Maths with Matt

Physics and maths resources

OCR AS Level Further Mathematics A

H235/Y532 - Statistics

Monday 23 May 2022

Questions and Fully Worked Model Solutions

This booklet places the **question and the solution together**. The mathematics is set out using proper fractions, powers, roots and exact notation, with extra written explanation so that a student can see *why* each step is being taken.

Solutions have been checked against the OCR Monday 23 May 2022 mark scheme. The aim is a clear **full-mark model response**, not merely a list of final answers.

Question wording is reproduced from the user-provided 2022 paper for study/reference. OCR retains copyright in the original examination material.

Contents

Question 1 2

Question 2 5

Question 3 6

Question 4 7

Question 5 8

Question 6 10

Question 7 11

www.mathswithmatt.co.uk

Question 1

Question 1

A geography student chose a certain point in a stream and took measurements of the speed of flow, $v\,\text{m s}^{-1}$, of water at various depths, $d\,\text{m}$, below the surface at that point. The results are shown in the table.

d	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
v	0.8	0.5	0.7	1.2	1.1	1.3	1.6	1.4	0.4

$$n = 9, \quad \sum d = 2.7, \quad \sum v = 9.0, \quad \sum d^2 = 0.96, \quad \sum v^2 = 10.4, \quad \sum dv = 2.85.$$

(a) (i) Explain why d is an example of an independent, controlled variable. **([1])**

(ii) Use two relevant terms to describe the variable v in a similar way. **([1])**

A statistician believes that the point $(0.5, 0.4)$ may be an anomaly.

(b) Calculate the equation of the least squares regression line of v on d for all the points in the table apart from $(0.5, 0.4)$. **([2])**

(c) Use the equation of the line found in part (b) to estimate the value of v when $d = 0.5$. **([1])**

(d) Use your answer to part (c) to comment on the statistician's belief. **([1])**

(e) Use the diagram in the Printed Answer Booklet (which does not illustrate the data in this question) to explain what is meant by "least squares regression line".



([2])

Model solution

(a)(i) The depth is the quantity the student chooses at which to take a measurement; it is not caused by another variable in this experiment.

d is an independent, controlled variable because its value can be selected by the experimenter. *([1])*

(a)(ii) The speed is measured in response to the chosen depth.

v is a dependent, response variable. *([1])*

(b) Remove the suspected point first, then use the regression formula $v = a + bd$. The six supplied sums are adjusted by subtracting that point's contributions.

$$n = 8, \quad \sum d = 2.2, \quad \sum v = 8.6, \quad \sum d^2 = 0.71, \quad \sum dv = 2.65.$$

$$b = \frac{\sum dv - \frac{(\sum d)(\sum v)}{n}}{\sum d^2 - \frac{(\sum d)^2}{n}} = \frac{2.65 - \frac{2.2(8.6)}{8}}{0.71 - \frac{2.2^2}{8}} = \frac{0.285}{0.105} = 2.714285\ldots$$

$$a = \bar{v} - b\bar{d} = \frac{8.6}{8} - 2.714285\ldots\frac{2.2}{8} = 0.328571\ldots$$

The least squares regression line is $v = 0.329 + 2.71d$ (approximately). *([2])*

(c) Substitute the requested depth into the fitted line; this gives the speed predicted by the other eight observations.

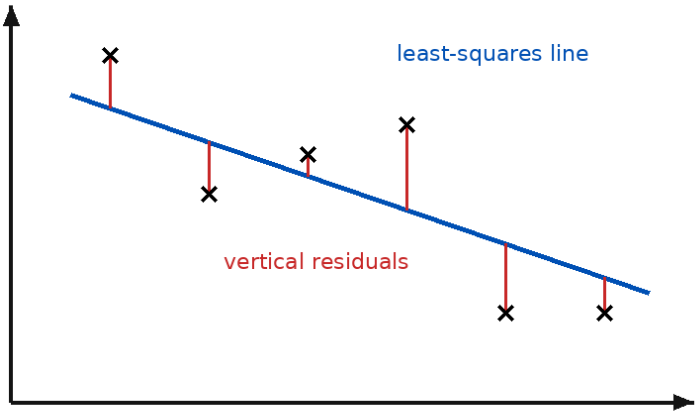
$$v = 0.328571\ldots + 2.714285\ldots(0.5) = 1.685714\ldots$$

Estimated $v = 1.69\,\mathrm{m\,s^{-1}}$. *([1])*

(d) Compare the observed value at $d = 0.5$ with the value predicted by the pattern in the remaining data.

