Topic 7 Part 1 *[328 marks]*

The random variable X has probability distribution $\mathrm{Po}(8)$.

1a. (i) Find $\mathrm{P}(X = 6)$.

(ii) Find $\mathrm{P}(X = 6 | 5 \leqslant X \leqslant 8)$.

[5 marks]

Markscheme

(i)

$\mathrm{P}(X = 6) = 0.122$ ***(M1)A1***

(ii)

$\mathrm{P}(X = 6 | 5 \leqslant X \leqslant 8) = \frac{\mathrm{P}(X=6)}{\mathrm{P}(5 \leqslant X \leqslant 8)} = \frac{0.122\ldots}{0.592\ldots - 0.0996\ldots}$ ***(M1)(A1)***

$= 0.248$ ***A1***

[5 marks]

Examiners report

[N/A]

1b. $\bar{X}$ denotes the sample mean of $n > 1$ independent observations from X.

(i) Write down $\mathrm{E}(\bar{X})$ and $\mathrm{Var}(\bar{X})$.

(ii) Hence, give a reason why $\bar{X}$ is not a Poisson distribution.

[3 marks]

Markscheme

(i)

$\mathrm{E}(\bar{X}) = 8$ ***A1***

$\mathrm{Var}(\bar{X}) = \frac{8}{n}$ ***A1***

(ii)

$\mathrm{E}(\bar{X}) \neq \mathrm{Var}(\bar{X})$

(for $n > 1$) ***R1***

Note: Only award the ***R1*** if the two expressions in (b)(i) are different.

[3 marks]

Examiners report

[N/A]

1c. A random sample of 40 observations is taken from the distribution for X. *[6 marks]*

(i) Find $\mathrm{P}(7.1 < \bar{X} < 8.5)$.

(ii) Given that $\mathrm{P}\left(\left|\bar{X} - 8\right| \leqslant k\right) = 0.95$, find the value of k.

Markscheme

(i) **EITHER**

$\bar{X} \sim \mathrm{N}(8,\, 0.2)$ ***(M1)A1***

Note: ***M1*** for normality, ***A1*** for parameters.

$\mathrm{P}(7.1 < \bar{X} < 8.5) = 0.846$ ***A1***

OR

The expression is equivalent to $\mathrm{P}(283 \leqslant \sum X \leqslant 339)$ where $\sum X$ is $\mathrm{Po}(320)$ ***M1A1***

$= 0.840$ ***A1***

Note: Accept 284, 340 instead of 283, 339

Accept any answer that rounds correctly to 0.84 or 0.85.

(ii) **EITHER**

$k = 1.96\frac{\sigma}{\sqrt{n}}$ or $1.96\,\mathrm{std}(\bar{X})$ ***(M1)(A1)***

$k = 0.877$ or $1.96\sqrt{0.2}$ ***A1***

OR

The expression is equivalent to

$P(320 - 40k \leqslant \sum X \leqslant 320 + 40k) = 0.95$ ***(M1)***

$k = 0.875$ ***A2***

Note: Accept any answer that rounds to 0.87 or 0.88.

Award ***M1A0*** if modulus sign ignored and answer obtained rounds to 0.74 or 0.75

[6 marks]

Examiners report

[N/A]

2. The following table gives the average yield of olives per tree, in kg, and the rainfall, in cm, for nine separate regions of Greece. You may assume that these data are a random sample from a bivariate normal distribution, with correlation coefficient ρ. *[16 marks]*

Rainfall (x)	11	10	15	13	7	18	22	20	28
Yield (y)	56	53	67	61	54	78	86	88	78

A scientist wishes to use these data to determine whether there is a positive correlation between rainfall and yield.

(a) State suitable hypotheses.

(b) Determine the product moment correlation coefficient for these data.

(c) Determine the associated p-value and comment on this value in the context of the question.

(d) Find the equation of the regression line of y on x.

(e) Hence, estimate the yield per tree in a tenth region where the rainfall was 19 cm.

(f) Determine the angle between the regression line of y on x and that of x on y . Give your answer to the nearest degree.

Markscheme

(a)

$H_0 : \rho = 0$ ***A1***

$H_1 : \rho > 0$ ***A1***

[2 marks]

(b) 0.853 ***A2***

Note: Accept any answer that rounds to 0.85.

[2 marks]

(c) p-value = 0.00173 (1-tailed) ***A1***

Note: Accept any answer that rounds to 0.0017.
Accept any answer that rounds to 0.0035 obtained from 2-tailed test.

strong evidence to reject the hypothesis that there is no correlation between rainfall and yield or to accept the hypothesis that there is correlation between rainfall and yield ***R1***

Note: Follow through the p-value for the conclusion.

[2 marks]

(d)

$y = 1.78x + 40.5$ ***A1A1***

Note: Accept numerical coefficients that round to 1.8 and 41.

[2 marks]

(e)

$y = 1.77\ldots(19) + 14.5\ldots$ ***M1***

74.3 ***A1***

Note: Accept any answer that rounds to 74 or 75.

