Maths with Matt

Physics and maths resources

Edexcel A Level Further Mathematics

9FM0/4B - Further Statistics 2

Tuesday 25 June 2019

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 Pearson Edexcel Summer 2019 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 2019 paper for study/reference. Pearson retains copyright in the original examination material.

Contents

Question 1 2

Question 2 3

Question 3 5

Question 4 6

Question 5 8

Question 6 9

Question 7 10

Question 8 11

www.mathswithmatt.co.uk

Question 1

Question 1

A machine is intended to fill yoghurt pots with mean weight 505 g. A random sample of eight weights g is

508, 510, 500, 500, 498, 503, 508, 505.

The population is normal with standard deviation 5.4 g.
a Find a 95% confidence interval for the population mean μ, to 2 decimal places. b Comment on whether the machine is working properly. c State the probability that a 95% confidence interval does not contain μ. d If the standard deviation were unknown, state and explain the changes needed to part a, assuming normality still holds.

Model solution

a *Because the population standard deviation is known and the population is normal, a normal (z) confidence interval for a mean is appropriate. First find the sample mean.*

$$\bar{x} = \frac{508 + 510 + 500 + 500 + 498 + 503 + 508 + 505}{8} = 504.$$

For 95% confidence, the central normal critical value is $z = 1.96$. Thus

$$\mu \in \bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}} = 504 \pm 1.96\frac{5.4}{\sqrt{8}} = 504 \pm 3.7418\ldots$$

The 95% confidence interval is $500.26,\ 507.74$g.

(4)

(b) Compare the target mean with the interval: values inside the interval are consistent with the sample at this confidence level.

Since 505g lies in $((500.26, 507.74))$, there is evidence that the machine is working pr

(1)

c *A 95% confidence procedure misses the true mean in the remaining 5% of repeated samples.*

0.05, or 5%.

(1)

d *With σ unknown, estimate it by the sample standard deviation s. Since this is a small sample from a normal population, the extra uncertainty means that a t, rather than normal, critical value is required.*

Use $\bar{x} \pm t_7\, s/\sqrt{8}$, replacing σ by s and 1.96 by the appropriate $t-$value.

(2)

Question 2

Question 2

Eight equal wheat plots receive fertiliser amount f g m^{-2}, and their yields w tonnes are recorded. The summary statistics are

$$\sum f = 28, \quad \sum w = 303, \quad \sum w^2 = 13\,447, \quad S_{ff} = 42, \quad S_{fw} = 269.5.$$

a Calculate the product moment correlation coefficient. b Interpret it. c Find the regression line of w on f, (w=a+bf). d Use it to estimate the decrease in yield for a decrease of 0.5 g m^{-2}. A residual plot against f has a clear approximately n-shaped pattern. e Comment on the suitability of the model. f Suggest a refinement.

Model solution

a *The product moment correlation coefficient uses the centred sums* S_{ff}, S_{ww}, S_{fw}. *We first obtain the missing centred sum for yield.*

$$S_{ww} = \sum w^2 - \frac{(\sum w)^2}{n} = 13447 - \frac{303^2}{8} = 1970.875.$$

$$r = \frac{S_{fw}}{\sqrt{S_{ff}S_{ww}}} = \frac{269.5}{\sqrt{42(1970.875)}} = 0.9367\ldots$$

$$(r \approx 0.937).$$

(2)

(b) The positive value close to 1 *indicates a strong positive linear association.*

As the amount of fertiliser increases, the wheat yield tends to increase.

(1)

c *For the regression of yield on fertiliser, the gradient predicts changes in* w *from* f, *so it is* S_{fw}/S_{ff}. *The line must then pass through* $(\bar{f}, \overline{w})$.

$$b = \frac{S_{fw}}{S_{ff}} = \frac{269.5}{42} = 6.4166\ldots, \qquad \bar{f} = \frac{28}{8} = 3.5, \quad \overline{w} = \frac{303}{8} = 37.875.$$

$$a = \overline{w} - b\bar{f} = 37.875 - (6.4166\ldots)(3.5) = 15.4166\ldots$$

$$w = 15.4 + 6.42f (more accurately, w = 15.4167 + 6.4167f).$$

(3)

d *A decrease of* 0.5 *in* f *changes the fitted yield by* $6.4166\ldots \times (-0.5)$.

$$\Delta w = -3.2083\ldots$$

Theestimateddecreaseis3.21tonnes.

