

Pearson Edexcel Level 3 Advanced GCE in Further Mathematics (9FM0)



Sample Assessment Materials Model Answers – Further Statistics 1&2

First teaching from September 2017
First certification from June 2019

Sample Assessment Materials Model Answers – Further Statistics 1&2

Contents

Introduction	5
Content of Further Statistics.....	5
Further Statistics 1	7
Question 1.....	7
Question 2.....	8
Question 3.....	10
Question 4.....	13
Question 5.....	14
Question 6.....	16
Question 7.....	19
Further Statistics 2	22
Question 1.....	22
Question 2.....	25
Question 3.....	27
Question 4.....	29
Question 5.....	31
Question 6.....	33
Question 7.....	36

Introduction

This booklet has been produced to support mathematics teachers delivering the new Pearson Edexcel Level 3 Advanced GCE in Mathematics (9FM0) specification for first teaching from September 2017.

This booklet looks at Sample Assessment Materials for A level Further Mathematics qualification, specifically at further statistics questions, and is intended to offer model solutions with different methods explored.

Content of Further Statistics

Further Statistics 1	
Discrete probability distribution	Mean and variance of discrete probability distributions. Extension of expected value function to include $E(g(X))$.
Poisson & binomial distributions	The Poisson distribution. The additive property of Poisson distributions. The mean and variance of the binomial and the Poisson distributions. Use of Poisson distribution as an approximation to the binomial distribution. Extend ideas of hypothesis tests to test for the mean of Poisson distribution.
Geometric and negative binomial distributions	Geometric and negative binomial distributions. Mean and variance of a geometric distribution with parameter p . Mean and variance of negative binomial distribution with $P(X = x) = \binom{x-1}{r-1} p^r (1-p)^{(x-r)}$
Hypothesis Testing	Extend ideas of hypothesis tests to test for the mean of a Poisson distribution. Extend hypothesis testing to test for the parameter p of a geometric distribution.
Central Limit Theorem	Applications of the Central Limit Theorem to other distributions.
Chi Squared Tests	Goodness of fit tests and Contingency tables. The null and alternative hypotheses. The use of $\sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$ as approximate χ^2 statistic. Degrees of freedom.
Probability generating functions	Definitions, derivations and applications. Use of the probability generating function for the negative binomial, geometric, binomial and Poisson distributions. Use to find the mean and variance. Probability generating function of the sum of independent random variables.
Quality of tests	Type I and Type II errors. Size and Power of Test. The power function.

Further Statistics 2	
Linear Regression	<p>Least squares linear regression. The concept of residuals and minimising the sum of squares of residuals.</p> <p>Residuals.</p> <p>The residual sum of squares (RSS).</p>
Continuous probability distributions	<p>The concept of a continuous random variable.</p> <p>The probability density function and the cumulative distribution function for a continuous random variable.</p> <p>Relationship between probability density and cumulative distribution functions.</p> <p>Mean and variance of continuous random variables.</p> <p>Extension of expected value function to include $E(g(X))$.</p> <p>Mode, median and percentiles of continuous random variables.</p> <p>Idea of skewness.</p> <p>The continuous uniform (rectangular) distribution.</p>
Correlation	<p>Use of formulae to calculate the product moment correlation.</p> <p>Knowledge of conditions for the use of the product moment correlation.</p> <p>A knowledge of effects of coding.</p> <p>Spearman's rank correlation coefficient, its use and interpretation.</p> <p>Testing the hypothesis that a correlation is zero using either Spearman's rank correlation or the product moment correlation coefficient.</p>
Combinations of random variables	<p>Distribution of linear combinations of independent Normal random variables.</p>
Estimation, confidence intervals and tests using a normal distribution	<p>Concepts of standard error, estimator, bias.</p> <p>Quality of estimators.</p> <p>Concept of a confidence interval and its interpretation.</p> <p>Confidence limits for a Normal mean, with variance known.</p> <p>Hypothesis test for the difference between the means of two Normal distributions with variances known.</p> <p>Use of large sample results to extend to the case in which the population variances are unknown.</p>
Other Hypothesis Tests and confidence intervals	<p>Hypothesis test and confidence interval for the variance of a Normal distribution.</p> <p>Hypothesis test that two independent random samples are from Normal populations with equal variances.</p>
Confidence intervals and tests using the t – distribution	<p>Hypothesis test and confidence interval for the mean of a Normal distribution with unknown variance.</p> <p>Paired t-test.</p> <p>Hypothesis test and confidence interval for the difference between two means from independent Normal distributions when the variances are equal but unknown.</p> <p>Use of the pooled estimate of variance.</p>

Further Statistics 1

Question 1

Bacteria are randomly distributed in a river at a rate of 5 per litre of water. A new factory opens and a scientist claims it is polluting the river with bacteria. He takes a sample of 0.5 litres of water from the river near the factory and finds that it contains 7 bacteria. Stating your hypotheses clearly test, at the 5% level of significance, whether there is evidence that the level of pollution has increased.