[2 marks]

(f) the gradient of the regression line y on x is 1.78 or equivalent ***A1***

the regression line of x on y is

$x = 0.409y - 12.2$ ***(A1)***

the gradient of the regression line x on y is

$\frac{1}{0.409}\ (= 2.44)$ ***(M1)A1***

calculate

$\arctan(2.44) - \arctan(1.78)$ ***(M1)***

angle between regression lines is 7 degrees ***A1***

Note: Accept any answer which rounds to ±7 degrees.

[6 marks]

Total [16 marks]

Examiners report

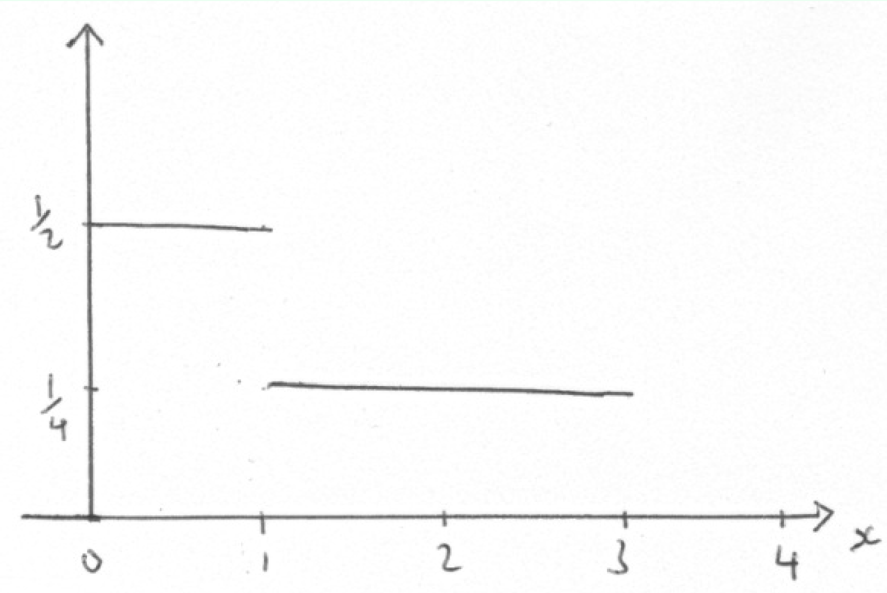
[N/A]

A random variable X has probability density function

$$f(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{2} & 0 \le x < 1 \\ \frac{1}{4} & 1 \le x < 3 \\ 0 & x \ge 3 \end{cases}$$

3a. Sketch the graph of $y = f(x)$. *[1 mark]*

Markscheme



A1

Note: Ignore open / closed endpoints and vertical lines.

Note: Award ***A1*** for a correct graph with scales on both axes and a clear indication of the relevant values.

[1 mark]

Examiners report

Part (a) was correctly answered by most candidates. Some graphs were difficult to mark because candidates drew their lines on top of the ruled lines in the answer book. Candidates should be advised not to do this. Candidates should also be aware that the command term ‘sketch’ requires relevant values to be indicated.

3b. Find the cumulative distribution function for X. *[5 marks]*

Markscheme

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{x}{2} & 0 \le x < 1 \\ \frac{x}{4} + \frac{1}{4} & 1 \le x < 3 \\ 1 & x \ge 3 \end{cases}$$

considering the areas in their sketch or using integration ***(M1)***

$F(x) = 0,\ x < 0,\ F(x) = 1,\ x \ge 3$ ***A1***

$F(x) = \frac{x}{2},\ 0 \le x < 1$ ***A1***

$F(x) = \frac{x}{4} + \frac{1}{4},\ 1 \le x < 3$ ***A1A1***

Note: Accept $<$ for $\le$ in all places and also $>$ for $\ge$ first ***A1***.

[5 marks]

Examiners report

In (b), most candidates realised that the cumulative distribution function had to be found by integration but the limits were sometimes incorrect.

3c. Find the interquartile range for X. *[3 marks]*

Markscheme

$Q_3 = 2,\ Q_1 = 0.5$ ***A1A1***

$\text{IQR is } 2 - 0.5 = 1.5$ ***A1***

[3 marks]

Total [9 marks]

Examiners report

In (c), candidates who found the upper and lower quartiles correctly sometimes gave the interquartile range as $[0.5,\ 2]$. It is important for candidates to realise that that the word range has a different meaning in statistics compared with other branches of mathematics.

Eric plays a game at a fairground in which he throws darts at a target. Each time he throws a dart, the probability of hitting the target is 0.2. He is allowed to throw as many darts as he likes, but it costs him $\$1$ a throw. If he hits the target a total of three times he wins $\$10$.