(1)

e *For a suitable linear model, residuals should be randomly scattered around zero. The plotted n-shaped pattern is systematic, rather than random.*

The linear model is unlikely to be suitable.

(2)

f *The curved residual pattern suggests that the mean relationship itself is curved.*

Fit a curve, for example a quadratic regression, rather than a straight line.

(1)

www.mathswithmatt.co.uk

Question 3

Question 3

For random samples of potato yields x kg:

	n	$\sum x$	$\sum x^2$
A	25	194.7	1637.37
B	26	227.5	2031.19

a At the 10% level, test whether the population variances are the same, stating hypotheses. b State an assumption made for this test.

Model solution

a *The question compares two variances, so use an* F*-test. The alternative is two-tailed because it asks whether they are the same or not.*

$$H_0 : \sigma_A^2 = \sigma_B^2, \qquad H_1 : \sigma_A^2 \neq \sigma_B^2.$$

Calculate each unbiased sample variance from the supplied summaries.

$$s_A^2 = \frac{1}{24}\left(1637.37 - \frac{194.7^2}{25}\right) = 5.0436\ldots, \qquad s_B^2 = \frac{1}{25}\left(2031.19 - \frac{227.5^2}{26}\right) = 1.6226\ldots$$

Put the larger variance on top, so the test statistic is in the upper tail. Under H_0*, it has an* $F_{24,25}$ *distribution.*

$$F = \frac{s_A^2}{s_B^2} = 3.108\ldots, \qquad F_{24,25;0.05} = 1.96.$$

For a 10% two-tailed test, the quoted value is the 5% upper-tail critical value. Since $3.108 > 1.96$*, reject* H_0*.*

There is evidence at the 10% level that the variances of the two potato yields are diff

(7)

(b) The F*-test for a variance ratio requires the sampled variable in each population to be normally distributed.*

The yields of each variety are normally distributed.

(1)

Question 4

Question 4

The continuous random variable X has cdf

$$F(x) = \begin{cases} 0, & x \leqslant 0, \\ k\left(x^3 - \dfrac{3x^4}{8}\right), & 0 < x \leqslant 2, \\ 1, & x > 2. \end{cases}$$

where k is a constant. (a) Show that $k = \frac{1}{2}$. (b) Using calculus, find (i) $E(X)$ and (ii) the mode. (c) Describe the skewness, with a reason.

Model solution

(a) At the upper end of the continuous support, the cdf must equal 1. Substituting $x = 2$ therefore determines k.

$$1 = F2 = k\left(2^3 - \frac{32^4}{8}\right) = k8 - 6 = 2k.$$

$$(k = \tfrac{1}{2}).$$

(1)

(b)(i) For a continuous random variable, expectation is calculated from the pdf, so first differentiate the cdf on $0 \leqslant x \leqslant 2$.

$$fx = F'x = k\left(3x^2 - \frac{3}{2}x^3\right).$$

Now integrate $xf(x)$ over the support, using $k = \frac{1}{2}$.

$$EX = \int_0^2 xfx\,dx = \int_0^2 \left(\frac{3}{2}x^3 - \frac{3}{4}x^4\right) dx = \left[\frac{3x^4}{8} - \frac{3x^5}{20}\right]_0^2 = 6 - \frac{24}{5} = \frac{6}{5}.$$

$$(EX = \tfrac{6}{5} = 1.2).$$

(3)

b*(ii) The mode is where the pdf is greatest. A stationary value of the pdf is found by differentiating it and setting the result to zero.*

$$f'x = 3x - \frac{9}{4}x^2 = 0 \quad \Longrightarrow \quad \frac{9}{4}x\left(\frac{4}{3} - x\right) = 0.$$

$$x = 0 \quad \textit{or} \quad x = \frac{4}{3}.$$

At $x = 0$, $f(0) = 0$, an endpoint minimum; the interior stationary point gives the maximum density.

