Oxford Cambridge and RSA

Accredited

A Level Further Mathematics B (MEI)
Y422 Statistics Major
Sample Question Paper

Date – Morning/Afternoon
Time allowed: 2 hours 15 minutes

OCR supplied materials:
- Printed Answer Booklet
- Formulae Further Mathematics B (MEI)

You must have:
- Printed Answer Booklet
- Formulae Further Mathematics B (MEI)
- Scientific or graphical calculator

Model Answers

INSTRUCTIONS
- Use black ink. HB pencil may be used for graphs and diagrams only.
- Complete the boxes provided on the Printed Answer Booklet with your name, centre number and candidate number.
- Answer **all** the questions.
- **Write your answer to each question in the space provided in the Printed Answer Booklet.**
- Additional paper may be used if necessary but you must clearly show your candidate number, centre number and question number(s).
- Do **not** write in the bar codes.
- You are permitted to use a scientific or graphical calculator in this paper.
- Final answers should be given to a degree of accuracy appropriate to the context.

INFORMATION
- The total number of marks for this paper is **120**.
- The marks for each question are shown in brackets **[ ]**.
- You are advised that an answer may receive **no marks** unless you show sufficient detail of the working to indicate that a correct method is used. You should communicate your method with correct reasoning.
- The Printed Answer Booklet consists of **20** pages. The Question Paper consists of **16** pages.

Section A (30 marks)

Answer **all** the questions.

1 In a promotion for a new type of cereal, a toy dinosaur is included in each pack. There are three different types of dinosaur to collect. They are distributed, with equal probability, randomly and independently in the packs. Sam is trying to collect all three of the dinosaurs.

(i) Find the probability that Sam has to open only 3 packs in order to collect all three dinosaurs. **[1]**

$P = 1 \times \frac{2}{3} \times \frac{1}{3} = \frac{2}{9}$

Sam continues to open packs until she has collected all three dinosaurs, but once she has opened 6 packs she gives up even if she has not found all three. The random variable X represents the number of packs which Sam opens.

(ii) Complete the table below, using the copy in the Printed Answer Booklet, to show the probability distribution of X.

r	3	4	5	6
$P(X = r)$	$\frac{2}{9}$	$\frac{2}{9}$	$\frac{14}{81}$	$\frac{31}{81}$

$\sum P(X=r) = 1$ so $P(X=6) = 1 - \frac{2}{9} - \frac{2}{9} - \frac{14}{81} = \frac{31}{81}$

[1]

(iii) In this question you must show detailed reasoning.

Find

- $E(X)$ and
- $\text{Var}(X)$. **[5]**

$E(X) = \sum r P(X=r) = \left(3 \times \frac{2}{9}\right) + \left(4 \times \frac{2}{9}\right) + \left(5 \times \frac{14}{81}\right) + \left(6 \times \frac{31}{81}\right)$

$= \frac{382}{81} / 4.716$

$E(X^2) = \sum r^2 P(X=r) = \left(3^2 \times \frac{2}{9}\right) + \left(4^2 \times \frac{2}{9}\right) + \left(5^2 \times \frac{14}{81}\right) + \left(6^2 \times \frac{31}{81}\right)$

$= \frac{1916}{81} / 23.654$

$\text{Var} X = E(X^2) - [E(X)]^2 = \frac{1916}{81} - \left(\frac{382}{81}\right)^2$

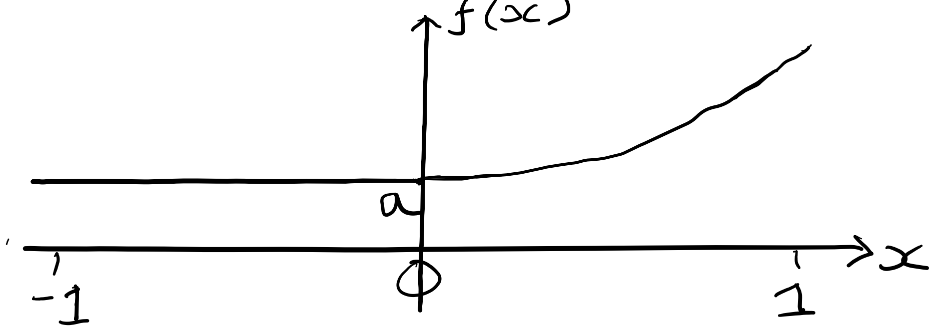
$= 1.413$

2 The continuous random variable X takes values in the interval $-1 \le x \le 1$ and has probability density function

$$f(x) = \begin{cases} a & -1 \le x < 0 \\ a + x^2 & 0 \le x \le 1 \end{cases}$$

where a is a constant.

