frac{700 - 80x}{10\sqrt{x}}\right) < 0.025$$

$$\frac{700 - 80x}{10\sqrt{x}} > 1.96$$

$700 - 80x > 19.6\sqrt{x} \Rightarrow 80x + 19.6\sqrt{x} - 700 < 0$

$$\sqrt{x} < \frac{-19.6 + \sqrt{(19.6^2 + 4 \times 80 \times 700)}}{2 \times 80} = \frac{-19.6 + 473.69\ldots}{160} = 2.8380\ldots$$

$x < 2.8380\ldots^2 = 8.0547\ldots$

So $c = 8$ (people)

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

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 10 marks)

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

7 The following table shows observed frequencies, where x is an integer, from an experiment to test whether or not a six-sided die is biased.

Number on die	1	2	3	4	5	6
Observed frequency	$x + 6$	$x - 8$	$x + 8$	$x - 5$	$x + 4$	$x - 5$

A goodness of fit test is conducted to determine if there is evidence that the die is biased.

(a) Write down suitable null and alternative hypotheses for this test.

(1)

It is found that the null hypothesis is not rejected at the 5% significance level.

(b) Hence

(i) find the minimum value of x

(8)

(ii) determine the minimum number of times the die was rolled.

(2)

(a) H_0: the die is not biased (discrete uniform distribution is a suitable model)

H_1: the die is biased (discrete uniform distribution is not a suitable model)

(b)(i) Total frequency = $6x$, so each expected frequency = $6x \div 6 = x$

O	$x + 6$	$x - 8$	$x + 8$	$x - 5$	$x + 4$	$x - 5$	Total $6x$
E	x	x	x	x	x	x	Total $6x$
$O - E$	6	-8	8	-5	4	-5	
$(O - E)^2$	36	64	64	25	16	25	

$$X^2 = \Sigma\frac{(O-E)^2}{E} = \frac{36 + 64 + 64 + 25 + 16 + 25}{x} = \frac{230}{x}$$

$\nu = 6 - 1 = 5$

Critical value $\chi^2_5(0.05) = 11.070$, critical region $X^2 \geqslant 11.070$

H_0 not rejected so $\frac{230}{x} \leqslant 11.070$

$x \geqslant \frac{230}{11.070} = 20.7768...$

x is an integer so minimum $x = 21$

(check: all expected frequencies = 21 > 5 and smallest $O = 13 > 0$)

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 7 continued

(ii) Number of rolls = total frequency = $6x = 6 \times 21 = 126$

Minimum number of times the die was rolled = 126

Question 7 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

Question 7 continued

Question 7 continued

(Total for Question 7 is 11 marks)

TOTAL FOR PAPER: 75 MARKS

DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA