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 Statistics Tables (Yellow), calculator

Total Marks

Candidates may use any calculator permitted by Pearson regulations. Calculators must not have the facility for symbolic 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 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 ▶

P72906A

©2023 Pearson Education Ltd.
J:1/1/1/1/1/

P 7 2 9 0 6 A 0 1 2 8

Pearson

1 A machine fills bottles with mineral water.

The machine is checked every day to ensure that it is working correctly. On a particular day a random sample of 100 bottles is taken. The volume of water, x millilitres, for each bottle is measured and each measurement is coded using

$$y = x - 1000$$

The results are summarised below

$$\sum y = 847 \qquad \sum y^2 = 13\ 510.09$$

(a) (i) Show that the value of the unbiased estimate of the mean of x is 1008.47

(ii) Calculate the unbiased estimate of the variance of x

(4)

The machine was initially set so that the volume of water in a bottle had a mean value of 1010 millilitres.

Later, a test at the 5% significance level is used to determine whether or not the mean volume of water in a bottle has changed. If it has changed then the machine is stopped and reset.

(b) Write down suitable null and alternative hypotheses for a 2-tailed test.

(1)

(c) Find the critical region for $\bar{X}$ in the above test.

(4)

(d) Using your answer to part (a) and your critical region found in part (c), comment on whether or not the machine needs to be stopped and reset.
Give a reason for your answer.

(2)

(e) Explain why the use of $\sigma^2 = s^2$ is reasonable in this situation.

(1)

(a)(i) $\bar{y} = \frac{847}{100} = 8.47$

$\bar{x} = \bar{y} + 1000 = 8.47 + 1000 = 1008.47$ as required

(ii) $s_x^2 = s_y^2 = \frac{13\ 510.09 - 100 \times 8.47^2}{99} = \frac{13\ 510.09 - 7174.09}{99} = \frac{6336}{99}$

$s_x^2 = 64$

(b) H_0: $\mu = 1010$ H_1: $\mu \neq 1010$ (μ = mean volume of water in a bottle, ml)

(c) $\bar{X} \sim N\left(1010, \frac{8^2}{100}\right)$, standard error $= \frac{8}{\sqrt{100}} = 0.8$

2-tailed 5% test: critical values $z = \pm 1.96$

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

$\frac{\bar{X} - 1010}{0.8} = -1.96 \Rightarrow \bar{X} = 1010 - 1.96 \times 0.8 = 1008.432$

$\frac{\bar{X} - 1010}{0.8} = 1.96 \Rightarrow \bar{X} = 1010 + 1.96 \times 0.8 = 1011.568$

Critical region: $\bar{X} \leqslant 1008.432$ or $\bar{X} \geqslant 1011.568$

(d) $1008.432 < 1008.47 < 1011.568$, so $\bar{x} = 1008.47$ is not in the critical region (not significant, do not reject H_0).

There is no evidence the mean volume has changed, so the machine does not need to be stopped and reset.

(e) The sample size ($n = 100$) is large, so s^2 will be a good (reliable) estimate of σ^2.

Question 1 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 1 continued

(Total for Question 1 is 12 marks)

2 The table shows the season's best times, x seconds, for the 8 athletes who took part in the 200 m final in the 2021 Tokyo Olympics. It also shows their finishing position in the race.

Athlete	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*
Season's best time	19.89	19.83	19.74	19.84	19.91	19.99	20.13	20.10
Finishing position	1	2	3	4	5	6	7	8

Given that the fastest season's best time is ranked number 1

(a) calculate the value of the Spearman's rank correlation coefficient for these data.

(4)

(b) Stating your hypotheses clearly, test, at the 1% level of significance, whether or not there is evidence of a positive correlation between the rank of the season's best time and the finishing position for these athletes.

(4)

Chris suggests that it would be better to use the actual finishing time, y seconds, of these athletes rather than their finishing position.

Given that

$$S_{xx} = 0.1286875 \qquad S_{yy} = 0.55275 \qquad S_{xy} = 0.225175$$

(c) calculate the product moment correlation coefficient between the season's best time and the finishing time for these athletes.
Give your answer correct to 3 decimal places.

(2)

(d) Use your value of the product moment correlation coefficient to test, at the 1% level of significance, whether or not there is evidence of a positive correlation between the season's best time and the finishing time for these athletes.

