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 6 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 ▶

P72405A

©2022 Pearson Education Ltd.

Q:1/1/1/1/

P 7 2 4 0 5 A 0 1 2 4

Pearson

1. The weights, x kg, of each of 10 watermelons selected at random from Priya's shop were recorded. The results are summarised as follows

$$\sum x = 114.2 \qquad \sum x^2 = 1310.464$$

(a) Calculate unbiased estimates of the mean and the variance of the weights of the watermelons in Priya's shop.

(3)

Priya researches the weight of watermelons, for the variety she has in her shop, and discovers that the weights of these watermelons are normally distributed with a standard deviation of 0.8 kg

(b) Calculate a 95% confidence interval for the mean weight of watermelons in Priya's shop. Give the limits of your confidence interval to 2 decimal places.

(4)

Priya claims that the confidence interval in part (b) suggests that nearly all of the watermelons in her shop weigh more than 10.5 kg

(c) Use your answer to part (b) to estimate the smallest proportion of watermelons in her shop that weigh less than 10.5 kg

(3)

(a) $\bar{x} = 114.2 \div 10 = 11.42$

$s^2 = \dfrac{1310.464 - 10 \times 11.42^2}{9} = \dfrac{6.3}{9} = 0.7$

Mean = 11.42 kg, variance = $0.7\ kg^2$

(b) 95% $\Rightarrow z = 1.96$

$11.42 \pm 1.96 \times \dfrac{0.8}{\sqrt{10}} = 11.42 \pm 0.49584...$

$= (10.9241..., 11.9158...)$

Confidence interval = (10.92, 11.92) kg

(c) Take the largest plausible mean, the upper limit 11.9158...

$Y \sim N(11.9158..., 0.8^2)$

$P(Y < 10.5) = P\left(Z < \dfrac{10.5 - 11.9158...}{0.8}\right) = P(Z < -1.7698...)$

$= 0.038379... \approx 0.038$

So at least about 0.038 (3.8%) of the watermelons weigh less than 10.5 kg

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

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. Secondary schools in a region conduct ability testing at the start of Year 7 and the start of Year 8. Each year a regional education officer randomly selects 240 Year 7 students and 240 Year 8 students from across the region. The results for last year are summarised in the table below.

	Mean score	**Variance of scores**
Year 7	101	38
Year 8	103	42

The regional education officer claims that there is no difference between the mean scores of these two year groups.

(a) Test the regional education officer's claim at the 1% significance level. You should state your hypotheses, test statistic and critical value clearly.

(7)

(b) Explain the significance of the Central Limit Theorem in part (a).

(1)

(a) H_0: $\mu_7 = \mu_8$ H_1: $\mu_7 \neq \mu_8$ (μ_7, μ_8 = mean score of Year 7, Year 8)

$SE = \sqrt{\left(\frac{38}{240} + \frac{42}{240}\right)} = \sqrt{(1/3)} = 0.57735...$

$z = \frac{103 - 101}{0.57735...} = 3.4641...$ $(= 2\sqrt{3})$

Two-tailed 1% test: critical value $z = 2.5758$ (reject H_0 if $|z| > 2.5758$)

$3.464 > 2.5758$ so z is in the critical region: significant, reject H_0

There is sufficient evidence to suggest the regional education officer's claim is not correct, i.e. there is a difference between the mean scores of the two year groups.

(b) The samples are large (240), so by the Central Limit Theorem the sample means are approximately normally distributed (even though the distribution of scores is unknown), so the z test can be used.

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 2 continued

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 8 marks)

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

3. A medical research team carried out an investigation into the metabolic rate, MR, of men aged between 30 years and 60 years.

A random sample of 10 men was taken from this age group.

The table below shows for each man his MR and his body mass index, BMI. The table also shows the rank for the level of daily physical activity, DPA, which was assessed by the medical research team.

Rank 1 was assigned to the man with the highest level of daily physical activity.