(i) (*A*) Sketch the probability density function. **[2]**



(*B*) Show that $a = 1/3$ **[3]**

Total area under pdf = 1

$(a \times 1) + \int_0^1 a + x^2 \, dx = 1$

$a + \left[ax + \frac{x^3}{3}\right]_0^1 = 1$ so $a + \left(a + \frac{1}{3} - 0\right) = 1$

$2a + \frac{1}{3} = 1$ so $2a = \frac{2}{3} \Rightarrow a = \frac{1}{3}$

(ii) Find

(*A*) $P\left(X < \frac{1}{2}\right)$ $= P(-1 < X < 0) + P\left(X < \frac{1}{2}\right)$ **[2]**

$= \frac{1}{3} + \int_0^{1/2} \frac{1}{3} + x^2 \, dx = \frac{1}{3} + \left[\frac{1}{3}x + \frac{1}{3}x^3\right]_0^{1/2}$

$= \frac{1}{3} + \frac{1}{3}\left(\frac{1}{2}\right) + \frac{1}{3}\left(\frac{1}{2}\right)^3 = \frac{1}{3} + \frac{1}{6} + \frac{1}{24}$

$= \frac{13}{24}$

(*B*) the mean of X. **[2]**

$$E(x) = \int x f(x)\,dx = \int_{-1}^{0} \frac{1}{3}x\,dx + \int_{0}^{1} \frac{1}{3}x + x^3\,dx$$

$$= \left[\frac{1}{6}x^2\right]_{-1}^{0} + \left[\frac{1}{6}x^2 + \frac{1}{4}x^4\right]_{0}^{1}$$

$$= -\frac{1}{6} + \frac{1}{6} + \frac{1}{4} = \frac{1}{4}$$

(iii) Show that the median of X satisfies the equation $2m^3 + 2m - 1 = 0$. **[3]**

Area from -1 to 0 is $\frac{1}{3}$ so

$$\frac{1}{3} + \int_{0}^{m} \frac{1}{3} + x^2\,dx = 0.5$$

$$\int_{0}^{m} \frac{1}{3} + x^2\,dx = \frac{1}{6} \text{ so } \left[\frac{1}{3}x + \frac{1}{3}x^3\right]_{0}^{m} = \frac{1}{6}$$

$$\frac{1}{3}m + \frac{1}{3}m^3 = \frac{1}{6} \xRightarrow{\times 6} 2m + 2m^3 = 1$$

$$\therefore 2m^3 + 2m - 1$$

3 A researcher is investigating factors that might affect how many hours per day different species of mammals spend asleep.

First she investigates human beings. She collects data on body mass index, x, and hours of sleep, y, for a random sample of people. A scatter diagram of the data is shown in Fig. 3.1 together with the regression line of y on x.