(2)

(a) Rank season's best time (fastest = 1):

Athlete	A	B	C	D	E	F	G	H
SBT rank	4	2	1	3	5	6	8	7
Position	1	2	3	4	5	6	7	8
d	3	0	−2	−1	0	0	1	−1
d^2	9	0	4	1	0	0	1	1

$\Sigma d^2 = 9 + 0 + 4 + 1 + 0 + 0 + 1 + 1 = 16$

$r_s = 1 - \frac{6\Sigma d^2}{n(n^2-1)} = 1 - \frac{6 \times 16}{8 \times 63} = \frac{17}{21} = 0.80952...$

$r_s = 0.810$ (3 s.f.)

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

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

$n = 8$, 1% one-tailed: critical value $r_s = 0.8333$ (CR: $r_s \geqslant 0.8333$)

$0.810 < 0.8333$ so not significant, do not reject H_0

There is no evidence of a positive correlation between the rank of season's best time and finishing position for these athletes.

(c) $r = \dfrac{S_{xy}}{\sqrt{(S_{xx}S_{yy})}} = \dfrac{0.225175}{\sqrt{(0.1286875 \times 0.55275)}} = 0.844282...$

$r = 0.844$ (3 d.p.)

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

$n = 8$, 1% one-tailed: critical value $r = 0.7887$ (CR: $r \geqslant 0.7887$)

$0.844 > 0.7887$ so significant, reject H_0

There is evidence of a positive correlation between season's best time and finishing time for these athletes.

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

Question 2 continued

(Total for Question 2 is 12 marks)

3 A mobile phone company offers an insurance policy to its customers when they purchase a mobile phone. The company conducted a survey on the age of the customers and whether or not claims were made.

A random sample of 1200 customers from this company was investigated for 2020 and the results are shown in the table below.

		Claim made in 2020	No claim made in 2020	Total
Age	**17 – 20 years**	24	176	200
	21 – 50 years	48	652	700
	51 years and over	14	286	300
	Total	86	1114	1200

The data are to be used to determine whether or not making a claim is independent of age.

(a) Calculate the expected frequencies for the age group 51 years and over that

(i) made a claim in 2020

(ii) did not make a claim in 2020

(2)

The 4 classes of customers aged between 17 and 50 give a value of

$\sum \frac{(O-E)^2}{E} = 7.123$ correct to 3 decimal places.

(b) Test, at the 1% level of significance, whether or not making a claim is independent of age. Show your working clearly, stating your hypotheses, the degrees of freedom, the test statistic and the critical value used.

(7)

(a)(i) $E = \frac{86 \times 300}{1200} = 21.5$

(ii) $E = \frac{1114 \times 300}{1200} = 278.5$

(b) H_0: making a claim and age are independent

H_1: making a claim and age are not independent

O	E	$(O-E)^2/E$
14	21.5	$(14-21.5)^2 \div 21.5 = 2.61627...$
286	278.5	$(286-278.5)^2 \div 278.5 = 0.20197...$

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

$X^2 = \Sigma\frac{(O-E)^2}{E} = 7.123 + 2.61627... + 0.20197... = 9.9412...$

$X^2 = 9.94$

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

Critical value $\chi^2_2(0.01) = 9.210$ (CR: $X^2 \geqslant 9.210$)

$9.94 > 9.210$ so in the critical region, significant, reject H_0

There is sufficient evidence to suggest that making a claim is not independent of age (making a claim and age are associated).

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 9 marks)

4 A research student is investigating the number of children who are girls in families with 4 children.

The table below shows her results for 200 such families.

Number of girls	0	1	2	3	4
Frequency	15	68	69	38	10

The research student suggests that a binomial distribution with $p = \frac{1}{2}$ could be a suitable model for the number of children who are girls in a family of 4 children.

(a) Using her results and a 5% significance level, test the research student's claim. You should state your hypotheses, expected frequencies, test statistic and the critical value used.

(8)

The research student decides to refine the model and retains the idea of using a binomial distribution but does not specify the probability that the child is a girl.