Man	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*
MR (*x*)	6.24	5.94	6.83	6.53	6.31	7.44	7.32	8.70	7.88	7.78
BMI (*y*)	19.6	19.2	23.6	21.4	20.2	20.8	22.9	25.5	23.3	25.1
DPA rank	10	7	9	8	6	3	1	4	5	2

[You may use $S_{xy} = 15.1608$ $S_{xx} = 6.901\,81$ $S_{yy} = 45.304$]

(a) Calculate the value of the product moment correlation coefficient between MR and BMI for these 10 men.

(2)

(b) Use your value of the product moment correlation coefficient to test, at the 5% significance level, whether or not there is evidence of a positive correlation between MR and BMI.
State your hypotheses clearly.

(3)

(c) State an assumption that must be made to carry out the test in part (b).

(1)

(d) Calculate the value of Spearman's rank correlation coefficient between MR and DPA for these 10 men.

(4)

(e) Use a two-tailed test and a 5% level of significance to assess whether or not there is evidence of a correlation between MR and DPA.

(2)

(a) $r = \frac{S_{xy}}{\sqrt{(S_{xx}S_{yy})}} = \frac{15.1608}{\sqrt{(6.90181 \times 45.304)}} = 0.85737...$

$r = 0.857$

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

$n = 10$, 5% one-tailed: critical value = 0.5494

$0.857 > 0.5494$ so significant, reject H_0

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

Question 3 continued

There is evidence of a positive correlation between MR and BMI.

(c) MR and BMI are (bivariate) normally distributed.

(d) Rank MR with the highest MR as rank 1 (same direction as DPA):

Man	A	B	C	D	E	F	G	H	I	J
MR rank	9	10	6	7	8	4	5	1	2	3
DPA rank	10	7	9	8	6	3	1	4	5	2
d	−1	3	−3	−1	2	1	4	−3	−3	1
d^2	1	9	9	1	4	1	16	9	9	1

$\Sigma d^2 = 1 + 9 + 9 + 1 + 4 + 1 + 16 + 9 + 9 + 1 = 60$

$r_s = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)} = 1 - \frac{6 \times 60}{10 \times 99} = 0.63636...$

$r_s = 0.636$

(e) H_0: $\rho = 0$ H_1: $\rho \neq 0$

$n = 10$, 5% two-tailed: critical value = 0.6485

$0.636 < 0.6485$ so not significant, do not reject H_0

There is insufficient evidence of a correlation between MR and DPA.

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

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 12 marks)

4. A survey was carried out with students that had studied Maths, Physics and Chemistry at a college between 2016 and 2020. The students were divided into two groups *A* and *B*.

(a) Explain how a sample could be obtained from this population using quota sampling.

(2)

The students were asked which of the three subjects they enjoyed the most.
The results of the survey are shown in the table.

	Subject enjoyed the most			
	Maths	Physics	Chemistry	Total
Group *A*	16	10	13	39
Group *B*	38	13	10	61
Total	54	23	23	100

(b) Test, at the 5% level of significance, whether the subject enjoyed the most is independent of group.
You should state your hypotheses, expected frequencies, test statistic and the critical value used for this test.

(8)

The Headteacher discovered later that the results were actually based on a random sample of 200 students but had been recorded in the table as percentages.

(c) For the test in part (b), state with reasons the effect, if any, that this information would have on

(i) the null and alternative hypotheses,

(ii) the critical value,

(iii) the value of the test statistic,

(iv) the conclusion of the test.

(4)

(a) Quota sampling is non-random: decide how many students are needed from each group (A and B), e.g. in proportion to group sizes. The interviewer then chooses students from each group until each quota has been met.