4a. Find the probability he has his third success of hitting the target on his sixth throw. *[3 marks]*

Markscheme

METHOD 1

let X be the number of throws until Eric hits the target three times

$X \sim \mathrm{NB}(3, 0.2)$ ***(M1)***

$\mathrm{P}(X = 6) = \binom{5}{2} 0.8^3 \times 0.2^3$ ***(A1)***

$= 0.04096 \quad \left(= \frac{128}{3125}\right)$ (exact) ***A1***

METHOD 2

let X be the number of hits in five throws

X is $\mathrm{B}(5,\ 0.2)$ ***(M1)***

$\mathrm{P}(X = 2) = \binom{5}{2} 0.2^2 \times 0.8^3 \quad (0.2048)$ ***(A1)***

P(3rd hit on 6th throw) $= \binom{5}{2} 0.2^2 \times 0.8^3 \times 0.2 = 0.04096 \left(= \frac{128}{3125}\right)$ (exact) ***A1***

[3 marks]

Examiners report

Part (a) was well answered, using the negative binomial distribution $NB(3,\ 0.2)$, by many candidates. Some candidates began by using the binomial distribution $B(5,\ 0.2)$ which is a valid method as long as it is followed by multiplying by 0.2 but this final step was not always carried out successfully.

4b. (i) Find the expected number of throws required for Eric to hit the target three times. *[3 marks]*

(ii) Write down his expected profit or loss if he plays until he wins the $\$10$.

Markscheme

(i) $\text{expected number of throws} = \frac{3}{0.2} = 15$ ***(M1)A1***

(ii) $\text{profit} = (10 - 15) = -\$5 \text{ or loss} = \$5$ ***A1***

[3 marks]

Examiners report

Part (b) was well answered by the majority of candidates.

4c. If he has just $\$8$, find the probability he will lose all his money before he hits the target three times. *[3 marks]*

Markscheme

METHOD 1

let Y be the number of times the target is hit in 8 throws

$Y \sim \mathrm{B}(8,\ 0.2)$ ***(M1)***

$\mathrm{P}(Y \le 2)$ ***(M1)***

$= 0.797$ ***A1***

METHOD 2

let the 3^{rd} hit occur on the Y^{th} throw

Y is $\mathrm{NB}(3,\ 0.2)$ ***(M1)***

$\mathrm{P}(Y > 8) = 1 - \mathrm{P}(Y \le 8)$ ***(M1)***

$= 0.797$ ***A1***

[3 marks]

Total [9 marks]

Examiners report

In (c), candidates who used the binomial distribution $B(8,\ 0.2)$ were generally successful. Candidates who used the negative binomial distribution $Y \approx NB(3,\ 0.2)$ to evaluate $P(Y > 8)$ were usually unsuccessful because of the large amount of computation involved.

5a. If X and Y are two random variables such that $\mathrm{E}(X) = \mu_X$ and $\mathrm{E}(Y) = \mu_Y$ then $\mathrm{Cov}(X,\ Y) = \mathrm{E}\left((X - \mu_X)(Y - \mu_Y)\right)$. *[3 marks]*

Prove that if X and Y are independent then $\mathrm{Cov}(X,\ Y) = 0$.

Markscheme

METHOD 1

$\mathrm{Cov}(X,\ Y) = \mathrm{E}\left((X - \mu_X)(Y - \mu_Y)\right)$

$= \mathrm{E}(XY - X\mu_Y - Y\mu_X + \mu_X\mu_Y)$ ***(M1)***

$= \mathrm{E}(XY) - \mu_Y\mathrm{E}(X) - \mu_X\mathrm{E}(Y) + \mu_X\mu_Y$

$= \mathrm{E}(XY) - \mu_X\mu_Y$ ***A1***

as X and Y are independent $\mathrm{E}(XY) = \mu_X\mu_Y$ ***R1***

$\mathrm{Cov}(X,\ Y) = 0$ ***AG***

METHOD 2

$\mathrm{Cov}(X,\ Y) = \mathrm{E}\left((X - \mu_x)(Y - \mu_y)\right)$

$= \mathrm{E}(X - \mu_x)\mathrm{E}(Y - \mu_y)$ ***(M1)***

since X, Y are independent ***R1***

$= (\mu_x - \mu_x)(\mu_y - \mu_y)$ ***A1***

$= 0$ ***AG***

[3 marks]

Examiners report

Solutions to (a) were often disappointing with few candidates gaining full marks, a common error being failure to state that

$E(XY) = E(X)E(Y)$ or $\mathrm{E}\left((X - \mu_x)(Y - \mu_y)\right) = \mathrm{E}(X - \mu_x)\mathrm{E}(Y - \mu_y)$ in the case of independence.

5b. In a particular company, it is claimed that the distance travelled by employees to work is independent of their salary. To test this, 20 randomly selected employees are asked about the distance they travel to work and the size of their salaries. It is found that the product moment correlation coefficient, r, for the sample is -0.35. *[8 marks]*

You may assume that both salary and distance travelled to work follow normal distributions.

Perform a one-tailed test at the 5% significance level to test whether or not the distance travelled to work and the salaries of the employees are independent.

Markscheme

$H_0 : \rho = 0 \quad H_1 : \rho < 0$ ***A1***

Note: The hypotheses must be expressed in terms of ρ.

test statistic $t_{test} = -0.35\sqrt{\frac{20-2}{1-(-0.35)^2}}$ ***(M1)(A1)***

$= -1.585\ldots$ ***(A1)***

$\text{degrees of freedom} = 18$ ***(A1)***

EITHER

$p\text{ - value} = 0.0652$ ***A1***

this is greater than 0.05 ***M1***

OR

$t_{5\%}(18) = -1.73$ ***A1***

this is less than -1.59 ***M1***

THEN

hence accept H_0 or reject H_1 or equivalent or contextual equivalent ***R1***

Note: Allow follow through for the final ***R1*** mark.