(5)

Hypotheses:

Either:

$$H_0 : \lambda = 5 \quad H_1 : \lambda > 5$$

Or:

$$H_0 : \lambda = 2.5 \quad H_1 : \lambda > 2.5$$

B1

Model this situation as $X \sim \text{Po}(2.5)$

B1

Either:

$$\begin{aligned} P(X \geq 7) &= 1 - P(X \leq 6) \\ &= 1 - 0.9858 \\ &= 0.0142 \end{aligned}$$

Or:

$$\begin{aligned} P(X \geq 5) &= 1 - P(X \leq 4) \\ &= 1 - 0.8912 = 0.1088 \\ P(X \geq 6) &= 1 - P(X \leq 5) \\ &= 1 - 0.9580 = 0.042 \end{aligned}$$

M1

Critical Region is $X \geq 6$

A1

$$0.0142 < 0.05$$

7 is in the critical region

Reject H_0 .

There is evidence at the 5% significance level that the level of pollution has increased.

Or: There is evidence to support the scientists claim is justified.

A1

Question 2

A call centre routes incoming telephone calls to agents who have specialist knowledge to deal with the call. The probability of a caller, chosen at random, being connected to the wrong agent is p .

The probability of at least 1 call in 5 consecutive calls being connected to the wrong agent is 0.049.

The call centre receives 1000 calls each day.

(a) Find the mean and variance of the number of wrongly connected calls a day.

(7)

$$P(X \geq 1) = 1 - P(X = 0) = 0.049 \quad \text{B1}$$

$$P(X = 0) = 0.951 \quad \text{B1}$$

$$x^5 = 0.951$$

$$x = 0.9900020716... = 0.99 \text{ (2dp)} \quad \text{M1}$$

$$p = 1 - 0.99 = 0.01 \quad \text{A1}$$

$$\text{Model this situation as } X \sim B(1000, 0.01) \quad \text{M1}$$

Formula book:

Binomial distribution:

$$\text{Mean} = np \quad \text{Variance} = np(1 - p)$$

$$\text{Mean} = 1000 \times 0.01 = 10 \quad \text{A1}$$

$$\text{Variance} = 1000 \times 0.01 \times 0.99 = 9.9 \quad \text{A1}$$

(b) Use a Poisson approximation to find, to 3 decimal places, the probability that more than 6 calls each day are connected to the wrong agent.

(2)

Use the model $X \sim \text{Po}(10)$

$$P(X > 6) = 1 - P(X \leq 6) \quad \text{M1}$$

$$= 1 - 0.1301$$

$$= 0.870 \text{ (3dp)} \quad \text{A1}$$

(c) Explain why the approximation used in part (b) is valid.

(2)

The approximation is valid as: the number of calls is large,
The probability of connecting to the wrong agent is small.

B1

B1

The probability that more than 6 calls each day are connected to the wrong agent using the binomial distribution is 0.8711 to 4 decimal places.

(d) Comment on the accuracy of your answer in part (b).

(1)

The answer is accurate to 2 decimal places.

B1

Question 3

Bags of £1 coins are paid into a bank. Each bag contains 20 coins.

The bank manager believes that 5% of the £1 coins paid into the bank are fakes. He decides to use the distribution $X \sim B(20, 0.05)$ to model the random variable X , the number of fake £1 coins in each bag.

The bank manager checks a random sample of 150 bags of £1 coins and records the number of fake coins found in each bag. His results are summarised in Table 1. He then calculates some of the expected frequencies, correct to 1 decimal place.

Number of fake coins in each bag	0	1	2	3	4 or more
Observed frequency	43	62	26	13	6
Expected frequency	53.8	56.6		8.9	

Table 1

- (a) Carry out a hypothesis test, at the 5% significance level, to see if the data supports the bank manager's statistical model. State your hypotheses clearly.

(10)

$$\begin{aligned}
 \text{Expected value for '2'} &= 150 \times P(X = 2) \\
 &= 150 \times 0.1886768013\dots && \text{M1} \\
 &= 28.30152\dots && \text{A1} \\
 &= 28.3 \text{ (1dp)}
 \end{aligned}$$

$$\begin{aligned}
 \text{Expected value for 4 or more} &= 150 - (53.8 + 56.6 + 28.3 + 8.9) \\
 &= 2.4 && \text{A1}
 \end{aligned}$$

Hypotheses:

$$\begin{aligned}
 H_0: X \sim B(20, 0.05) \text{ is a suitable model} \\
 H_1: X \sim B(20, 0.05) \text{ is not a suitable model} &&& \text{B1}
 \end{aligned}$$

The last 2 groups need to be combined since $2.4 < 5$

$$\begin{array}{ll}
 \text{No of fake coins} & \geq 3 \\
 \text{Observed freq} & 19 \\
 \text{Expected freq} & 11.3 && \text{M1}
 \end{array}$$

$$\text{Degrees of freedom } \nu = 4 - 1 = 3 \quad \text{B1}$$

$$\text{Critical value from tables, } \chi^2_{(5\%)} = 7.815 \quad \text{B1}$$

Formula book:

Goodness of fit: $\sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_v$

$$\frac{\sum (O_i - E_i)^2}{E_i} = \frac{(43 - 53.8)^2}{53.8} + \frac{(62 - 56.6)^2}{56.6} + \frac{(26 - 28.3)^2}{28.3} + \frac{(19 - 11.3)^2}{11.3} \quad \text{M1}$$

$$= 2.1680 + 0.5152 + 0.1869 + 5.2469$$

$$= 8.117 \text{ (3dp)} \quad \text{A1}$$

This is in the critical region, there is sufficient evidence to reject H_0 , accept H_1 .

There is significant evidence at 5% level to reject the manager's model. A1

The assistant manager thinks that a binomial distribution is a good model but suggests that the proportion of fake coins is higher than 5%. She calculates the actual proportion of fake coins in the sample and uses this value to carry out a new hypothesis test on the data. Her expected frequencies are shown in Table 2.