The observed value 0.4 is very different from the estimate 1.69, so the statistician is likely to be right that the point may be an anomaly. *([1])*

(e) A regression line is chosen to make the vertical prediction errors, called residuals, as small as possible overall in the least-squares sense. The completed diagram shows a line and vertical residuals.



It is the line that minimises the sum of the squares of the vertical distances (residuals) from the points to the line. *([2])*

www.mathswithmatt.co.uk

Question 2

Question 2

Eight runners took part in two races. The positions in which the runners finished in the two races are shown in the table.

Runner	A	B	C	D	E	F	G	H
First race	3	1	5	6	2	8	7	4
Second race	4	3	8	7	2	5	6	1

Test at the 5% significance level whether those runners who do better in one race tend to do better in the other. **([7])**

Model solution

The data are paired ranks, so Spearman's rank correlation test is appropriate. A smaller finishing position means doing better, so a positive correlation represents the stated tendency.

$$H_0 : \rho_s = 0, \qquad H_1 : \rho_s > 0,$$

where ρ_s is the population Spearman's rank correlation coefficient.

Runner	A	B	C	D	E	F	G	H
$d = ($*first rank*$) - ($*second rank*$)$	-1	-2	-3	-1	0	3	1	3
d^2	1	4	9	1	0	9	1	9

$$\sum d^2 = 34, qquad r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6(34)}{8(8^2-1)} = \frac{25}{42} = 0.595238\ldots$$

For $n = 8$, the one-tailed 5% critical value is 0.6429. The test statistic does not reach this critical value.

$$0.595238\ldots < 0.6429.$$

Do not reject H_0. There is insufficient evidence at the 5% level that runners who do better in one race tend to do better in the other. **([7])**

Question 3

Question 3

A discrete random variable X has the following probability distribution.

x	1	2	3	4
$P(X = x)$	p	0.31	0.3	p^2

(a) Determine the value of p. **([3])**

(b) It is given that $E(aX + b) = \text{Var}(aX + b) = 23.19$, where a and b are positive constants. Determine the value of a and the value of b. **([6])**

Model solution

(a) Probabilities in a distribution must total 1.

$$p + 0.31 + 0.3 + p^2 = 1 \quad \Rightarrow \quad p^2 + p - 0.39 = 0.$$

$$(p - 0.3)(p + 1.3) = 0,$$

The root -1.3 *cannot be a probability.*

$p = 0.3$. *([3])*

(b) First find the mean and variance of X*. Then use the linear-transformation rules* $E(aX + b) = aE(X) + b$ *and* $\text{Var}(aX + b) = a^2\,\text{Var}(X)$.

$$E(X) = 1(0.3) + 2(0.31) + 3(0.3) + 4(0.09) = 2.18,$$

$$E(X^2) = 1^2(0.3) + 2^2(0.31) + 3^2(0.3) + 4^2(0.09) = 5.68,$$

$$\text{Var}(X) = 5.68 - (2.18)^2 = 0.9276.$$

$$0.9276a^2 = 23.19 \quad \Rightarrow \quad a^2 = 25.$$

Since a *is positive, take the positive square root.*

$$a = 5.$$

$$2.18(5) + b = 23.19 \quad \Rightarrow \quad b = 12.29.$$

$a = 5$ *and* $b = 12.29$. *([6])*

Question 4

Question 4

A school pupil keeps a note of whether her journeys to school and from school are delayed. The results for a random sample of journeys are shown in the table.

Direction of journey	To school	From school
Delayed	64	56
Not delayed	74	106

Test at the 10% significance level whether there is association between delays and the direction of the journey. ([7])

Model solution

This is a two-way frequency table, so use a chi-squared test of association.

H_0 : *delay and direction of journey are independent*; H_1 : *there is an association.*

The column totals are 138 *and* 162, *the row totals are* 120 *and* 180, *and the grand total is* 300. *Under* H_0, *each expected frequency is* $\frac{(\textit{row total})(\textit{column total})}{300}$.