(b) Use the data in the table to show that the probability that a child is a girl is 0.45

(2)

The research student uses the probability from part (b) to calculate a new set of expected frequencies, none of which are less than 5

The statistic $\sum \frac{(O-E)^2}{E}$ is evaluated and found to be 2.47

(c) Test, at the 5% significance level, whether using a binomial distribution is suitable to model the number of children who are girls in a family of 4 children. You should state your hypotheses and the critical value used.

(4)

(a) H_0: B(4, 0.5) is a suitable model for the number of girls

H_1: B(4, 0.5) is not a suitable model for the number of girls

$E = 200 \times P(G = g)$, e.g. $200 \times 0.5^4 = 12.5$, $200 \times 4 \times 0.5^4 = 50$, $200 \times 6 \times 0.5^4 = 75$

Girls	0	1	2	3	4
O	15	68	69	38	10
E	12.5	50	75	50	12.5
$(O-E)^2/E$	0.5	6.48	0.48	2.88	0.5

All $E \geq 5$ so no classes need combining.

$X^2 = \Sigma \frac{(O-E)^2}{E} = 0.5 + 6.48 + 0.48 + 2.88 + 0.5 = 10.84$

$\nu = 5 - 1 = 4$ (no parameters estimated)

Critical value $\chi^2_4(0.05) = 9.488$ (CR: $X^2 \geq 9.488$)

$10.84 > 9.488$ so significant, reject H_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 4 continued

There is sufficient evidence to say the research student's claim is not supported:
B(4, 0.5) is not a suitable model.

(b) Total number of girls = $0 \times 15 + 1 \times 68 + 2 \times 69 + 3 \times 38 + 4 \times 10$
$= 0 + 68 + 138 + 114 + 40 = 360$

Total children = $200 \times 4 = 800$, $p = \frac{360}{800} = 0.45$ as required

(c) H_0: a binomial distribution is a suitable model
H_1: a binomial distribution is not a suitable model
$\nu = 5 - 1 - 1 = 3$ (p estimated from the data)
Critical value $\chi^2_3(0.05) = 7.815$ (CR: $X^2 \geq 7.815$)
$2.47 < 7.815$ so not significant, do not reject H_0
There is no significant evidence to say that a binomial distribution is not a suitable model for the number of girls in a family of 4 children.

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 Claire grows strawberries on her farm. She wants to compare two brands of fertiliser, brand *A* and brand *B*.

She grows two sets of plants of the same variety of strawberries under the same conditions, fertilising one set with brand *A* and the other with brand *B*.

The yields per plant, in grams, from each set of plants are summarised below.

	Mean	**Standard deviation**	**Number of plants**
Fertiliser *A*	1377	17.8	50
Fertiliser *B*	1368	18.4	40

(a) Stating your hypotheses clearly, carry out a suitable test to assess whether the mean yield from plants using fertiliser *A* is greater than the mean yield from plants using fertiliser *B*.
Use a 1% level of significance and state your test statistic and critical value.

(7)

The total cost of fertiliser *A* for Claire's 50 plants was £75

The total cost of fertiliser *B* for Claire's 40 plants was £50

Claire sells all her strawberries at £3 per kilogram.

(b) Use this information, together with your answer in part (a), to advise Claire on which of the two brands of fertiliser she should use next year in order to maximise her expected profit per plant, giving a reason for your answer.

(3)

(a) H_0: $\mu_A = \mu_B$ H_1: $\mu_A > \mu_B$ (mean yield per plant, g)

$se = \sqrt{\left(\frac{17.8^2}{50} + \frac{18.4^2}{40}\right)} = \sqrt{14.8008} = 3.84718...$

$z = \frac{1377 - 1368}{3.84718...} = 2.33937... = 2.34$

1% one-tailed: critical value $z = 2.3263$ (CR: $z \geq 2.3263$)

$2.34 > 2.3263$ so in the critical region, significant, reject H_0

There is sufficient evidence to support that the mean yield from plants using fertiliser A is greater than the mean yield from plants using fertiliser B.

(b) Expected profit per plant = income − fertiliser cost per plant

A: $3 \times 1.377 - \frac{75}{50} = 4.131 - 1.50 = £2.63(1)$

B: $3 \times 1.368 - \frac{50}{40} = 4.104 - 1.25 = £2.85(4)$

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

Question 5 continued

Although A gives a (significantly) greater mean yield, the extra yield is small,
and B gives the greater expected profit per plant (£2.85 > £2.63),
so Claire should use fertiliser B next year.