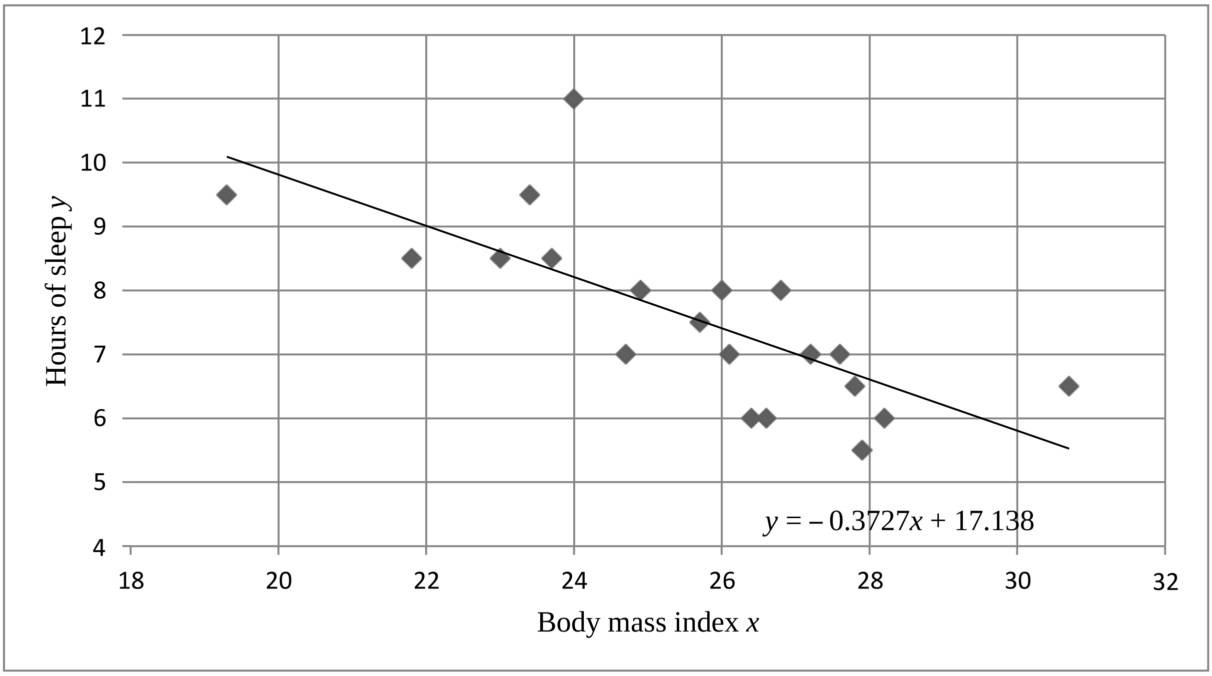


Fig. 3.1

(i) Calculate the residual for the data point which has the residual with the greatest magnitude. **[3]**

The point with the largest residual, the one furthest from the line, is (24, 11).
When $x = 24$, $y = -0.3727(24) + 17.138$
$= 8.1932$
Residual $= 11 - 8.1932 = 2.8068$

(ii) Use the equation of the regression line to estimate the mean number of hours spent asleep by a person with body mass index commenting briefly on each of your predictions.

(A) 26, $y = -0.3727(26) + 17.138 = 8.1932$.
This is interpolated and points lie close to the line so a good estimate

(B) 16, $y = -0.3727(16) + 17.138 = 11.1748$.
This is extrapolated so less reliable estimate

[4]

The researcher then collects additional data for a large number of species of mammals and analyses different factors for effect size. Definitions of the variables measured for a typical animal of the species, the correlations between these variables, and guidelines often used when considering effect size are given in Fig. 3.2.

Variable	**Definition**
Body mass	Mass of animal in kg
Brain mass	Mass of brain in g
Hours of sleep/day	Number of hours per day spent asleep
Life span	How many years the animal lives
Danger	A measure of how dangerous the animal's situation is when asleep, taking into account predators and how protected the animal's den is: higher value indicates greater danger.

Correlations (pmcc)	Body Mass	Brain Mass	Hours of sleep/day	Life span	Danger
Body Mass	1.00				
Brain Mass	0.93	1.00			
Hours of sleep/day	−0.31	−0.36	1.00		
Life span	0.30	0.51	−0.41	1.00	
Danger	0.13	0.15	−0.59	0.06	1.00

Product moment correlation coefficient	**Effect size**
0.1	Small
0.3	Medium
0.5	Large

Fig. 3.2

(iii) State two conclusions the researcher might draw from these tables, relevant to her investigation into how many hours mammals spend asleep. **[2]**

The only factor with a large effect size when correlated with hours of sleep is danger. it seems the more dangerous the situation, the less time the animal sleeps.

One of the researcher's students notices the high correlation between body mass and brain mass and produces a scatter diagram for these two variables, shown in Fig. 3.3 below.