	To school	*From school*
Delayed	55.2	64.8
Not delayed	82.8	97.2

Use the continuity-corrected statistic for this 2×2 *table.*

$$\chi^2 = \frac{(|64-55.2|-0.5)^2}{55.2} + \frac{(|56-64.8|-0.5)^2}{64.8} + \frac{(|74-82.8|-0.5)^2}{82.8} + \frac{(|106-97.2|-0.5)^2}{97.2}$$

$$= 1.248 + 1.063 + 0.832 + 0.709 = 3.85.$$

There is $(2-1)(2-1) = 1$ *degree of freedom. At the* 10% *level the critical value is* 2.706.

$$3.85 > 2.706.$$

Reject H_0. *There is significant evidence at the* 10% *level of an association between delays and direction of journey.* ([7])

Question 5

Question 5

The manager of an emergency response hotline believes that calls are made to the hotline independently and at constant average rate throughout the day. From a small random sample of the population, the manager finds that the mean number of calls made in a 1-hour period is 14.4.
Let R denote the number of calls made in a randomly chosen 1-hour period.

(a) Using evidence from the small sample, state a suitable distribution with which to model R. You should give the value(s) of any parameter(s). **([1])**

(b) In this part of the question, use the distribution and value(s) of the parameter(s) from your answer to part (a).

(i) Find $P(R > 20)$. **([2])**

(ii) Given that $P(R = r) > P(R = r + 1)$, show algebraically that $r > 13.4$. **([2])**

(iii) Hence write down the mode of the distribution. **([1])**

(c) The manager also finds, from records over many years, that the modal value of R is 10. Use this result to comment on the validity of the distribution used in part (b). **([1])**

(d) Assume now that the type of distribution used in part (b) is valid. Find the range(s) of values of the parameter(s) of this distribution that would correspond to the modal value of R being 10. **([2])**

Model solution

(a) Counts in a fixed time period with independent events at a constant average rate are modelled by a Poisson distribution. Its parameter is the mean count.

$R \sim \text{Po}(14.4)$. *([1])*

(b)(i) Use the complement because cumulative Poisson probabilities are readily evaluated up to 20.

$$P(R > 20) = 1 - P(R \leq 20) = 0.060351\ldots$$

$P(R > 20) = 0.0604$. *([2])*

(b)(ii) Compare consecutive Poisson probabilities; cancelling their common factors identifies where the probabilities begin to decrease.

$$\frac{e^{-14.4}14.4^r}{r!} > \frac{e^{-14.4}14.4^{r+1}}{(r+1)!}$$

$$1 > \frac{14.4}{r+1} \quad \Rightarrow \quad r + 1 > 14.4 \quad \Rightarrow \quad r > 13.4.$$

$r > 13.4$, *as required.* *([2])*

(b)(iii) The probabilities increase up to the integer after 13.4, then decrease.

The mode is 14. *([1])*

(c) The long-run records give much stronger evidence about the population than the small sample. Compare their mode with the model's mode.

The model gives mode 14, which is some way from the long-run modal value 10; therefore a Poisson distribution with $\lambda = 14.4$ is unlikely to be valid. *([1])*

(d) For a Poisson distribution, a mode of 10 requires $P(R = 10)$ to exceed both neighbouring probabilities.

$$P(R = 10) > P(R = 9) \Rightarrow \lambda > 10, \qquad P(R = 10) > P(R = 11) \Rightarrow \lambda < 11.$$

$10 < \lambda < 11$. *([2])*

www.mathswithmatt.co.uk

Question 6

Question 6

A teacher has 10 different mathematics books. Of these books, 5 are on Algebra, 3 are on Calculus and 2 are on Trigonometry.
The teacher chooses 5 of the books at random.

(a) Find the probability that 3 of the books are on Algebra. **([3])**

The teacher now arranges all 10 books in random order on a shelf.