Number of fake coins in each bag	0	1	2	3	4 or more
Observed frequency	43	62	26	13	6
Expected frequency	44.5	55.7	33.2	12.5	4.1

Table 2

(b) Explain why there are 2 degrees of freedom in this case.

(2)

$$v = 4 - 1 - 1 = 2$$

There are 4 classes due to combining the last 2 columns ($4.1 < 5$). B1

There are 2 restrictions, totals must equal and the mean has been calculated. B1

(c) Given that she obtains a χ^2 test statistic of 2.67, test the assistant manager's hypothesis that the binomial distribution is a good model for the number of fake coins in each bag. Use a 5% level of significance and state your hypotheses clearly.

(2)

Hypotheses:

H_0 : Binomial distribution is a good model

H_1 : Binomial distribution is not a good model

B1

Critical value from tables, $\chi^2_{(5\%)} = 5.991$

The test statistic is not in critical region, insufficient evidence to reject H_0 .

There is evidence that the Binomial distribution is a good model.

B1

Question 4

A random sample of 100 observations is taken from a Poisson distribution with mean 2.3.
 Estimate the probability that the mean of the sample is greater than 2.5.

(4)

$$X \sim \text{Po}(2.3), n = 100, \mu = 2.3, \sigma^2 = 2.3$$

Using the Central limit Theorem (CLT) with mean 2.3: M1

$$X \approx N\left(2.3, \frac{2.3}{100}\right) \quad \text{A1}$$

$$\begin{aligned} P(X > 2.5) &= P\left(Z > \frac{2.5 - 2.3}{\sqrt{0.023}}\right) && \text{M1} \\ &= P(Z > 1.31876) \\ &= 1 - 0.906375\dots \\ &= 0.093625\dots \\ &= 0.0936 \text{ (4dp)} && \text{A1} \end{aligned}$$

Note:

Candidates are expected to have a calculator that can access probabilities from statistical distributions including the normal distribution. It may not be necessary to standardise before obtaining the probability.

Question 5

The probability of Richard winning a prize in a game at the fair is 0.15.

Richard plays a number of games.

(a) Find the probability of Richard winning his second prize on his 8th game,

(2)

Formula book:

$$\text{Binomial: } P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

$$\text{Negative Binomial: } P(X = x) = \binom{x-1}{r-1} p^r (1 - p)^{x-r}$$

Either:

Negative binomial with
 $x = 8, r = 2$

$$\binom{8-1}{2-1} \times 0.15^2 \times 0.85^{8-2}$$

$$= 7 \times 0.15^2 \times 0.85^6$$

$$= 0.0594$$

Or:

$X \sim B(7, 0.15)$ with an extra
success in 8th trial

$$\binom{7}{1} \times 0.15^1 \times 0.85^{7-1} \times 0.15$$

$$= 7 \times 0.15^2 \times 0.85^6$$

$$= 0.0594$$

M1

A1

(b) State two assumptions that have to be made, for the model used in part (a) to be valid.

(2)

The model is only valid if:

The games (trials) are independent

B1

The probability of winning a prize, 0.15, is constant for each game

B1

Mary plays the same game, but has a different probability of winning a prize. She plays until she has won r prizes. The random variable G represents the total number of games Mary plays.

(c) Given that the mean and standard deviation of G are 18 and 6 respectively, determine whether Richard or Mary has the greater probability of winning a prize in a game.

(4)

Formula book:

Negative Binomial: Mean = $\frac{r}{p}$ Variance = $\frac{r(1-p)}{p^2}$

$$\frac{r}{p} = 18$$

M1

$$\frac{r(1-p)}{p^2} = 6^2$$

A1

$$r = 18p$$

$$r = \frac{36p^2}{1-p}$$

$$\frac{36p^2}{1-p} = 18p$$

$$2p = 1 - p$$

M1

$$p = \frac{1}{3}$$

$$\frac{1}{3} > 0.15 \text{ so Mary has the greater chance of winning a prize.}$$

A1

Question 6

The probability generating function of the discrete random variable X is given by

$$G_X(t) = k(3 + t + 2t^2)^2$$

(a) Show that $k = \frac{1}{36}$

(2)

Let $t = 1$ then $G_X(1) = 1$

M1

$$k(3 + 1 + 2)^2 = 1$$

$$36k = 1$$

$$k = \frac{1}{36}$$

A1

(b) Find $P(X = 3)$

(2)

$P(X = 3) =$ coefficient of t^3 term in expansion of $\frac{1}{36}(3 + t + 2t^2)(3 + t + 2t^2)$

$$= \frac{1}{36}(\dots + 2t^3 + 2t^3 + \dots)$$

M1

$$= \frac{4}{36}$$

$$= \frac{1}{9}$$

A1

Note:

Complete expansion would be

$$\frac{1}{36}(9 + 6t + 13t^2 + 4t^3 + 4t^4)$$

(c) Show that $\text{Var}(X) = \frac{29}{18}$

(8)

Formula book:

Probability generating function (PGF):

$$E(X) = G'_X(1) \quad \text{Var}(X) = G''_X(1) + G'_X(1) - [G'_X(1)]^2$$

Differentiate using chain rule:

$$\begin{aligned} G'_X(t) &= \frac{1}{36} \times 2(3+t+2t^2)^1 \times (1+4t) && \text{M1} \\ &= \frac{1}{18} (3+t+2t^2)(1+4t) \end{aligned}$$

$$\begin{aligned} E(X) = G'_X(1) &= \frac{1}{18} (3+1+2)(1+4) && \text{M1} \\ &= \frac{5}{3} && \text{A1} \end{aligned}$$