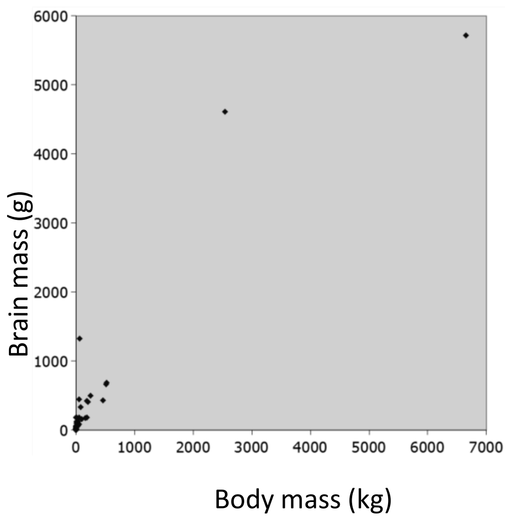


Fig. 3.3

(iv) Comment on the suitability of a linear model for these two variables. **[2]**

There are outliers which makes the rest of the data clustered in the corner. If these outliers are removed then a linear model may fit the data.

Section B (90 marks)

Answer **all** the questions.

4 A fair six-sided dice is rolled repeatedly. Find the probability of the following events.

(i) A five occurs for the first time on the fourth roll. **[1]**

$P(4\text{th roll}) = P(3 \text{ fails}) \times P(1 \text{ success}) = \frac{5}{6}^3 \times \frac{1}{6}$
$= 0.0965$

(ii) A five occurs at least once in the first four rolls. **[2]**

$P(\text{at least 1 five}) = 1 - P(\text{no fives})$
$= 1 - \frac{5}{6}^4 = 0.5177$

(iii) A five occurs for the second time on the third roll. **[2]**

$P(\text{2nd on 3rd roll}) = P(\text{5 on 1st and 3rd}) + P(\text{5 on 2nd and 3rd})$
$= \left(\frac{1}{6} \times \frac{5}{6} \times \frac{1}{6}\right) + \left(\frac{5}{6} \times \frac{1}{6} \times \frac{1}{6}\right) = \frac{5}{108} / 0.0463$

(iv) At least two fives occur in the first three rolls. **[2]**

$P(\geq 2 \text{ fives}) = P(2 \text{ fives}) + P(3 \text{ fives})$
$= \frac{5}{108} + \left(\frac{1}{6}\right)^3 = \frac{2}{27} / 0.0741$

The dice is rolled repeatedly until a five occurs for the second time.

(v) Find the expected number of rolls required for two fives to occur. Justify your answer. **[3]**

$X \sim Geo(1/6)$. $E(X) = \frac{1}{p} = \frac{1}{1/6} = 6$

$E(X_1 + X_2) = 6 + 6 = 12$ rolls

5 A particular brand of pasta is sold in bags of two different sizes. The mass of pasta in the large bags is advertised as being 1500 g; in fact it is Normally distributed with mean 1515 g and standard deviation 4.7 g. The mass of pasta in the small bags is advertised as being 500 g; in fact it is Normally distributed with mean 508 g and standard deviation 3.3 g.

(i) Find the probability that the total mass of pasta in 5 randomly selected small bags is less than 2550 g. **[3]**

$S \sim N(508, 3.3^2)$. $E(S_1+S_2+S_3+S_4+S_5)$
$= 5 \times E(S) = 5 \times 508 = 2540$
$Var(S_1+S_2+S_3+S_4+S_5) = 5 \times Var(S) = 5 \times 3.3^2$
$= 54.45$
$S_1+S_2+S_3+S_4+S_5 \sim N(2540, 54.45)$
$P(S_1+S_2+S_3+S_4+S_5 < 2550) = 0.9123$

(ii) Find the probability that the mass of pasta in a randomly selected large bag is greater than three times the mass of pasta in a randomly selected small bag. **[4]**

$S \sim N(508, 3.3^2)$. $L \sim N(1515, 4.7^2)$
$E(L-3S) = E(L) - 3E(S) = 1515 - 3(508)$
$= -9$
$Var(L-3S) = Var L - 3^2 Var S = 4.7^2 - 9(3.3^2)$
$= 120.1$
$L-3S \sim N(-9, 120.1)$ so
$P(L-3S > 0) = 0.2058$

6 Fig. 6 shows the wages earned in the last 12 months by each of a random sample of American males aged between 16 and 65.

