Maths with Matt

Physics and maths resources

OCR AS Further Mathematics

H235/Y532 - Statistics

Thursday 08 October 2020

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 supplied official OCR mark scheme. The original paper carries 60 marks. The aim is a clear **full-mark model response**, not merely a list of final answers.

Question wording is reproduced from the user-provided 2020 paper for study/reference. OCR retains copyright in the original examination material.

Contents

Question 1 2

Question 2 3

Question 3 4

Question 4 6

Question 5 7

Question 6 9

Question 7 11

www.mathswithmatt.co.uk

Question 1

Question 1

Five observations of bivariate data (x, y) are given in the table.

x	7	8	12	6	4
y	20	16	7	17	23

(a) Find the value of Pearson's product-moment correlation coefficient. **(2)**
(b) State what your answer to part (a) tells you about a scatter diagram representing the data. **(2)**
(c) A new variable a is defined by $a = 3x + 4$. Dee says "The value of Pearson's product-moment correlation coefficient between a and y will not be the same as the answer to part (a)." State with a reason whether you agree with Dee. **(1)**

Model solution

(a) Pearson's correlation coefficient is appropriate because both variables are numerical. We calculate it from the deviations from their means.

$$\bar{x} = \frac{37}{5} = 7.4, \qquad \bar{y} = \frac{83}{5} = 16.6,$$

$$S_{xx} = \sum(x - \bar{x})^2 = 35.2, \quad S_{yy} = \sum(y - \bar{y})^2 = 145.2, \quad S_{xy} = \sum(x - \bar{x})(y - \bar{y}) = -68.2.$$

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{-68.2}{\sqrt{35.2(145.2)}} = -0.954 \textit{ (3 s.f.)}.$$

$$r = -0.954.$$

(2)

(b) The sign gives the direction of the linear trend and the magnitude, close to 1, indicates that the points are close to a line.

The points lie close to a straight line with negative gradient.

(2)

(c) Adding a constant and multiplying by a positive constant are linear changes of scale. They do not alter Pearson's correlation coefficient.

Disagree: $a = 3x + 4$ is a positive linear transformation of x, so the correlation coefficient with y is unchanged.

(1)

Question 2

Question 2

Every time a spinner is spun, the probability that it shows the number 4 is 0.2, independently of all other spins.
(**a**) A pupil spins the spinner repeatedly until it shows the number 4. Find the mean of the number of spins required. **(2)**
(**b**) Calculate the probability that the number of spins required is between 3 and 10 inclusive. **(2)**
(**c**) Each pupil in a class of 30 spins the spinner until it shows the number 4. Out of the 30 pupils, the number of pupils who require at least 10 spins is denoted by X. Determine the variance of X. **(4)**

Model solution

(a) Each spin is an independent trial with a constant probability of first success, so the number of spins has a geometric distribution with success probability 0.2.

$$N \sim \text{Geo}(0.2), \qquad \text{E}(N) = \frac{1}{0.2} = 5.$$

5 ***spins.***

(2)

(b) Between 3 *and* 10 *inclusive means no* 4 *in the first* 2 *spins, but at least one* 4 *by the tenth spin. Using complementary probabilities gives the result directly.*

$$\Pr(3 \leq N \leq 10) = \Pr(N > 2) - \Pr(N > 10) = 0.8^2 - 0.8^{10} = 0.5326258\ldots$$

$$\Pr(3 \leq N \leq 10) = 0.533 \textbf{ (3 s.f.).}$$

(2)

(c) A pupil requires at least 10 *spins exactly when the first nine spins fail to show* 4. *The 30 pupils give independent identical Bernoulli trials, so a binomial model applies.*

$$\Pr(N \geq 10) = 0.8^9 = 0.134217728, \qquad X \sim \text{Bin}(30, 0.8^9).$$

$$\text{Var}(X) = np(1-p) = 30(0.8^9)(1-0.8^9) = 3.486\ldots$$

$$\text{Var}(X) = 3.49 \textbf{ (3 s.f.).}$$

(4)