[8 marks]

Total [11 marks]

Examiners report

In (b), the hypotheses were sometimes given incorrectly. Some candidates gave H_1 as $\rho \neq 0$, not seeing that a one-tailed test was required. A more serious error was giving the hypotheses as $H_0 : r = 0,\ H_1 : r < 0$ which shows a complete misunderstanding of the situation. Subsequent parts of the question were well answered in general.

If X is a random variable that follows a Poisson distribution with mean $\lambda > 0$ then the probability generating function of X is $G(t) = e^{\lambda(t-1)}$.

6a. (i) Prove that $\mathrm{E}(X) = \lambda$. *[6 marks]*

(ii) Prove that $\mathrm{Var}(X) = \lambda$.

Markscheme

(i) $G'(t) = \lambda e^{\lambda(t-1)}$ ***A1***

$\mathrm{E}(X) = G'(1)$ ***M1***

$= \lambda$ ***AG***

(ii) $G''(t) = \lambda^2 e^{\lambda(t-1)}$ ***M1***

$\Rightarrow G''(1) = \lambda^2$ ***(A1)***

$\mathrm{Var}(X) = G''(1) + G'(1) - (G'(1))^2$ ***(M1)***

$= \lambda^2 + \lambda - \lambda^2$ ***A1***

$= \lambda$ ***AG***

[6 marks]

Examiners report

Solutions to the different parts of this question proved to be extremely variable in quality with some parts well answered by the majority of the candidates and other parts accessible to only a few candidates. Part (a) was well answered in general although the presentation was sometimes poor with some candidates doing the differentiation of $G(t)$ and the substitution of $t = 1$ simultaneously.

6b. Y is a random variable, independent of X, that also follows a Poisson distribution with mean λ. *[3 marks]*

If $S = 2X - Y$ find

(i) $\mathrm{E}(S)$;

(ii) $\mathrm{Var}(S)$.

Markscheme

(i) $\mathrm{E}(S) = 2\lambda - \lambda = \lambda$ ***A1***

(ii) $\mathrm{Var}(S) = 4\lambda + \lambda = 5\lambda$ ***(A1)A1***

Note: First ***A1*** can be awarded for either 4λ or λ.

[3 marks]

Examiners report

Part (b) was well answered in general, the most common error being to state that $\mathrm{Var}(2X - Y) = \mathrm{Var}(2X) - \mathrm{Var}(Y)$.

6c. Let $T = \frac{Y}{2} + \frac{Y}{2}$. *[3 marks]*

(i) Show that T is an unbiased estimator for λ.

(ii) Show that T is a more efficient unbiased estimator of λ than S.

Markscheme

(i) $\mathrm{E}(T) = \frac{\lambda}{2} + \frac{\lambda}{2} = \lambda$ (so T is an unbiased estimator) ***A1***

(ii) $\mathrm{Var}(T) = \frac{1}{4}\lambda + \frac{1}{4}\lambda = \frac{1}{2}\lambda$ ***A1***

this is less than $\mathrm{Var}(S)$, therefore T is the more efficient estimator ***R1AG***

Note: Follow through their variances from (b)(ii) and (c)(ii).

[3 marks]

Examiners report

Parts (c) and (d) were well answered by the majority of candidates.

6d. Could either S or T model a Poisson distribution? Justify your answer. *[1 mark]*

Markscheme

no, mean does not equal the variance ***R1***

[1 mark]

Examiners report

Parts (c) and (d) were well answered by the majority of candidates.

6e. By consideration of the probability generating function, $G_{X+Y}(t)$, of $X+Y$, prove that $X+Y$ follows a Poisson distribution with mean 2λ. *[3 marks]*

Markscheme

$G_{X+Y}(t) = e^{\lambda(t-1)} \times e^{\lambda(t-1)} = e^{2\lambda(t-1)}$ ***M1A1***

which is the probability generating function for a Poisson with a mean of 2λ ***R1AG***

[3 marks]

Examiners report

Solutions to (e), however, were extremely disappointing with few candidates giving correct solutions. A common incorrect solution was the following:

$$G_{X+Y}(t) = G_X(t)G_Y(t)$$

Differentiating,

$$G'_{X+Y}(t) = G'_X(t)G_Y(t) + G_X(t)G'_Y(t)$$

$$\mathrm{E}(X+Y) = G'_{X+Y}(1) = \mathrm{E}(X) \times 1 + \mathrm{E}(Y) \times 1 = 2\lambda$$

This is correct mathematics but it does not show that $X+Y$ is Poisson and it was given no credit. Even the majority of candidates who showed that $G_{X+Y}(t) = \mathrm{e}^{2\lambda(t-1)}$ failed to state that this result proved that $X+Y$ is Poisson and they usually differentiated this function to show that $\mathrm{E}(X+Y) = 2\lambda$.

6f. Find [2 marks]

(i) $G_{X+Y}(1)$;

(ii) $G_{X+Y}(-1)$.