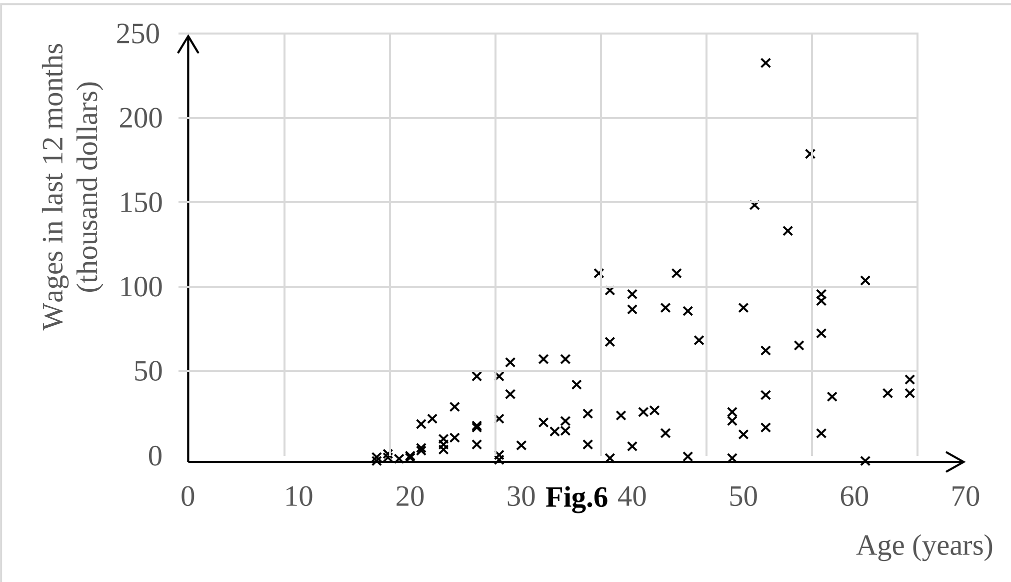


Fig.6

A researcher wishes to test whether the sample provides evidence of a tendency for higher wages to be earned by older men in the age range 16 to 65 in America.

(i) The researcher needs to decide whether to use a test based on Pearson's product moment correlation coefficient or Spearman's rank correlation coefficient. Use the information in Fig. 6 to decide which test is more appropriate. **[2]**

The shape of the scatter diagram is not roughly elliptical so no evidence of a Bivariate Normal population and thus a pmcc test wouldn't be valid so spearman's is more valid.

(ii) Should it be a one-tail or a two-tail test? Justify your answer. **[1]**

It should be a one-tail test as we looking for evidence of a positive relationship.

7 A newspaper reports that the average price of unleaded petrol in the UK is 110.2 p per litre.

The price, in pence, of a litre of unleaded petrol at a random sample of 15 petrol stations in Yorkshire is shown below together with some output from software used to analyse the data.

116.9	114.9	110.9	113.9	114.9
117.9	112.9	99.9	114.9	103.9
123.9	105.7	108.9	102.9	112.7

Statistics	
n	15
Mean	111.6733
σ	6.1877
s	6.4048
Σx	1675.1
Σx²	187638.31
Min	99.9
Q1	105.7
Median	112.9
Q3	114.9
Max	123.9

Fig. 7.1

n	15
Kolmogorov-Smirnov test	$p > 0.15$
Null hypothesis	The data can be modelled by a Normal distribution
Alternative hypothesis	The data cannot be modelled by a Normal distribution

Fig. 7.2

(i) Select a suitable hypothesis test to investigate whether there is any evidence that the average price of unleaded petrol in Yorkshire is different from 110.2 p. Justify your choice of test. **[3]**

The Kolmogorov-Smirnov p value is greater than 0.15, which suggests data could be from Normal distribution. As the sample is small and population variance is unknown; a t distribution test will be suitable

(ii) Conduct the hypothesis test at the 5% level of significance. **[8]**

$H_0: \mu = 110.2$ and $H_1: \mu \neq 110.2$ where μ is the population mean petrol price for Yorkshire.

Test statistic; $t = \dfrac{\bar{x} - \mu}{s/\sqrt{n}} = \dfrac{111.6733 - 110.2}{6.4048/\sqrt{15}}$

$= 0.8909$

Critical value for $v = 15 - 1 = 14$ and $p = 5\%$ is 2.145. As $0.8909 < 2.145$, this result is not significant so insufficient evidence to reject H_0, which would suggest the average price of petrol in Yorkshire is different from that in the UK.