(b) H_0: subject enjoyed the most and group are independent

H_1: subject enjoyed the most and group are not independent

Expected frequency $= \frac{\text{row total} \times \text{column total}}{\text{grand total}}$, e.g. $\frac{39 \times 54}{100} = 21.06$

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

Question 4 continued

Expected	Maths	Physics	Chemistry	Total
Group A	21.06	8.97	8.97	39
Group B	32.94	14.03	14.03	61
Total	54	23	23	100

All $E_i > 5$ so no classes need combining.

O	E	$(O - E)^2/E$
16	21.06	1.21575...
10	8.97	0.11827...
13	8.97	1.81058...
38	32.94	0.77728...
13	14.03	0.07562...
10	14.03	1.15758...

$X^2 = \Sigma\frac{(O - E)^2}{E} = 5.1550... = 5.16$

$\nu = (2 - 1)(3 - 1) = 2$

Critical value $\chi^2_2(0.05) = 5.991$

$5.16 < 5.991$ so not in the critical region, do not reject H_0

There is insufficient evidence to suggest that subject enjoyed the most and group are not independent (they appear to be independent).

(c)(i) No change: the hypotheses are the same, the test is still about independence of subject enjoyed and group.

(ii) No change: the table is still 2×3 so $\nu = 2$ and critical value is still 5.991

(iii) All observed and expected frequencies double, so each $(O - E)^2/E$ doubles:

$X^2 = 2 \times 5.155... = 10.31$

(iv) $10.31 > 5.991$ so now significant, reject H_0: there is sufficient evidence that subject enjoyed the most and group are not independent (conclusion is reversed).

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

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 14 marks)

5. Charlie is training for three events: a 1500 m swim, a 40 km bike ride and a 10 km run.

From past experience his times, in minutes, for each of the three events independently have the following distributions.

$S \sim N(41, 5.2^2)$ represents the time for the swim
$B \sim N(81, 4.2^2)$ represents the time for the bike ride
$R \sim N(57, 6.6^2)$ represents the time for the run

(a) Find the probability that Charlie's total time for a randomly selected swim, bike ride and run exceeds 3 hours.

(5)

(b) Find the probability that the time for a randomly selected swim will be at least 20 minutes quicker than the time for a randomly selected run.

(4)

Given that $P(S + B + R > t) = 0.95$

(c) find the value of t

(3)

A triathlon consists of a 1500 m swim, immediately followed by a 40 km bike ride, immediately followed by a 10 km run.

Charlie uses the answer to part (a) to find the probability that, in 6 successive independent triathlons, his time will exceed 3 hours on at least one occasion.

(d) Find the answer Charlie should obtain.

(3)

Jane says that Charlie should not have used the answer to part (a) for the calculation in part (d).

(e) Explain whether or not Jane is correct.

(2)

(a) Let $T = S + B + R$

$T \sim N(41 + 81 + 57, 5.2^2 + 4.2^2 + 6.6^2)$, so $T \sim N(179, 88.24)$

3 hours = 180 minutes

$$P(T > 180) = P\left(Z > \frac{180 - 179}{\sqrt{88.24}}\right) = P(Z > 0.10645\ldots)$$

$= 1 - 0.54239\ldots = 0.45761\ldots \approx 0.458$

(b) Let $D = R - S$, $D \sim N(57 - 41, 6.6^2 + 5.2^2)$, $D \sim N(16, 70.6)$

Need $P(S \leq R - 20) = P(D \geq 20)$

$$P(D \geq 20) = P\left(Z \geq \frac{20 - 16}{\sqrt{70.6}}\right) = P(Z \geq 0.47606\ldots)$$

$= 1 - 0.68298\ldots = 0.31702\ldots \approx 0.317$

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

(c) $P(T > t) = 0.95 \Rightarrow P(T < t) = 0.05$

$\frac{t - 179}{\sqrt{88.24}} = -1.6449$

$t = 179 - 1.6449 \times 9.39361... = 163.548...$

$t = 164$ (minutes) (3 s.f.)