Markscheme

(i) $G_{X+Y}(1) = 1$ ***A1***

(ii) $G_{X+Y}(-1) = e^{-4\lambda}$ ***A1***

[2 marks]

Examiners report

In (f), most candidates stated that $G_{X+Y}(1) = 1$ even if they were unable to determine $G_{X+Y}(t)$ but many candidates were unable to evaluate $G_{X+Y}(-1)$. Very few correct solutions were seen to (g) even if the candidates correctly evaluated $G_{X+Y}(1)$ and $G_{X+Y}(-1)$.

6g. Hence find the probability that $X+Y$ is an even number. [3 marks]

Markscheme

$G_{X+Y}(1) = p(0) + p(1) + p(2) + p(3) \ldots$

$G_{X+Y}(-1) = p(0) - p(1) + p(2) - p(3) \ldots$

so $2\mathrm{P}(\text{even}) = G_{X+Y}(1) + G_{X+Y}(-1)$ ***(M1)(A1)***

$\mathrm{P}(\text{even}) = \frac{1}{2}(1 + e^{-4\lambda})$ ***A1***

[3 marks]

Total [21 marks]

Examiners report

[N/A]

Two species of plant, A and B, are identical in appearance though it is known that the mean length of leaves from a plant of species A is 5.2 cm, whereas the mean length of leaves from a plant of species B is 4.6 cm. Both lengths can be modelled by normal distributions with standard deviation 1.2 cm.

In order to test whether a particular plant is from species A or species B, 16 leaves are collected at random from the plant. The length, x, of each leaf is measured and the mean length evaluated. A one-tailed test of the sample mean, $\bar{X}$, is then performed at the 5% level, with the hypotheses: $H_0 : \mu = 5.2$ and $H_1 : \mu < 5.2$.

7a. Find the critical region for this test. [3 marks]

Markscheme

$\bar{X} \sim N\left(5.2,\ \frac{1.2^2}{16}\right)$ ***(M1)***

critical value is $5.2 - 1.64485\ldots \times \frac{1.2}{4} = 4.70654\ldots$ ***(A1)***

critical region is $]-\infty,\ 4.71]$ ***A1***

Note: Allow follow through for the final ***A1*** from their critical value.

Note: Follow through previous values in (b), (c) and (d).

[3 marks]

Examiners report

Solutions to this question were generally disappointing.

In (a), the standard error of the mean was often taken to be $\sigma(1.2)$ instead of $\frac{\sigma}{\sqrt{n}}(0.3)$ and the solution sometimes ended with the critical value without the critical region being given.

7b. It is now known that in the area in which the plant was found 90% of all the plants are of species A and 10% are of species B. *[2 marks]*

Find the probability that $\bar{X}$ will fall within the critical region of the test.

Markscheme

$0.9 \times 0.05 + 0.1 \times (1 - 0.361\ldots) = 0.108875997\ldots = 0.109$ ***M1A1***

Note: Award ***M1*** for a weighted average of probabilities with weights $0.1, 0.9$.

[2 marks]

Examiners report

In (c), the question was often misunderstood with candidates finding the weighted mean of the two means, ie $0.9 \times 5.2 + 0.1 \times 4.6 = 5.14$ instead of the weighted mean of two probabilities.

7c. If, having done the test, the sample mean is found to lie within the critical region, find the probability that the leaves came from a plant of species A. *[3 marks]*

Markscheme

attempt to use conditional probability formula ***M1***

$\frac{0.9 \times 0.05}{0.108875997\ldots}$ ***(A1)***

$= 0.41334\ldots = 0.413$ ***A1***

[3 marks]

Total [10 marks]

Examiners report

Without having the solution to (c), part (d) was inaccessible to most of the candidates so that very few correct solutions were seen.

Engine oil is sold in cans of two capacities, large and small. The amount, in millilitres, in each can, is normally distributed according to Large $\sim \mathrm{N}(5000,\ 40)$ and Small $\sim \mathrm{N}(1000,\ 25)$.

8a. A large can is selected at random. Find the probability that the can contains at least 4995 millilitres of oil. *[2 marks]*

Markscheme

$\mathrm{P}(L \geq 4995) = 0.785$ ***(M1)A1***

Note: Accept any answer that rounds correctly to 0.79.

Award ***M1A0*** for 0.78.

Note: Award ***M1A0*** for any answer that rounds to 0.55 obtained by taking $\mathrm{SD} = 40$.

[2 marks]

Examiners report

Most candidates solved (a) correctly. In (b) and (c), however, many candidates made the usual error of confusing $\sum_{i=1}^{n} X_i$ and nX. Indeed some candidates even use the second expression to mean the first. This error leads to an incorrect variance and of course an incorrect answer. Some candidates had difficulty in converting the verbal statements into the correct probability statements, particularly in (c).