2nd derivative using product rule:

$$G''_X(t) = \frac{1}{18} [(3+t+2t^2) \times 4 + (1+4t)(1+4t)] \quad \text{M1 A1}$$

$$\begin{aligned} G''_X(1) &= \frac{1}{18} [6 \times 4 + 5^2] && \text{M1} \\ &= \frac{49}{18} \end{aligned}$$

$$\begin{aligned} \text{Var}(X) &= G''_X(1) + G'_X(1) - [G'_X(1)]^2 \\ &= \frac{49}{18} + \frac{5}{3} - \frac{25}{9} && \text{M1} \\ &= \frac{29}{18} && \text{A1} \end{aligned}$$

(d) Find the probability generating function of $2X + 1$

(2)

Original PGF has X values 0, 1, 2, 3, 4 with probabilities found from coefficients of t^0, t^1, t^2, t^3, t^4 .

So PGF for $2X + 1$ has X values 1, 3, 5, 7, 9 with probabilities found from coefficients of t^1, t^3, t^5, t^7, t^9 .

So the original PGF needs to be amended by substituting t^2 for t then multiplying by t . M1

$$G_{2X+1}(t) = \frac{1}{36}(3 + t^2 + 2(t^2)^2)^2 \times t$$

$$G_{2X+1}(t) = \frac{1}{36}t(3 + t^2 + 2t^4)^2 \quad \text{A1}$$

Or: $G_{2X+1}(t) = \frac{1}{36}(9t + 6t^3 + 13t^5 + 4t^7 + 4t^9)$

Question 7

Sam and Tessa are testing a spinner to see if the probability, p , of it landing on red is less than $\frac{1}{5}$.

They both use a 10% significance level.

Sam decides to spin the spinner 20 times and record the number of times it lands on red.

(a) Find the critical region for Sam's test.

(2)

Need to model this situation as $X \sim B(20, 0.2)$ and find c such that $P(X \leq c) < 0.10$

$$P(X \leq 1) = 0.0692$$

$$P(X \leq 2) = 0.2061$$

M1

So Critical Region is $X \leq 1$

A1

(b) Write down the size of Sam's test.

(1)

$$\text{Size} = 0.0692$$

B1

Tessa decides to spin the spinner until it lands on red and she records the number of spins.

(c) Find the critical region for Tessa's test.

(6)

Let Y = no of spins until red obtained

so this is modelled by the geometric distribution $Y \sim \text{Geo}(0.2)$

M1

Formula book:

Geometric distribution:

$$\text{mean} = \frac{1}{p}$$

So if $p < 0.2$ then the mean is larger so need to find d such that $P(X \leq d) < 0.10$

M1

$$P(Y \geq d) = (0.8)^{d-1}$$

M1

$$(0.8)^{d-1} < 0.10$$

$$(d - 1) \log 0.8 < \log 0.10$$

$$d - 1 > \frac{\log 0.10}{\log 0.8} \quad \text{M1}$$

$$d > 11.3 \dots \quad \text{A1}$$

So Critical Region is $Y \geq 12$ A1

(d) Find the size of Tessa's test. (1)

$$\text{Size} = 0.8^{11} = 0.0859 \quad \text{B1}$$

(e) (i) Show that the power function for Sam's test is given by

$$(1 - p)^{19} (1 + 19p)$$

(ii) Find the power function for Tessa's test. (4)

(i)
Power function = P(reject H_0 when it is false)

$$= P[X \leq 1 \mid X \sim B(20, p)] \quad \text{M1}$$

$$= P(X = 0) + P(X = 1)$$

$$= (1 - p)^{20} + 20(1 - p)^{19} p \quad \text{M1}$$

$$= (1 - p)^{19} [(1 - p) + 20p]$$

$$= (1 - p)^{19} (1 + 19p) \quad \text{A1}$$

(ii)
Power function = $(1 - p)^{11}$ B1

(f) With reference to parts (b), (d) and (e), state, giving your reasons, whether you would recommend Sam's test or Tessa's test when $p = 0.15$

(4)

$$0.0692 < 0.0859$$

Sam's test has a smaller P(Type 1 error) so is better.

B1

When $p = 0.15$,

$$\text{Power of Sam's test} = (1 - 0.15)^{19}(1 + 19 \times 0.15) = 0.176 \text{ (3dp)}$$

B1

$$\text{Power of Tessa's test} = (1 - 0.15)^{11} = 0.167 \text{ (3dp)}$$

B1

So for $p = 0.15$ Sam's test is recommended

B1

Further Statistics 2

Question 1

The three independent random variables A , B and C each have a continuous uniform distribution over the interval $[0, 5]$.

(a) Find the probability that A , B and C are all greater than 3.

(3)

$$P(A > 3) = P(B > 3) = P(C > 3) = \frac{2}{5}$$

B1

$$P(A, B, C \text{ all } > 3) = \left(\frac{2}{5}\right)^3$$

$$= \frac{8}{125}$$

M1

A1

The random variable Y represents the maximum value of A , B and C .