(b) Find the probability that the Calculus books are next to each other and the Trigonometry books are next to each other. **([3])**

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

Find the probability that 2 of the Calculus books are next to each other but the third Calculus book is separated from the other 2 by at least 1 other book. **([4])**

Model solution

(a) Select exactly three of the five Algebra books and exactly two of the five non-Algebra books, then divide by all equally likely selections of five books.

$$P = \frac{\binom{5}{3}\binom{5}{2}}{\binom{10}{5}} = \frac{100}{252} = \frac{25}{63}.$$

$\frac{25}{63}$. *([3])*

(b) Treat the three Calculus books as one block and the two Trigonometry books as one block. Together with the five Algebra books, this gives seven objects, whose internal orders must also be counted.

$$P = \frac{7!\,3!\,2!}{10!} = \frac{1}{60}.$$

$\frac{1}{60}$. *([3])*

(c) Arrange the seven non-Calculus books first. This makes eight gaps for the adjacent pair of Calculus books. Once that pair is placed, exactly seven allowable gaps remain for the lone Calculus book: choosing an adjacent position would make all three Calculus books consecutive, which is not allowed.

$$\underbrace{\times \quad \times \quad \times \quad \times \quad \times \quad \times \quad \times}_{\text{seven non-Calculus books}}$$

There are $7!$ orders of the non-Calculus books, 8 choices for the pair's gap, 7 choices for the separated book's gap, and 3×2 ways to choose and order the two books in the pair.

$$\textit{favourable arrangements} = 7! \times 8 \times 7 \times (3 \times 2).$$

$$P = \frac{7! \times 8 \times 7 \times 6}{10!} = \frac{7}{15}.$$

$\frac{7}{15}$. *([4])*

Question 7

Question 7

Each of three students, X, Y and Z, was given an identical pack of 48 cards, of which 12 cards were red and 36 were blue. They were each told to carry out a different experiment, as follows:
Student X: Choose a card from the pack, at random, 20 times altogether, with replacement. Record how many times you obtain a red card.
Student Y: Choose a card from the pack, at random, 20 times altogether, without replacement. Record how many times you obtain a red card.
Student Z: Choose single cards from the pack at random, with replacement, until you obtain the first red card. Record how many cards you have chosen, including the first red card.

(a) Find the probability that student Z has to choose more than 8 cards in order to obtain the first red card. **([3])**

Each student carries out their experiment 30 times.
The frequencies of the results recorded by each student are shown in the following table, but not necessarily with the rows in the order X, Y, Z:

	Number recorded	0	1	2	3	4	5	6	7	8	≥ 9	Observed Mean	Observed Variance
Observed Frequencies	Student 1	0	0	1	3	7	8	6	4	1	0	5.03	1.97
	Student 2	0	8	5	4	2	3	3	2	1	2	4.03	11.57
	Student 3	0	1	2	5	4	6	5	3	4	0	4.97	3.70

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

Two other students make the following statements about the results. For each of the statements, explain whether you agree with the statement. Do not carry out any hypothesis tests, but in each case you should give two justifications for your answer.

(i) "The second row is a good match with the expected results for student Z." **([4])**

(ii) "The third row is definitely student X's results." **([3])**

Model solution

(a) Each draw has probability $12/48 = 0.25$ of red and is replaced, so the number of cards until the first red is geometric. More than eight cards means the first eight are blue.

$$Z \sim \text{Geo}(0.25), \qquad P(Z > 8) = (1 - 0.25)^8 = 0.75^8 = \frac{6561}{65536}.$$

$P(Z > 8) = 0.100$ *(approximately).* *([3])*

(b)(i) Student Z has $Z \sim \text{Geo}(0.25)$. Compare the expected mean and variance of this distribution with the second row.

$$E(Z) = \frac{1}{0.25} = 4, \qquad \text{Var}(Z) = \frac{1 - 0.25}{0.25^2} = 12.$$

The observed mean 4.03 is close to 4, and the observed variance 11.57 is close to 12. The generally decreasing frequencies are also consistent with a geometric distribution.

Agree: the second row is a good match with the expected results for student Z. *([4])*

(b)(ii) For student X, the number of red cards in 20 independent draws with replacement has a binomial distribution. Compare its mean and variance with the third row, but remember that

sampling variation prevents certainty.

$$X \sim \text{Bin}(20, 0.25), \qquad E(X) = 20(0.25) = 5, \qquad \text{Var}(X) = 20(0.25)(0.75) = 3.75.$$

The third row has observed mean 4.97, close to 5, and observed variance 3.70, close to 3.75. These are strong indications, but another student's random results could happen to be similar.

Disagree with "definitely": the third row is quite likely to be student X's results, but it cannot be known with certainty. *([3])*

www.mathswithmatt.co.uk