8b. A large can and a small can are selected at random. Find the probability that the large can contains at least 30 milliliters more than five times the amount contained in the small can. *[6 marks]*

Markscheme

we are given that $L \sim \mathrm{N}(5000,\ 40)$ and $S \sim \mathrm{N}(1000,\ 25)$

consider $X = L - 5S$ (ignore ± 30) ***(M1)***

$\mathrm{E}(X) = 0$ (± 30 consistent with line above) ***A1***

$\mathrm{Var}(X) = \mathrm{Var}(L) + 25\mathrm{Var}(S) = 40 + 625 = 665$ ***(M1)A1***

require $\mathrm{P}(X \geq 30) \quad (\text{or } \mathrm{P}(X \geq 0) \text{ if } -30 \text{ above})$ ***(M1)***

obtain 0.122 ***A1***

Note: Accept any answer that rounds correctly to 2 significant figures.

[6 marks]

Examiners report

Most candidates solved (a) correctly. In (b) and (c), however, many candidates made the usual error of confusing $\sum_{i=1}^{n} X_i$ and nX. Indeed some candidates even use the second expression to mean the first. This error leads to an incorrect variance and of course an incorrect answer. Some candidates had difficulty in converting the verbal statements into the correct probability statements, particularly in (c).

8c. A large can and five small cans are selected at random. Find the probability that the large can contains at least 30 milliliters less than the total amount contained in the small cans. *[5 marks]*

Markscheme

consider $Y = L - (S_1 + S_2 + S_3 + S_4 + S_5)$ (ignore ± 30)   ***(M1)***

$\mathrm{E}(Y) = 0$ (± 30 consistent with line above)   ***A1***

$\mathrm{Var}(Y) = 40 + 5 \times 25 = 165$ ***A1***

require $\mathrm{P}(Y \le -30)$ (or $\mathrm{P}(Y \le 0)$ if $+30$ above)   ***(M1)***

obtain 0.00976 ***A1***

Note: Accept any answer that rounds correctly to 2 significant figures.

Note: Condone the notation $Y = L - 5S$ if the variance is correct.

[5 marks]

Total [13 marks]

Examiners report

Most candidates solved (a) correctly. In (b) and (c), however, many candidates made the usual error of confusing $\sum_{i=1}^{n} X_i$ and nX. Indeed some candidates even use the second expression to mean the first. This error leads to an incorrect variance and of course an incorrect answer. Some candidates had difficulty in converting the verbal statements into the correct probability statements, particularly in (c).

Eleven students who had under-performed in a philosophy practice examination were given extra tuition before their final examination. The differences between their final examination marks and their practice examination marks were

$$10,\ -1,\ 6,\ 7,\ -5,\ -5,\ 2,\ -3,\ 8,\ 9,\ -2.$$

Assume that these differences form a random sample from a normal distribution with mean μ and variance σ^2.

9a. Determine unbiased estimates of μ and σ^2. *[4 marks]*

Markscheme

unbiased estimate of μ is $2.36(36\ldots)$ $(26/11)$ ***(M1)A1***

unbiased estimate of σ^2 is $33.65(45\ldots) = (5.801^2)$ $(1851/55)$ ***(M1)A1***

Note: Accept any answer that rounds correctly to 3 significant figures.

Note: Award ***M1A0*** for any unbiased estimate of σ^2 that rounds to 5.80.

[4 marks]

Examiners report

Almost every candidate gave the correct estimate of the mean but some chose the wrong variance from their calculators to estimate σ^2. In (b)(i), the hypotheses were sometimes incorrectly written, usually with an incorrect symbol instead of μ, for example d, $\bar{x}$ and 'mean' were seen. Many candidates failed to make efficient use of their calculators in (b)(ii). The intention of the question was that candidates should simply input the data into their calculators and use the software to give the *p*-value. Instead, many candidates found the *p*-value by first evaluating t using the appropriate formula. This was a time consuming process and it gave opportunity for error. In (b)(iii), candidates were expected to refer to the claim so that the answers 'Accept H_0' or 'Reject H_1' were not accepted.

9b. (i) State suitable hypotheses to test the claim that extra tuition improves examination marks. *[8 marks]*

(ii) Calculate the p-value of the sample.

(iii) Determine whether or not the above claim is supported at the 5% significance level.

Markscheme

(i) $\mathrm{H}_0 : \mu = 0;\ \mathrm{H}_1 : \mu > 0$ ***A1A1***

Note: Award ***A1A0*** if an inappropriate symbol is used for the mean, *eg*, r, $\bar{\mathrm{d}}$.

(ii) attempt to use *t*-test ***(M1)***

$t = 1.35$ ***(A1)***

$\mathrm{DF} = 10$ ***(A1)***

p-value $= 0.103$ ***A1***

Note: Accept any answer that rounds correctly to 3 significant figures.

(iii) $0.103 > 0.05$ ***A1***

there is insufficient evidence at the 5% level to support the claim (that extra tuition improves examination marks)

OR

the claim (that extra tuition improves examination marks) is not supported at the 5% level (or equivalent statement) ***R1***

Note: Follow through the candidate's p-value.

Note: Do not award ***R1*** for Accept H_0 or Reject H_1.

[8 marks]

Total [12 marks]

Examiners report

Almost every candidate gave the correct estimate of the mean but some chose the wrong variance from their calculators to estimate σ^2. In (b)(i), the hypotheses were sometimes incorrectly written, usually with an incorrect symbol instead of μ, for example d, $\bar{x}$ and 'mean' were seen. Many candidates failed to make efficient use of their calculators in (b)(ii). The intention of the question was that candidates should simply input the data into their calculators and use the software to give the p-value. Instead, many candidates found the p-value by first evaluating t using the appropriate formula. This was a time consuming process and it gave opportunity for error. In (b)(iii), candidates were expected to refer to the claim so that the answers 'Accept H_0' or 'Reject H_1' were not accepted.