(d) Let X = number of the 6 triathlons with time over 3 hours

$X \sim B(6, 0.45761...)$

$P(X \geqslant 1) = 1 - P(X = 0) = 1 - (1 - 0.45761...)^6 = 1 - 0.54239...^6$

$= 0.97453... \approx 0.975$

(e) Jane is correct. In a triathlon the events follow on immediately, so the times are unlikely to be independent (e.g. Charlie gets tired after the swim and bike ride), and the total also ignores the transition time between events, so the model in (a) and hence the calculation in (d) is not valid.

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 17 marks)

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

6. A farmer sells strawberries in baskets. The contents of each of 100 randomly selected baskets were weighed and the results, given to the nearest gram, are shown below.

Weight of strawberries (grams)	Number of baskets
302 – 303	5
304 – 305	13
306 – 307	10
308 – 309	18
310 – 311	25
312 – 313	20
314 – 315	5
316 – 317	4

The farmer proposes that the weight of strawberries per basket, in grams, should be modelled by a normal distribution with a mean of 310 g and standard deviation 4 g.

Using his model, the farmer obtains the following expected frequencies.

Weight of strawberries (s, grams)	Expected frequency
$s \leqslant 303.5$	a
$303.5 < s \leqslant 305.5$	7.8
$305.5 < s \leqslant 307.5$	13.6
$307.5 < s \leqslant 309.5$	18.4
$309.5 < s \leqslant 311.5$	19.6
$311.5 < s \leqslant 313.5$	16.3
$313.5 < s \leqslant 315.5$	10.6
$s > 315.5$	b

(a) Find the value of a and the value of b. Give your answers correct to one decimal place.

(5)

Question 6 continues on page 23

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

Question 6 continued

Before $s \leqslant 303.5$ and $s > 315.5$ are included, for the remaining cells,

$$\sum \frac{(O-E)^2}{E} = 9.71$$

(b) Using a 5% significance level, test whether the data are consistent with the model. You should state your hypotheses, the test statistic and the critical value used.

(7)

An alternative model uses estimates for the population mean and standard deviation from the data given.

Using these estimated values no expected frequency is below 5

Another test is to be carried out, using a 5% significance level, to assess whether the data are consistent with this alternative model.

(c) State the effect, if any, on the critical value for this test. Give a reason for your answer.

(2)

(a) $S \sim N(310, 4^2)$

$P(S < 303.5) = P\left(Z < \frac{303.5 - 310}{4}\right) = P(Z < -1.625) = 0.052081...$

$a = 100 \times 0.052081... = 5.2081...$ so $a = 5.2$

$P(S > 315.5) = P\left(Z > \frac{315.5 - 310}{4}\right) = P(Z > 1.375) = 0.084566...$

$b = 100 \times 0.084566... = 8.4566...$ so $b = 8.5$

(check: $100 - 7.8 - 13.6 - 18.4 - 19.6 - 16.3 - 10.6 - 5.2 = 8.5$)

$a = 5.2$, $b = 8.5$

(b) H_0: $N(310, 4^2)$ is a suitable model for the weights (data consistent with model)

H_1: $N(310, 4^2)$ is not a suitable model for the weights

All expected frequencies ≥ 5 so no cells need combining (8 cells).

$X^2 = 9.71 + \frac{(5 - 5.2)^2}{5.2} + \frac{(4 - 8.5)^2}{8.5} = 9.71 + 0.00769... + 2.38235...$

$X^2 = 12.100... = 12.1$

$\nu = 8 - 1 = 7$ (no parameters estimated)

Critical value $\chi^2_7(0.05) = 14.067$

$12.1 < 14.067$ so not significant, do not reject H_0

There is insufficient evidence to suggest $N(310, 16)$ is not a suitable model: the data are consistent with the farmer's model.

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

Question 6 continued

(c) Two parameters (mean and standard deviation) are now estimated from the data, so $\nu = 8 - 1 - 2 = 5$. The critical value is reduced, to $\chi^2_5(0.05) = 11.070$

(Total for Question 6 is 14 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