The cumulative distribution function of Y is

$$F(y) = \begin{cases} 0 & y < 0 \\ \frac{y^3}{125} & 0 \leq y \leq 5 \\ 1 & y > 5 \end{cases}$$

(b) Using algebraic integration, show that $\text{Var}(Y) = 0.9375$

(4)

Differentiate $F(y)$:

$$f(y) = \frac{3y^2}{125}$$

M1

$$E(Y) = \int_0^5 y \times \frac{3y^2}{125} dy = \int_0^5 \frac{3y^3}{125} dy$$

$$= \left[\frac{3y^4}{500} \right]_0^5$$

M1

$$= 3.75$$

$$E(Y^2) = \int_0^5 y^2 \times \frac{3y^2}{125} dy = \int_0^5 \frac{3y^4}{125} dy$$

$$= \left[\frac{3y^5}{625} \right]_0^5$$

$$= 15$$

$$\text{Var}(Y) = E(Y^2) - [E(Y)]^2$$

$$= 15 - 3.75^2 \quad \text{M1}$$

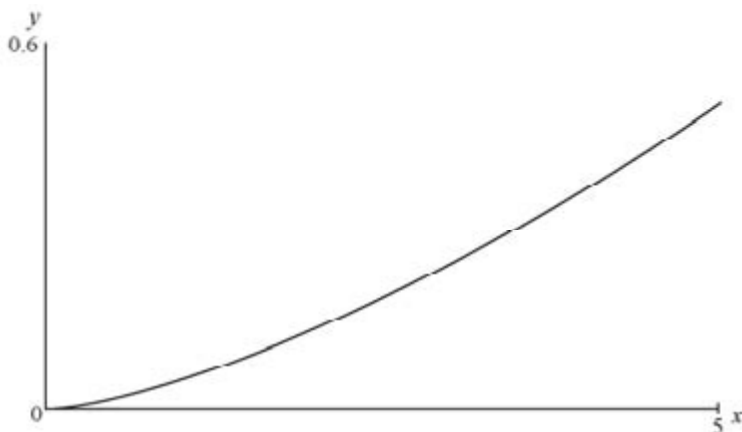
$$= 0.9375 \quad \text{A1}$$

(c) Find the mode of Y , giving a reason for your answer. (2)

Mode = 5 B1

Either:

A sketch of the quadratic function $f(y)$ shows that the maximum is at $(5, 0.6)$



Or:

$$\frac{d f(y)}{d y} = \frac{6y}{125} > 0 \text{ for } 0 < y < 5$$

So $f(y)$ is an increasing function for $0 < y < 5$, so the maximum is at 5. B1

(d) Describe the skewness of the distribution of Y . Give a reason for your answer. (1)

Either:

mode $>$ mean, $5 > 3.75$
therefore it has negative skew

Or:

The shape of the above sketch
shows it has negative skew

B1

(e) Find the value of k such that $P(k < Y < 2k) = 0.189$ (3)

$$P(k < Y < 2k) = \int_k^{2k} \frac{3y^2}{125} dy = 0.189$$

$$\left[\frac{y^3}{125} \right]_k^{2k} = 0.189$$

$$\frac{(2k)^3}{125} - \frac{k^3}{125} = 0.189$$

M1

$$\frac{7k^3}{125} = 0.189$$

A1

$$k^3 = 3.375$$

$$k = 1.5$$

A1

Question 2

A researcher claims that, at a river bend, the water gradually gets deeper as the distance from the inner bank increases. He measures the distance from the inner bank, b cm, and the depth of a river, s cm, at 7 positions. The results are shown in the table below.

Position	A	B	C	D	E	F	G
Distance from inner bank b cm	100	200	300	400	500	600	700
Depth s cm	60	75	85	76	110	120	104

The Spearman's rank correlation coefficient between b and s is $\frac{6}{7}$

- (a) Stating your hypotheses clearly, test whether or not the data provides support for the researcher's claim. Use a 1% level of significance.

(4)

Hypotheses:

$$H_0 : \rho = 0, \quad H_1 : \rho > 0$$

B1

From tables, critical value at 1% level is 0.8929

B1

$$r_s = \frac{6}{7} = 0.857... < 0.8929$$

so no significant evidence to reject H_0

M1

(Need a conclusion in context), e.g.

The researcher's claim is not correct (at 1% level)

or insufficient evidence for researcher's claim

or there is insufficient evidence that water gets deeper further from inner bank

or no (positive) correlation between depth of water and distance from inner bank

A1

- (b) Without re-calculating the correlation coefficient, explain how the Spearman's rank correlation coefficient would change if
- (i) the depth for G is 109 instead of 104
 - (ii) an extra value H with distance from the inner bank of 800 cm and depth 130 cm is included.

(3)

(i)
The ranks will remain the same therefore there will be no change to the Spearman's rank correlation coefficient

B1

(ii)
Spearman's rank correlation coefficient will increase
since e.g.

B1

The ranks are the same for both distance and depth therefore $d = 0$ however, n has increased.

Or:

the new position follows the pattern that large b is associated with large s and so r_s will increase.

B1

The researcher decided to collect extra data and found that there were now many tied ranks.

- (c) Describe how you would find the correlation with many tied ranks.