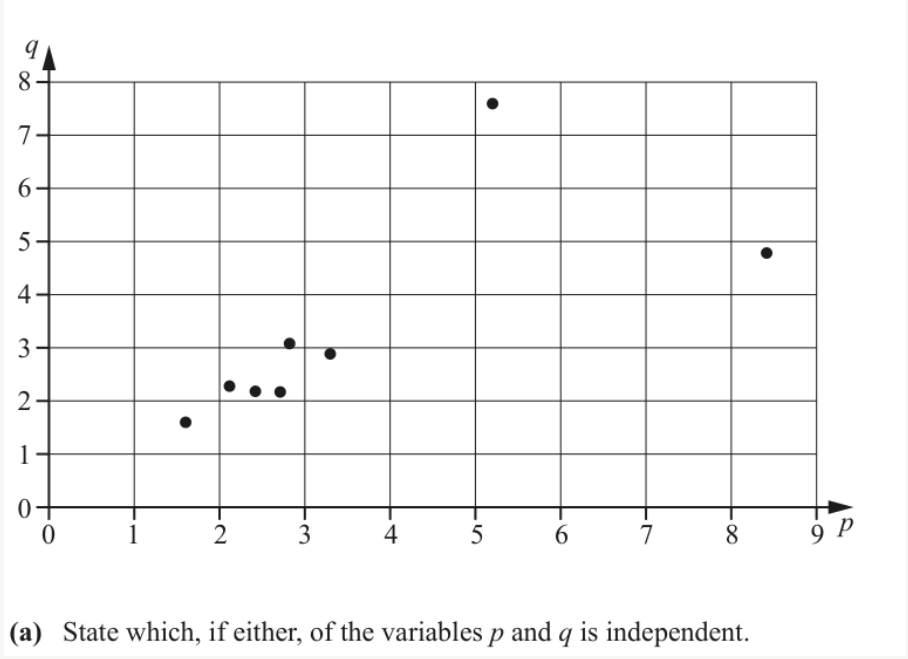
Question 3

Question 3

An investor obtains data about the profits of 8 randomly chosen investment accounts over two one-year periods. The profit in the first year for each account is p% and the profit in the second year is q%.

Account	A	B	C	D	E	F	G	H
p	1.6	2.1	2.4	2.7	2.8	3.3	5.2	8.4
q	1.6	2.3	2.2	2.2	3.1	2.9	7.6	4.8

$$n = 8, \quad \sum p = 28.5, \quad \sum q = 26.7, \quad \sum p^2 = 136.35, \quad \sum q^2 = 116.35, \quad \sum pq = 116.70.$$



(a) State which, if either, of the variables p and q is independent.

(**a**) State which, if either, of p and q is independent. **(1)**
(**b**) Calculate the regression line of q on p. **(3)**
(**c**)(**i**) Use it to estimate q when $p = 2.5$. **(1)**
(**c**)(**ii**) Give two reasons why this estimate could be reliable. **(2)**
(**d**) Comment on the reliability of predicting q when $p = 7.0$. **(2)**

Model solution

(a) The data are paired measurements, with neither variable specified as explanatory.

Neither p nor q is independent.

(1)

(b) For $q = a + bp$, calculate the least-squares gradient S_{pq}/S_{pp}.

$$\bar{p} = 3.5625, \qquad \bar{q} = 3.3375,$$

$$S_{pp} = 136.35 - \frac{28.5^2}{8} = 34.81875,$$

$$S_{pq} = 116.70 - \frac{28.5(26.7)}{8} = 21.58125.$$

$$b = \frac{21.58125}{34.81875} = 0.620\ldots, \qquad a = \bar{q} - b\bar{p} = 1.13\ldots$$

$$q = 1.13 + 0.620p.$$

(3)

(c)(i) Substitute the given value of p into the fitted line.

$$q = 1.13 + 0.620(2.5) = 2.68.$$

$$q = 2.68\%.$$

(1)

(c)(ii) This is interpolation and the plotted points show a reasonably close linear pattern.

2.5 is within the data range, and the points near it are close to the line.

(2)

(d) There is little data near $p = 7.0$, and the points in that region are scattered.

The prediction is not very reliable.

(2)

Question 4

Question 4

After a holiday, scores out of 10 for six aspects were obtained from 100 couples and 100 single travellers.

Aspect	Couples	Single travellers
Organisation	884	867
Travel	710	633
Food	692	675
Leader	898	898
Included visits	561	736
Optional visits	683	712

Fred wishes to test whether there is significant positive correlation between the scores.

(**a**) Explain why Pearson's product-moment correlation coefficient is probably not appropriate. **(1)**

(**b**) Carry out an appropriate test at the 1% level. **(7)**

(**c**) Explain what is meant by a non-parametric test. **(1)**

Model solution

(a) The distributional assumptions for Pearson correlation are not justified for these six aspect scores.

Use ranks rather than assuming bivariate normal data.

(1)

(b) Spearman's rank test is appropriate for association between two rankings.

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

Aspect	*Organisation*	*Travel*	*Food*	*Leader*	*Included*	*Optional*
Couples rank	5	4	3	6	1	2
Singles rank	5	1	2	6	4	3
d^2	0	9	1	0	9	1

$$\sum d^2 = 20, \qquad r_s = 1 - \frac{6(20)}{6(6^2-1)} = \frac{3}{7} = 0.42857\ldots$$

The 1% *one-tailed critical value for* $n = 6$ *is* 0.9429*, and* $0.42857 < 0.9429$*.*

Do not reject H_0*: insufficient evidence of positive association between the rankings.*

(7)

(c) Such a test makes no assumed parametric population distribution.

It does not depend on distributional assumptions.

(1)

Question 5

Question 5

At a cinema there are three Saturday sessions: early, middle and late. Table 1 gives the audience numbers on a randomly chosen Saturday.