A manufacturer of stopwatches employs a large number of people to time the winner of a 100 metre sprint. It is believed that if the true time of the winner is μ seconds, the times recorded are normally distributed with mean μ seconds and standard deviation 0.03 seconds.

The times, in seconds, recorded by six randomly chosen people are

$$9.765,\ 9.811,\ 9.783,\ 9.797,\ 9.804,\ 9.798.$$

10a. Calculate a 99% confidence interval for μ. Give your answer correct to three decimal places. *[4 marks]*

Markscheme

the (unbiased) estimate of μ is 9.793 ***(A1)***

the 99% CI is $9.793 \pm 2.576\frac{0.03}{\sqrt{6}}$ ***(M1)(A1)***

$= [9.761,\ 9.825]$ ***A1***

Note: Accept 9.762 and 9.824.

[4 marks]

Examiners report

The intention in (a) was that candidates should input the data into their calculators and use the software to give the confidence interval. However, as in Question 2, many candidates calculated the mean and variance by hand and used the appropriate formulae to determine the confidence limits. Again valuable time was used up and opportunity for error introduced.

10b. Interpret the result found in (a). *[2 marks]*

Markscheme

if this process is carried out a large number of times ***A1***

(approximately) 99% of the intervals will contain μ ***A1***

Note: Award ***A1A1*** for a consideration of any specific large value of times $(n \geq 100)$.

[2 marks]

Examiners report

Answers to (b) were extremely disappointing with the vast majority giving an incorrect interpretation of a confidence interval. The most common answer given was along the lines of 'There is a 99% probability that the interval [9.761, 9.825] contains μ'. This is incorrect since the interval and μ are both constants; the statement that the interval [9.761, 9.825] contains μ is either true or false, there is no question of probability being involved. Another common response was 'I am 99% confident that the interval [9.761, 9.825] contains μ'. This is unsatisfactory partly because 99% confident is really a euphemism for 99% probability and partly because it answers the question 'What is a 99% confidence interval for μ' by simply rearranging the words without actually going anywhere. The expected answer was that if the sampling was carried out a large number of times, then approximately 99% of the calculated confidence intervals would contain μ. A more rigorous response would be that a 99% confidence interval for μ is an observed value of a random interval which contains μ with probability 0.99 just as the number $\bar{x}$ is an observed value of the random variable $\bar{X}$. The concept of a confidence interval is a difficult one at this level but confidence intervals are part of the programme and so therefore is their interpretation. In view of the widespread misunderstanding of confidence intervals, partial credit was given on this occasion for interpretations involving 99% probability or confidence but this will not be the case in future examinations.

10c. Find the confidence level of the interval that corresponds to halving the width of the 99% confidence interval. Give your answer as a percentage to the nearest whole number. *[3 marks]*

Markscheme

METHOD 1

If the interval is halved, 2.576 becomes 1.288 ***M1***

normal tail probability corresponding to $1.288 = 0.0988\ldots$ ***A1***

confidence level $= 80\%$ ***A1***

METHOD 2

half width $= 0.5 \times 0.063$ or 0.062 or $0.064 = 0.0315$ or 0.031 or 0.032 ***M1***

$\frac{2z \times 0.03}{\sqrt{6}} = 0.0315$ or 0.031 or 0.032

giving $z = 1.285\ldots$ or $1.265\ldots$ or $1.306\ldots$ ***A1***

confidence level $= 80\%$ or
79% or
81% ***A1***

Note: Follow through values from (a).

[3 marks]

Total [9 marks]

Examiners report

Many candidates solved (c) correctly, mostly using Method 2 in the mark scheme.

A random variable
X has a population mean μ.

11a. Explain briefly the meaning of *[3 marks]*

(i) an estimator of μ;

(ii) an unbiased estimator of μ.

Markscheme

(i) an estimator T is a formula (or statistic) that can be applied to the values in any sample, taken from X ***A1***

to estimate the value of μ ***A1***

(ii) an estimator is unbiased if $\mathrm{E}(T) = \mu$ ***A1***

[3 marks]

Examiners report

In general, solutions to (a) were extremely disappointing with the vast majority unable to give correct explanations of estimators and unbiased estimators. Solutions to (b) were reasonably good in general, indicating perhaps that the poor explanations in (a) were due to an inability to explain what they know rather than a lack of understanding.

11b. A random sample $X_1,\ X_2,\ X_3$ of three independent observations is taken from the distribution of X. *[12 marks]*

An unbiased estimator of $\mu,\ \mu \neq 0$, is given by $U = \alpha X_1 + \beta X_2 + (\alpha - \beta)X_3$,

where $\alpha,\ \beta \in \mathbb{R}$.

(i) Find the value of α.

(ii) Show that $\mathrm{Var}(U) = \sigma^2\left(2\beta^2 - \beta + \frac{1}{2}\right)$ where $\sigma^2 = \mathrm{Var}(X)$.