(2)

The mean of the values for the tied ranks is given to each value
so the PMCC (product moment correlation coefficient) must be used

B1

B1

Question 3

A nutritionist studied the levels of cholesterol, X mg/cm³, of male students at a large college. She assumed that X was distributed $N(\mu, \sigma^2)$ and examined a random sample of 25 male students. Using this sample she obtained unbiased estimates of μ and σ^2 as $\hat{\mu}$ and $\hat{\sigma}^2$

A 95% confidence interval for μ was found to be (1.128, 2.232)

(a) Show that $\hat{\sigma}^2 = 1.79$ (correct to 3 significant figures)

(4)

The student's t -distribution must be used as a model here.

Degrees of freedom, $\nu = 25 - 1 = 24$.

95% confidence interval uses the 0.025 column in the tables.

So t value used is 2.064

B1

$$\hat{\mu} = \frac{1}{2}(2.232 + 1.128) = 1.68$$

Using $\hat{\mu} \pm \frac{\hat{\sigma}}{\sqrt{n}} \times (t \text{ value})$:

Either:

$$1.68 + \frac{\hat{\sigma}}{\sqrt{25}} \times 2.064 = 2.232$$

Or:

$$1.68 - \frac{\hat{\sigma}}{\sqrt{25}} \times 2.064 = 1.128$$

Alternative:

using half interval width

$$\frac{\hat{\sigma}}{\sqrt{25}} \times 2.064 = \frac{1}{2}(2.232 - 1.128)$$

M1

Rearranging any of these to solve,

$$\hat{\sigma} = 1.3372\dots$$

M1

$$\hat{\sigma}^2 = 1.778\dots = 1.79 \text{ (3sf)}$$

A1

(b) Obtain a 95% confidence interval for σ^2

(3)

Formula book:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

Use χ^2 distribution,

$n = 25$, so need row $\nu = 24$,

95% confidence interval uses 0.975 and 0.025 columns in the tables.

So values used are 12.401 and 39.364

B1

$$12.401 < \frac{24 \times 1.788}{\sigma^2} < 39.364$$

M1

$$1.09 < \sigma^2 < 3.46$$

A1

Question 4

The times, x seconds, taken by the competitors in the 100 m freestyle events at a school swimming gala are recorded. The following statistics are obtained from the data.

	No of competitors	Sample mean \bar{x}	Σx^2
Girls	8	83.1	55 746
Boys	7	88.9	56 130

Following the gala, a mother claims that girls are faster swimmers than boys. Assuming that the times taken by the competitors are two independent random samples from normal distributions,

- (a) test, at the 10% level of significance, whether or not the variances of the two distributions are the same. State your hypotheses clearly.

(7)

Hypotheses:

$$H_0 : \sigma_G^2 = \sigma_B^2 \quad H_1 : \sigma_G^2 \neq \sigma_B^2 \quad \text{B1}$$

$$s^2 = \frac{1}{n-1} \left(\Sigma x^2 - \frac{(\Sigma x)^2}{n} \right) = \frac{\Sigma x^2 - n\bar{x}^2}{n-1}$$

$$s_B^2 = \frac{1}{6} (56130 - 7 \times 88.9^2) \quad \text{M1}$$

$$= \frac{1}{6} (807.53) = 134.6 \quad \text{A1}$$

$$s_G^2 = \frac{1}{7} (55746 - 8 \times 83.1^2)$$

$$= \frac{1}{7} (501.12) = 71.59 \quad \text{A1}$$

Formula book:

For a random sample of n_x observations from $N(\mu_x, \sigma_x^2)$ and, independently, a random sample of n_y observations from $N(\mu_y, \sigma_y^2)$

$$\frac{S_x^2 / \sigma_x^2}{S_y^2 / \sigma_y^2} \sim F_{n_x-1, n_y-1}$$

Using the F distribution as the model,

$$\frac{s_B^2}{s_G^2} = 1.880... \quad \text{M1}$$

F distribution tables:

0.05 section, $v_1 = 6$, $v_2 = 7$,

Critical value $F_{6,7} = 3.87$ B1

$1.88 < 3.87$, not significant, variances can be treated as the same. A1

(b) Stating your hypotheses clearly, test the mother's claim. Use a 5% level of significance. (6)

Hypotheses:

$H_0 : \mu_B = \mu_G$ $H_1 : \mu_B > \mu_G$ B1

Formula book:

If $\sigma_x^2 = \sigma_y^2 = \sigma^2$ (unknown) then

$$\frac{(X - Y) - (\mu_x - \mu_y)}{\sqrt{S_p^2 \left(\frac{1}{n_x} + \frac{1}{n_y} \right)}} \sim t_{n_x + n_y - 2} \quad \text{where} \quad S_p^2 = \frac{(n_x - 1)S_x^2 + (n_y - 1)S_y^2}{n_x + n_y - 2}$$

Pooled estimate of variance $s^2 = \frac{6 \times 134.6 + 7 \times 71.59}{13} = 100.6653846... \quad \text{M1}$

$s = 10.0332...$

Test statistic $t = \frac{88.9 - 83.1}{10.03 \sqrt{\frac{1}{7} + \frac{1}{8}}} \quad \text{M1}$

$$= 1.1169566... = 1.12 \text{ (3sf)} \quad \text{A1}$$

t distribution tables:

0.05 column, $v = 13$

Critical value $t = 1.771$ B1

$1.12 < 1.771$, not significant.

Insufficient evidence to support mother's claim. A1

Question 5

Scaffolding poles come in two sizes, long and short. The length L of a long pole has the normal distribution $N(19.6, 0.6^2)$. The length S of a short pole has the normal distribution $N(4.8, 0.3^2)$. The random variables L and S are independent.