8 Natural background radiation consists of various particles, including neutrons. A detector is used to count the number of neutrons per second at a particular location.

(i) State the conditions required for a Poisson distribution to be a suitable model for the number of neutrons detected per second. **[2]**

The neutrons occur independently and randomly at a constant average rate.

The number of neutrons detected per second due to background radiation only is modelled by a Poisson distribution with mean 1.1. $X \sim Po(1.1)$

(ii) Find the probability that the detector detects

(A) no neutrons in a randomly chosen second,

$P(X=0) = e^{-1.1} = 0.3329$ (4sf)

(B) at least 60 neutrons in a randomly chosen period of 1 minute. **[3]**

$\lambda_{new} = 1.1 \times 60 = 66$. $Y \sim Po(66)$.

$P(Y \geq 60) = 1 - P(Y \leq 59) = 1 - 0.214 = 0.786$

A neutron source is switched on. It emits neutrons which should all be contained in a protective casing. The detector is used to check whether any neutrons have not been contained; these are known as stray neutrons.

If the detector detects more than 8 neutrons in a period of 1 second, an alarm will be triggered in case this high reading is due to stray neutrons.

(iii) Suppose that there are no stray neutrons and so the neutrons detected are all due to the background radiation. Find the expected number of times the alarm is triggered in 1000 randomly chosen periods of 1 second. **[3]**

$P(X > 8) = 1 - P(X \leq 7) = 1 - 0.99999757 3$
$= 0.00000242 7$

Expected No $= 0.000002427 \times 1000$
$= 0.002427$

(iv) Suppose instead that stray neutrons are being produced at a rate of 3.4 per second in addition to the natural background radiation. Find the probability that at least one alarm will be triggered in 10 randomly chosen periods of 1 second. You should assume that all stray neutrons produced are detected. **[4]**

$\lambda_{new} = 1.1 + 3.4 = 4.5$ so $X \sim Po(4.5)$.

P(no alarm in 1 secs) $= P(X \leq 8) = 0.95974$.

P(≥ 1 alarm in 10 secs) $= 1 -$ P(no alarm in 10s)

$= 1 - 0.95974^{10} = 0.337$ (3sf)

9 A random sample of adults in the UK were asked to state their primary source of news: television (**T**), internet (**I**), newspapers (**N**) or radio (**R**). The responses were classified by age group, and an analysis was carried out to see if there is any association between age group and primary source of news.

Fig. 9 is a screenshot showing part of the spreadsheet used to analyse the data. Some values in the spreadsheet have been deliberately omitted.

	A	B	C	D	E	F
1	Source	Age group				
2	of news	18-32	33-47	48-64	65+	
3	T	63	61	71	80	275
4	I	33	33	22	12	100
5	N	9	8	11	20	48
6	R	4	9	9	5	27
7		109	111	113	117	450
8						
9		Expected frequencies				
10		66.61	67.83	69.06	71.50	
11		24.22	24.67		26.00	
12		11.63	11.84	12.05	12.48	
13		6.54	6.66	6.78	7.02	
14						
15		Contributions to the test statistic				
16		0.20	0.69	0.05	1.01	
17		3.18	2.82		7.54	
18		0.59		0.09	4.53	
19		0.99	0.82	0.73	0.58	
20					test statistic	25.45

Fig. 9

(i) (*A*) State the sample size. **[1]**

Sample size = 450

(B) Give the name of the appropriate hypothesis test. **[1]**

Chi-Squared Test

(C) State the null and alternative hypotheses. **[1]**

H_0: No association between age and news source

H_1: Some association between age and news source

(ii) Showing your calculations, find the missing values in cells

- D11 $= \frac{113 \times 100}{450} = 25.11$
- D17 $= \frac{(8 - 11.84)^2}{11.84} = 1.25$
- C18. $= \frac{(22 - 25.11)^2}{25.11} = 0.39$ **[4]**

(iii) Complete the appropriate hypothesis test at the 5% level of significance. **[4]**

Test statistic (χ^2) = 25.245.
Critical value for $v = (n-1)(m-1) = (4-1)(4-1) = 9$
n and m are no of rows and columns
and p = 5% is 16.92.
25.45 > 16.92 so this result is significant so sufficient evidence to reject H_0, which would suggest there is some association between age and news source.