Age group \ Session	Early	Middle	Late
< 25	24	20	40
25 to 60	4	2	10
> 60	28	22	10

The manager tests for association between age group and session attended.

(a) Show that cells must be combined. **(2)**

The second and third rows are combined. The incomplete expected-frequency and contribution tables are

Age group	Early	Middle	Late
< 25	29.4	23.1	
≥ 25	26.6	20.9	

Age group	Early	Middle	Late
< 25	0.9918	0.4160	
≥ 25	1.0962	0.4598	

(b) Complete both tables. **(3)**

(c) Carry out the test at the 5% level. **(5)**

(d) Use Table 3 to comment on the audience numbers. **(2)**

Model solution

(a) Expected frequencies in a chi-squared test must all be at least 5.

$$E_{\textit{middle},\,25\ \textit{to}\ 60} = \frac{16(44)}{160} = 4.4 < 5.$$

Cells must be combined.

(2)

(b) With combined row totals $84, 76$, *column totals* $56, 44, 60$, *and total* 160, *use* $E = (\textit{row total})(\textit{column total})/160$, *then* $(O-E)^2/E$.

Age group	Early	Middle	Late
< 25	29.4	23.1	31.5
≥ 25	26.6	20.9	28.5

Age group	Early	Middle	Late
< 25	0.9918	0.4160	2.2937
≥ 25	1.0962	0.4598	2.5351

Missing entries: $31.5, 28.5, 2.2937, 2.5351$.

(3)

(c) Test independence in the resulting 2×3 *table.*

$$H_0: \textit{ no association}; \qquad H_1: \textit{ association},$$

$$\chi^2 = 7.7926\ldots, \qquad \nu = (2-1)(3-1) = 2, \qquad \chi^2_{0.95,2} = 5.991.$$

Reject H_0, since $7.793 > 5.991$. There is significant evidence of association.

(5)

(d) The largest contributions are both in the late-session column.

There are more under-25s and fewer people aged at least 25 at the late session than expected under independence.

(2)

www.mathswithmatt.co.uk

Question 6

Question 6

A statistician investigates the number F of signal failures per week on a railway network.
(**a**) Explain what it means to assume failures occur randomly. **(1)**
(**b**) State two further assumptions for a Poisson model. **(2)**
For 50 randomly sampled weeks, the mean is 1.61 and the standard deviation is 1.28.
(**c**) Explain whether this suggests a Poisson model is likely to be suitable. **(2)**
Assume first that $F \sim \text{Po}(1.61)$.
(**d**) Write down an exact expression for $\Pr(F = 0)$. **(1)**
(**e**) Complete the probabilities, to three significant figures.

Value of F	0	1	≥ 2
Probability	0.200		

(2)
In a new model, $\Pr(F = 0)$ is still 0.200, but after one failure further failures are more likely. **(f)**
Explain the effect on $\Pr(F = 1)$. **(2)**

Model solution

(a) Randomness means no predictable timing pattern.

Failures occur to no fixed pattern.

(1)

(b) A Poisson model needs independent occurrences at a constant average rate.

Failures are independent and occur at a constant average rate.

(2)

(c) Poisson mean and variance should be equal.

$$1.28^2 = 1.6384 \approx 1.61.$$

Yes; the sample variance is close to the sample mean.

(2)

(d) Put zero into the Poisson probability formula.

$$\Pr(F = 0) = e^{-1.61}.$$

(1)

(e) Calculate $P(F = 1)$, then use the complement for $F \geq 2$.

$$\Pr(F = 1) = 1.61e^{-1.61} = 0.3218\ldots, \qquad \Pr(F \geq 2) = 1 - 0.200 - 0.3218\ldots = 0.4783\ldots$$

Value of F	0	1	≥ 2
Probability	0.200	0.322	0.478

(2)

(f) Probability moves from exactly one failure to two or more failures.

$\Pr(F = 1)$ *decreases.*

(2)

www.mathswithmatt.co.uk

Question 7

Question 7

A bag contains $2m$ yellow and m green counters. Three are chosen at random, without replacement. The probability of exactly two yellow counters is $\frac{28}{55}$. Determine m. **(7)**

Model solution

Choose two yellow counters and one green counter, then divide by the number of selections of three from all $3m$ counters.

$$\frac{\binom{2m}{2}\binom{m}{1}}{\binom{3m}{3}} = \frac{2m(2m-1)}{(3m-1)(3m-2)} = \frac{28}{55}.$$

Clear denominators and solve the resulting quadratic.

$$55(4m^2 - 2m) = 28(9m^2 - 9m + 2),$$

$$16m^2 - 71m + 28 = 0, \qquad (16m - 7)(m - 4) = 0.$$

$$m = \frac{7}{16} \quad \textbf{or} \quad m = 4.$$

m is a number of counters and must be an integer.

$$m = 4.$$

(7)