The mode is $\frac{4}{3}$.

(3)

c *The mode $\frac{4}{3}$ is greater than the mean $\frac{6}{5}$, so the longer tail is to the left.*

The distribution is negatively skewed.

(1)

www.mathswithmatt.co.uk

Question 5

Question 5

Eight students take papers I and II. Their percentage scores are

Student	A	B	C	D	E	F	G	H
I	70	70	84	80	64	65	65	90
II	64	76	72	74	68	64	58	76

At the 1% level, test whether there is evidence to support the belief that students are equally likely to achieve the same percentage score on the two papers. State hypotheses and show working.

Model solution

Each student supplies a matched pair, so a paired t*-test is needed: it tests the mean of the within-student differences rather than treating the two samples as independent. Let* $d =$ *score on I* $-$ *score on II.*

$$d: \quad 6,\ -6,\ 12,\ 6,\ -4,\ 1,\ 7,\ 14.$$

$$H_0 : \mu_d = 0, \qquad H_1 : \mu_d \neq 0.$$

The two-tailed alternative reflects the question asking whether the papers differ in level. From the differences,

$$\bar{d} = \frac{36}{8} = 4.5, \qquad s_d^2 = \frac{\sum (d - \bar{d})^2}{7} = 50.2857\ldots, \qquad s_d = 7.09\ldots$$

Under H_0*, the paired test statistic has a* t_7 *distribution.*

$$t = \frac{\bar{d} - 0}{s_d/\sqrt{8}} = \frac{4.5\sqrt{8}}{7.09\ldots} = 1.7948\ldots$$

At the 1% two-tailed level, the critical values are ± 3.499*. Since* $|1.795| < 3.499$*, do not reject* H_0*.*

Thereisinsufficientevidencethatthepapersareofdifferentdifficulty; Alexa'sb

(7)

Question 6

Question 6

Bolt diameters are normally distributed with mean 5 mm. Stan takes a random sample of 10 bolts. Ink obscures parts of his working; the visible entries include

$$4.5,\ 4.5,\ 5.5,\ 4.8,\ 4.9,\ 4.7,\ 5, \qquad \sum x = 48.4,$$

and a 99% confidence interval for the population variance of $(0.01712, 0.23280)$. At the 5% level, test Stan's belief that the mean diameter is less than 5 mm, stating hypotheses.

Model solution

The population is normal but the variance is unknown, so the eventual mean test is a one-sample t-test. The confidence interval for variance allows us to recover s^2, using the chi-squared interval formula with $n - 1 = 9$ degrees of freedom.

$$\chi^2_{0.995} = 1.735, \qquad \chi^2_{0.005} = 23.589.$$

Using either end of the interval gives the same sample variance; using the upper endpoint,

$$0.23280 = \frac{9s^2}{1.735} \quad \Longrightarrow \quad s^2 = \frac{0.232801.735}{9} = 0.04487\ldots$$

The sample mean comes from the visible total.

$$\bar{x} = \frac{48.4}{10} = 4.84.$$

$$H_0 : \mu = 5, \qquad H_1 : \mu < 5.$$

This is a lower-tailed test because Stan's belief is specifically "less than". Under H_0, the test statistic has a t_9 distribution. The 5% lower-tail critical value is -1.833.

$$t = \frac{\bar{x} - 5}{s/\sqrt{10}} = \frac{4.84 - 5}{\sqrt{0.04487\ldots}/\sqrt{10}} = -2.39\ldots$$

Since $-2.39 < -1.833$, reject H_0.

Stan's belief is supported : there is evidence at the 5% level that the mean bolt diameter is less than 5 mm.

(9)

Question 7

Question 7

Wooden-toy weight $W \sim N(2.5, 0.7^2)$ kg and plastic-toy weight $X \sim N(1.27, 0.4^2)$ kg; W and X are independent.
a Find $P(W > 2X)$. b A container holds n wooden and $2n$ plastic toys, has maximum weight 252 kg, and has probability 0.2119 of being overweight. Find n.