(iv) Discuss briefly what the data suggest about primary source of news. You should make a comment for each age group. **[3]**

- For 18-32 and 33-47, the large contributions show that more than expected use the Internet as primary news source.
- For 48-64, the small contributions shows the values are as expected.
- For 65+, the large contributions of 7.54 and 4.53 shows that fewer than expected have the internet and more than expected have newspapers as primary source.

10 The label on a particular size of milk carton states that it contains 1.5 litres of milk. In an investigation at the packaging plant the contents, x litres, of each of 60 randomly selected cartons are measured. The data are summarised as follows.

$$\Sigma x = 89.758 \qquad \Sigma x^2 = 134.280$$

(i) Estimate the variance of the underlying population. **[2]**

Variance $= \dfrac{S_{xx}}{n-1}$. $S_{xx} = \Sigma x^2 - \dfrac{(\Sigma x)^2}{n} = 134.28 - \dfrac{89.758^2}{60}$

$= 0.00502\,3933$.

Variance $= \dfrac{0.0050239333}{60-1} = 0.00008515$ (4sf)

(ii) Find a 95% confidence interval for the mean of the underlying population. **[4]**

$\bar{x} = \dfrac{\Sigma x}{n} = \dfrac{89.758}{60} = 1.49597$.

For 95% interval: $\bar{x} \pm 1.96\sqrt{\dfrac{\sigma}{n}}$

$= 1.49597 \pm 1.96\left(\sqrt{\dfrac{0.00008515}{60}}\right)$

$\Rightarrow 1.4936$ to 1.4983

As n is large, you can use Normal distribution confidence interval

(iii) What does the confidence interval which you have calculated suggest about the statement on the carton? **[1]**

The confidence interval doesn't contain 1.5 litres which suggests the statement that the mean is 1.5 litres is not correct

Each day for 300 days a random sample of 60 cartons is selected and for each sample a 95% confidence interval is constructed.

(iv) Explain why the confidence intervals will not be identical. **[2]**

Each sample will be slightly different and will have different means and variances thus not giving identical confidence intervals.

(v) What is the expected number of confidence intervals to contain the population mean? **[1]**

95% of 300 = $0.95 \times 300 = 285$

11 Two girls, Lili and Hui, play a game with a fair six-sided dice. The dice is thrown 10 times.

$X_1, X_2, \ldots, X_{10}$ represent the scores on the 1st, 2nd, … , 10th throws of the dice.

L denotes Lili's score and $L = 10X_1$.

H denotes Hui's score and $H = X_1 + X_2 + X_3 + \ldots + X_{10}$.

(i) Calculate

for sum of 10 to = 60, each one must be 6

- $P(L = 60) = P(X_1 = 6) = \frac{1}{6}$
- $P(H = 60) = P(X_1 = \ldots = X_{10} = 6) = \frac{1}{6}^{10} = 1.65 \times 10^{8}$ **[3]**

(ii) Without doing any further calculations, explain which girl's score has the greater standard deviation. **[1]**

Lili's score will have a higher SD as Hui's score is a combination of 10 results so is likely to be closer to the mean

(iii) Write down

- the name of the probability distribution of X_1, = Discrete Uniform
- the value of $E(X_1)$, $= \frac{n+1}{2} = \frac{6+1}{2} = 3.5$
- the value of $\text{Var}(X_1)$. **[3]**

$\text{Var}\, X_1 = \frac{1}{12}(n^2-1) = \frac{1}{12}(6^2-1) = \frac{35}{12}$

(iv) Find

(A) $E(L)$, $= E(10X_1) = 10E(X_1) = 10 \times 3.5 = 35$

(B) $\text{Var}(L)$, $\text{Var}(10X_1) = 10^2\text{Var}(X_1) = 100 \times \frac{35}{12} = \frac{875}{3}$

(C) $E(H)$, $= E(X_1 + \ldots + X_{10}) = 10 \times E(X_1) = 10 \times 3.5 = 35$

(D) $\text{Var}(H) = \text{Var}(X_1 + \ldots + X_{10}) = 10 \times \text{Var}\, X_1 = 10 \times \frac{35}{12} = \frac{175}{6}$ **[5]**