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

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

Question 5 continued

(Total for Question 5 is 10 marks)

6 A garden centre sells bags of stones and large bags of gravel.

The weight, X kilograms, of stones in a bag can be modelled by a normal distribution with unknown mean μ and known standard deviation 0.4

The stones in each of a random sample of 36 bags from a large batch is weighed. The total weight of stones in these 36 bags is found to be 806.4 kg

(a) Find a 98% confidence interval for the mean weight of stones in the batch.

(4)

(b) Explain why the use of the Central Limit theorem is not required to answer part (a)

(1)

The manufacturer of these bags of stones claims that bags in this batch have a mean weight of 22.5 kg

(c) Using your answer to part (a), comment on the claim made by the manufacturer.

(2)

The weight, Y kilograms, of gravel in a large bag can be modelled by a normal distribution with mean 850 kg and standard deviation 5 kg

A builder purchases 10 large bags of gravel.

(d) Find the probability that the mean weight of gravel in the 10 large bags is less than 848 kg

(3)

(a) $\bar{x} = \frac{806.4}{36} = 22.4$

98% confidence: $z = 2.3263$

$\bar{x} \pm z\frac{\sigma}{\sqrt{n}} = 22.4 \pm 2.3263 \times \frac{0.4}{\sqrt{36}} = 22.4 \pm 0.155086...$

$= (22.2449..., 22.5550...)$

98% confidence interval = (22.2, 22.6) kg

(b) The weights X of stones are (given to be) normally distributed, so $\bar{X}$ is exactly normally distributed whatever the sample size.

(c) 22.5 lies within the confidence interval ($22.24 < 22.5 < 22.56$), so there is no reason to doubt the manufacturer's claim.

(d) $\bar{Y} \sim N\left(850, \frac{5^2}{10}\right)$, i.e. $\bar{Y} \sim N(850, 2.5)$

$P(\bar{Y} < 848) = P\left(Z < \frac{848 - 850}{\sqrt{2.5}}\right) = P(Z < -1.2649...)$

$= 0.10295... = 0.103$

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

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)

7 At a particular supermarket, the times taken to serve each customer in a queue at a standard checkout may be modelled by a normal distribution with mean 240 seconds and standard deviation 20 seconds.

There is a queue of 3 customers at a standard checkout.

Making a reasonable assumption about the times taken to serve these customers,

(a) find the probability that the total time taken to serve the 3 customers will be less than 11 minutes.

(3)

(b) State the assumption you have made in part (a)

(1)

In the supermarket there is also an express checkout, which is reserved for customers buying 10 or fewer items. The time taken to serve a customer at this express checkout may be modelled by a normal distribution with mean 100 seconds and standard deviation 8 seconds.

On a particular day Jiang has 8 items to pay for and has to choose whether to join a queue of 3 customers waiting at a standard checkout or a queue of 7 customers waiting at the express checkout.

Using a similar assumption to that made in part (a),

(c) find the probability that the total time taken to serve the 3 customers at the standard checkout will exceed the total time taken to serve the 7 customers at the express checkout.

(4)

(a) P = time (s) to serve a customer at a standard checkout

$P \sim N(240, 20^2)$

$Q = P_1 + P_2 + P_3 \sim N(3 \times 240, 3 \times 20^2)$, so $Q \sim N(720, 1200)$

11 minutes = 660 seconds

$$P(Q < 660) = P\left(Z < \frac{660 - 720}{\sqrt{1200}}\right) = P(Z < -1.7320\ldots)$$

$= 0.041632\ldots = 0.0416$

(b) The times taken to serve the customers are independent of each other.

(c) Let R = time (s) to serve a customer at the express checkout, $R \sim N(100, 8^2)$

$S = (P_1 + P_2 + P_3) - (R_1 + R_2 + \ldots + R_7)$

$E(S) = 720 - 7 \times 100 = 20$

$Var(S) = 1200 + 7 \times 8^2 = 1200 + 448 = 1648$, $S \sim N(20, 1648)$

$$P(S > 0) = P\left(Z > \frac{0 - 20}{\sqrt{1648}}\right) = P(Z > -0.49266\ldots)$$

$= 0.68887\ldots = 0.689$

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 8 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