Model solution

a *The event compares two independent normal variables. Rearranging it into one linear combination gives a single normal variable that can be standardised. Let* $T = W - 2X$; *then* $W > 2X$ *is* $T > 0$.

$$ET = EW - 2EX = 2.5 - 21.27 = -0.04,$$

$$\operatorname{Var} T = \operatorname{Var} W + 4 \operatorname{Var} X = 0.7^2 + 40.4^2 = 1.13.$$

Independence makes the covariance term zero, so $T \sim N(-0.04, 1.13)$.

$$PW > 2X = PT > 0 = P\left(Z > \frac{0 - -0.04}{\sqrt{1.13}}\right) = P(Z > 0.0376\ldots) = 0.484\ldots$$

$$(PW > 2X \approx 0.485).$$

(6)

(b) The total is a sum of independent toy weights, hence normal. Write B *for total weight. Means and variances add; there are* n *wooden and (2n) plastic toys.*

$$EB = n2.5 + 2n1.27 = 5.04n, \qquad \operatorname{Var} B = n0.7^2 + 2n0.4^2 = 0.81n.$$

Thus $B \sim N(5.04n, 0.81n)$. *The stated overweight probability is an upper-tail probability of* 0.2119, *whose standard normal boundary is* $z = 0.8$. *Therefore*

$$\frac{252 - 5.04n}{\sqrt{0.81n}} = 0.8.$$

Set $y = \sqrt{n}$, *so that* $n = y^2$, *to turn this into a quadratic.*

$$252 - 5.04y^2 = 0.8(0.9y) = 0.72y, \qquad 5.04y^2 + 0.72y - 252 = 0.$$

$$y = 7 \quad \textit{(the positive root)}, \qquad n = y^2 = 49.$$

$$n = 49 \text{woodentoysand} 98 \text{plastictoys}.$$

(8)

Question 8

Question 8

Nine athletes competed in a 100 m sprint and a long jump. Spearman's rank correlation coefficient was $r_s = 0.85$.
a At the 5% level, test whether a higher sprint position is associated with a higher long-jump position. b The sprint positions for (A,B,C,D,E,F,G,H,I) were respectively

$$4, 6, 7, 9, 2, 8, 3, 1, 5.$$

The agreed long-jump positions were $A = 5, E = 4, F = 9, G = 3, H = 1, I = 2$; the positions of B, C, D were disputed. Given no ties, find them and justify the answer. c Explain how r_s changes if H is disqualified from both events, without recalculating it.

Model solution

a *Spearman's coefficient tests association between ranked variables. The proposed direction is positive, so use a one-tailed test.*

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

For $n = 9$ *at the 5% level, the upper critical value is* 0.6. *Since* $0.85 > 0.6$, *reject* H_0.

There is evidence that a higher 100 m position is associated with a higher long-jump position.

(4)

(b) With no tied ranks, use Spearman's formula to find the total squared rank difference required by the given coefficient.

$$r_s = 1 - \frac{6\sum d^2}{nn^2 - 1} = 1 - \frac{6\sum d^2}{980}.$$

$$0.85 = 1 - \frac{6\sum d^2}{720} \quad \Longrightarrow \quad \sum d^2 = 18.$$

The six agreed athletes contribute

$$4 - 5^2 + 2 - 4^2 + 8 - 9^2 + 3 - 3^2 + 1 - 1^2 + 5 - 2^2 = 1 + 4 + 1 + 0 + 0 + 9 = 15.$$

Hence (B,C,D) must contribute $18 - 15 = 3$. *The unused long-jump positions are (6,7,8), while their sprint positions are (6,7,9). To make three squared differences totalling* 3, *each must be* 1. *Athlete* B *must therefore be at position* 5 *or* 7, *but* 5 *is already used, so* $B = 7$. *The remaining adjacent assignments are* $C = 6, D = 8$.

$$\text{Thecorrectlong} - \text{jumppositionsare} B = 7, C = 6, D = 8.$$

(5)

c *Athlete* H *is first in both events, so their rank difference is* 0. *Removing them leaves* $\sum d^2$ *unchanged, but reduces* n, *which makes the subtracted fraction in Spearman's formula larger.*

Spearman's rank correlation coefficient decreases.

(2)