The spreadsheet below shows a simulation of 25 plays of the game. The cell E3, highlighted, shows the score when the dice is thrown the fourth time in the first game.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1		**Throw of dice**											**Lili's**	**Hui's**
2		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**		**score**	**score**
3	**Game 1**	3	5	2	1	1	3	1	1	1	4		30	22
4	**Game 2**	6	3	2	4	4	3	5	3	3	5		60	38
5	**Game 3**	6	4	2	6	5	2	1	5	2	3		60	36
6	**Game 4**	1	5	1	6	6	3	1	4	6	2		10	35
7	**Game 5**	4	4	3	1	6	4	4	1	6	2		40	35
8	**Game 6**	2	1	5	1	2	5	1	5	2	3		20	27
9	**Game 7**	1	1	3	4	4	5	6	3	4	2		10	33
10	**Game 8**	1	1	3	6	3	4	4	5	2	3		10	32
11	**Game 9**	2	2	2	4	3	2	1	5	5	6		20	32
12	**Game 10**	3	5	3	3	5	3	4	3	1	1		30	31
13	**Game 11**	5	3	6	5	5	4	2	1	1	5		50	37
14	**Game 12**	6	4	3	2	4	1	3	3	5	3		60	34
15	**Game 13**	2	3	2	1	2	2	2	2	2	1		20	19
16	**Game 14**	4	1	3	3	1	2	6	6	1	3		40	30
17	**Game 15**	5	1	2	6	3	4	6	3	6	4		50	40
18	**Game 16**	3	6	1	1	5	3	1	3	3	3		30	29
19	**Game 17**	5	2	5	2	4	5	2	2	3	4		50	34
20	**Game 18**	3	6	3	5	5	2	3	1	1	2		30	31
21	**Game 19**	6	6	3	1	5	6	3	4	1	6		60	41
22	**Game 20**	2	6	4	5	6	5	2	4	3	3		20	40
23	**Game 21**	5	3	5	4	5	3	3	6	6	1		50	41
24	**Game 22**	6	3	5	5	6	3	5	6	1	1		60	41
25	**Game 23**	5	4	5	5	6	4	2	1	3	6		50	41
26	**Game 24**	3	5	2	3	2	4	3	2	3	3		30	30
27	**Game 25**	5	2	4	2	4	5	2	2	5	2		50	33
28														
29												mean	37.60	33.68
30												sd	17.39	5.77

Fig. 11

(v) Use the simulation to estimate $P(L > 40)$ and $P(H > 40)$. **[2]**

Out of 25 games, 11 have $L > 40$ and 4 have $H > 40$ so $P(L > 40) = \frac{11}{25}$. $P(H > 40) = \frac{4}{25}$

(vi) (*A*) Calculate the exact value of $\mathrm{P}(L > 40)$. **[1]**

$P(L>40) = P(X_1=5)+P(X_1=6) = \frac{1}{6}+\frac{1}{6} = \frac{1}{3}$

(*B*) Comment on how the exact value compares with your estimate of $\mathrm{P}(L > 40)$ in part **(v)**. **[1]**

Estimate is 0.44 and exact value is 0.33 so estimate is a bit off but not massively off considering only 25 trials

Hui wonders whether it is appropriate to use the Central Limit Theorem to approximate the distribution of $X_1 + X_2 + X_3 + \ldots + X_{10}$.

(vii) (*A*) State what type of diagram Hui could draw, based on the output from the spreadsheet, to investigate this. **[1]**

Normal Probability Plot of the 25 values of Hui's scores.

(*B*) Explain how she should interpret the diagram. **[2]**

If it appears as a roughly straight line then the values can be described as Normally distributed so Central Limit Theorem can be used.

(viii) (*A*) Calculate an approximate value of $\mathrm{P}(X_1 + X_2 + X_3 + \ldots + X_{10} > 40)$ using the Central Limit Theorem. **[3]**

$H \sim N(35, 350) \Rightarrow \bar{H} \sim N(35, \frac{350}{12})$

$P(\bar{H} > 40) = P(\bar{H} > 40.5) = 0.154$

continuity correction needed as $X_1 \ldots X_{10}$ were discrete variables and Normal distribution is continuous.

(*B*) Comment on how this value compares with your estimate of $\mathrm{P}(H > 40)$ in part **(v)**. **[1]**

It is very close to the estimate of 0.16.

END OF QUESTION PAPER