A long pole and a short pole are selected at random.

- (a) Find the probability that the length of the long pole is more than 4 times the length of the short pole. Show your working clearly.

(6)

$$\text{Let } X = L - 4S$$

$$\begin{aligned} E(X) &= E(L) - 4E(S) \\ &= 19.6 - 4 \times 4.8 \\ &= 0.4 \end{aligned} \quad \begin{array}{l} \text{M1} \\ \text{A1} \end{array}$$

$$\begin{aligned} \text{Var}(X) &= \text{Var}(L) + 4^2 \text{Var}(S) \\ &= 0.6^2 + 16 \times 0.3^2 \\ &= 1.8 \end{aligned} \quad \begin{array}{l} \text{M1} \\ \text{A1} \end{array}$$

For $L > 4S$ need $X > 0$

$$P(X > 0) = P\left(Z > \frac{0 - 0.4}{\sqrt{1.8}}\right) \quad \text{M1}$$

$$= P(Z > -0.298)$$

$$= 0.617202... \quad \text{A1}$$

$$= 0.617 \text{ (3dp)}$$

Four short poles are selected at random and placed end to end in a row. The random variable T represents the length of the row.

- (b) Find the distribution of T .

(3)

$$\text{Let } T = S_1 + S_2 + S_3 + S_4 \quad \text{M1}$$

$$E(T) = 4 \times 4.8 = 19.2 \quad \text{B1}$$

$$\text{Var}(T) = 4 \times 0.3^2 = 0.36 \quad \text{A1}$$

(c) Find $P(|L - T| < 0.2)$

(4)

Let $Y = L - T$

$$E(Y) = E(L) - E(T) = 0.4 \quad \text{M1}$$

$$\text{Var}(Y) = \text{Var}(L) + \text{Var}(T) = 0.72 \quad \text{M1}$$

$$P(|L - T| < 0.2) = P(-0.2 < Y < 0.2) \quad \text{M1}$$

$$= P(-0.7071 < Z < -0.2357)$$

$$= 0.16708... \quad \text{A1}$$

$$= 0.167 \text{ (3dp)}$$

Note:

Candidates are expected to have a calculator that can access probabilities from statistical distributions including the normal distribution. It may not be necessary to standardise before obtaining the probability.

Question 6

A random sample of 10 female pigs was taken. The number of piglets, x , born to each female pig and their average weight at birth, m kg, was recorded. The results were as follows:

Number of piglets, x	4	5	6	7	8	9	10	11	12	13
Average weight at birth, m kg	1.50	1.20	1.40	1.40	1.23	1.30	1.20	1.15	1.25	1.15

(You may use $S_{xx} = 82.5$ and $S_{mm} = 0.12756$ and $S_{xm} = -2.29$)

- (a) Find the equation of the regression line of m on x in the form $m = a + bx$ as a model for these results.

(2)

Formula book:

The regression coefficient of y on x is $b = \frac{S_{xy}}{S_{xx}}$

Least squares regression line of y on x is $y = a + bx$ where $a = \bar{y} - b\bar{x}$

$$b = \frac{-2.29}{82.5} = -0.0277576$$

$$\bar{x} = \frac{85}{10} = 8.5, \quad \bar{m} = \frac{12.78}{10} = 1.278$$

$$a = 1.278 - (-0.0277576) \times 8.5 = 1.5139$$

M1

$$m = 1.5139 - 0.02776x$$

A1

(b) Show that the residual sum of squares (RSS) is 0.064 to 3 decimal places.

(2)

Formula book:

$$\text{Residual Sum of Squares (RSS)} = S_{yy} - \frac{(S_{xy})^2}{S_{xx}}$$

$$\text{RSS} = 0.12756 - \frac{(-2.29)^2}{82.5} \quad \text{M1}$$

$$= 0.063995\dots = 0.064 \text{ (3dp)} \quad \text{A1}$$

(c) Calculate the residual values.

(2)

x	m	$m = a + bx$	ϵ
4	1.50	1.4029	+0.0971
5	1.20	1.3752	-0.1752
6	1.40	1.3474	+0.0526
7	1.40	1.3196	+0.0804
8	1.23	1.2919	-0.0619
9	1.30	1.2641	+0.0359
10	1.20	1.2364	-0.0364
11	1.15	1.2086	-0.0586
12	1.25	1.1808	+0.0692
13	1.15	1.1531	-0.0031

M1

A1

(d) Write down the outlier.

(1)

The point (5, 1.20) is an outlier.
(because 0.1752 is much larger than the other residual values)

B1

(e) (i) Comment on the validity of ignoring this outlier.
(ii) Ignoring the outlier, produce another model.
(iii) Use this model to estimate the average weight at birth if $x = 15$
(iv) Comment, giving a reason, on the reliability of your estimate.

(5)

(i)

Either:

It is a valid piece of data so should be used.

Or:

It does not follow the pattern according to the residuals so may contain an error making the result invalid so should be removed. B1

(ii)

Removing the outlier leaves $n = 9$

Using a calculator:

$$b = 0.03765$$

$$a = 1.6213$$

M1

$$m = 1.6213 - 0.03765x$$

A1

(iii)

$$m = 1.6213 - 0.03765 \times 15$$

$$= 1.06 \text{ (3sf)}$$

B1

(iv)

The model is only reliable if the values are limited to those in the given range, so probably not reliable. B1

Note:

Candidates are expected to have a calculator that can compute summary statistics, including those required for a regression line.

Question 7

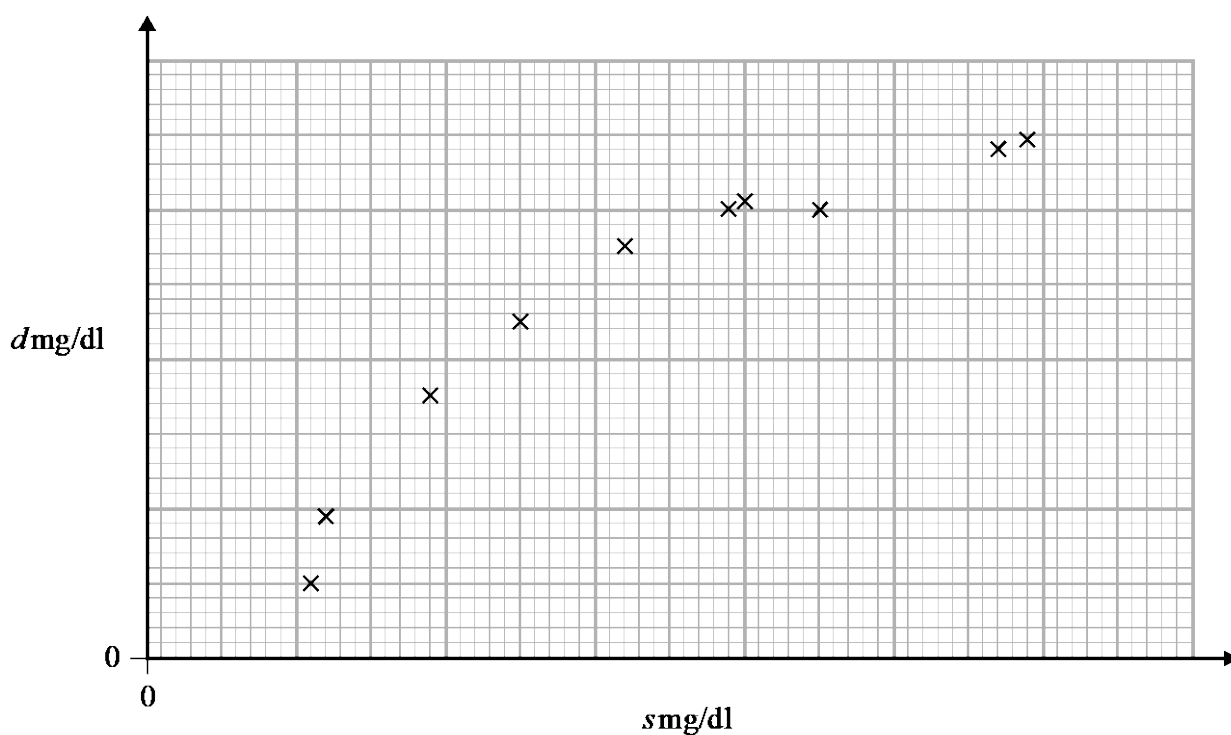
Over a period of time, researchers took 10 blood samples from one patient with a blood disease. For each sample, they measured the levels of serum magnesium, s mg/dl, in the blood and the corresponding level of the disease protein, d mg/dl. One of the researchers coded the data for each sample using $x = 10s$ and $y = 10(d - 9)$ but spilt ink over his work.

The following summary statistics and unfinished scatter diagram are the only remaining information.

$$\sum d^2 = 1081.74 \qquad S_{ds} = 59.524$$

and

$$\sum y = 64 \qquad S_{xx} = 2658.9$$



(a) Use the formula for S_{xx} to show that $S_{ss} = 26.589$

(3)

Formula book:

$$S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$$

$$x = 10s$$

$$S_{xx} = \sum(10s)^2 - \frac{(\sum 10s)^2}{10} \quad \text{M1}$$

$$2658.9 = 100\sum(s)^2 - 100\frac{(\sum s)^2}{10} \quad \text{M1}$$

$$2658.9 = 100\left\{\sum(s)^2 - \frac{(\sum s)^2}{10}\right\}$$

$$2658.9 = 100 S_{ss}$$

$$S_{ss} = 26.589 \quad \text{A1}$$

(b) Find the value of the product moment correlation coefficient between s and d . (4)

$$64 = \sum 10(d - 9) \quad \text{M1}$$

$$64 = 10\sum d - 900$$

$$\sum d = 96.4 \quad \text{A1}$$

$$S_{dd} = \sum d^2 - \frac{(\sum d)^2}{n}$$

$$S_{dd} = 1081.74 - \frac{(96.4)^2}{10} \quad \text{M1}$$

$$= 152.444$$

Formula book:

Product moment correlation coefficient:

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}$$

$$r = \frac{59.524}{\sqrt{26.589 \times 152.444}}$$

$$r = 0.935$$

A1

(c) With reference to the unfinished scatter diagram, comment on your result in part (b).

(1)

Linear correlation is significant but scatter diagram suggests a non linear relationship between the level of serum magnesium, and the level of the disease protein. B1

For more information on Edexcel and BTEC qualifications please visit our websites:
www.edexcel.com and www.btec.co.uk

Edexcel is a registered trademark of Pearson Education Limited

Pearson Education Limited. Registered in England and Wales No. 872828
Registered Office: 80 Strand, London WC2R 0RL.
VAT Reg No GB 278 537121