(iii) Find the value of β which gives the most efficient estimator of μ of this form.

(iv) Write down an expression for this estimator and determine its variance.

(v) Write down a more efficient estimator of μ than the one found in (iv), justifying your answer.

Markscheme

(i) using linearity and the definition of an unbiased estimator ***M1***

$\mu = \alpha\mu + \beta\mu + (\alpha - \beta)\mu$ ***A1***

obtain $\alpha = \frac{1}{2}$ ***A1***

(ii) attempt to compute $\mathrm{Var}(U)$ using correct formula ***M1***

$\mathrm{Var}(U) = \frac{1}{4}\sigma^2 + \beta^2\sigma^2 + \left(\frac{1}{2} - \beta\right)^2\sigma^2$ ***A1***

$\mathrm{Var}(U) = \sigma^2\left(2\beta^2 - \beta + \frac{1}{2}\right)$ ***AG***

(iii) attempt to minimise quadratic in β (or equivalent) ***(M1)***

$\beta = \frac{1}{4}$ ***A1***

(iv) $(U) = \frac{1}{2}X_1 + \frac{1}{4}X_2 + \frac{1}{4}X_3$ ***A1***

$\mathrm{Var}(U) = \frac{3}{8}\sigma^2$ ***A1***

(v) $\frac{1}{3}X_1 + \frac{1}{3}X_2 + \frac{1}{3}X_3$ ***A1***

$\mathrm{Var}\left(\frac{1}{3}X_1 + \frac{1}{3}X_2 + \frac{1}{3}X_3\right) = \frac{3}{9}\sigma^2$ ***A1***

$< \mathrm{Var}(U)$ ***R1***

Note: Accept $\sum\limits_{i=1}^{3} \lambda_i X_i$ if $\sum\limits_{i=1}^{3} \lambda_i = 1$ and $\sum\limits_{i=1}^{3} \lambda_i^2 < \frac{3}{8}$ and follow through to the variance if this is the case.

[12 marks]

Total [15 marks]

Examiners report

Solutions to (b) were reasonably good in general, indicating perhaps that the poor explanations in (a) were due to an inability to explain what they know rather than a lack of understanding.

12a. Determine the probability generating function for $X \sim \mathrm{B}(1,\ p)$. *[4 marks]*

Markscheme

$\mathrm{P}(X = 0) = 1 - p(= q);\ \mathrm{P}(X = 1) = p$ ***(M1)(A1)***

$\mathrm{G}_x(t) = \sum\limits_r \mathrm{P}(X = r)t^r$ (or writing out term by term) ***M1***

$= q + pt$ ***A1***

[4 marks]

Examiners report

Solutions to (a) were often disappointing with some candidates simply writing down the answer. A common error was to forget the possibility of X being zero so that $G(t) = pt$ was often seen.

12b. Explain why the probability generating function for $\mathrm{B}(n,\ p)$ is a polynomial of degree n. *[2 marks]*

Markscheme

METHOD 1

PGF for $B(n,\ p)$ is $(q + pt)^n$ ***R1***

which is a polynomial of degree n ***R1***

METHOD 2

in n independent trials, it is not possible to obtain more than n successes (or equivalent, *eg*, $\mathrm{P}(X > n) = 0$) ***R1***

so $a_r = 0$ for $r > n$ ***R1***

[2 marks]

Examiners report

Explanations in (b) were often poor, again indicating a lack of ability to give a verbal explanation.

12c. Two independent random variables X_1 and X_2 are such that $X_1 \sim \mathrm{B}(1,\ p_1)$ and $X_2 \sim \mathrm{B}(1,\ p_2)$. Prove that if $X_1 + X_2$ has a binomial distribution then $p_1 = p_2$. *[5 marks]*

Markscheme

let $Y = X_1 + X_2$

$G_Y(t) = (q_1 + p_1 t)(q_2 + p_2 t)$ ***A1***

$G_Y(t)$ has degree two, so if Y is binomial then

$Y \sim \mathrm{B}(2,\ p)$ for some p ***R1***

$(q + pt)^2 = (q_1 + p_1 t)(q_2 + p_2 t)$ ***A1***

Note: The LHS could be seen as $q^2 + 2pqt + p^2t^2$.

METHOD 1

by considering the roots of both sides, $\frac{q_1}{p_1} = \frac{q_2}{p_2}$ ***M1***

$\frac{1-p_1}{p_1} = \frac{1-p_2}{p_2}$ ***A1***

so $p_1 = p_2$ ***AG***

METHOD 2

equating coefficients,

$p_1p_2 = p^2,\ q_1q_2 = q^2$ or $(1 - p_1)(1 - p_2) = (1 - p)^2$ ***M1***

expanding,

$p_1 + p_2 = 2p$ so $p_1,\ p_2$ are the roots of $x^2 - 2px + p^2 = 0$ ***A1***

so $p_1 = p_2$ ***AG***

[5 marks]

Total [11 marks]

Examiners report

Very few complete solutions to (c) were seen with few candidates even reaching the result that $(q_1 + p_1 t)(q_2 + p_2 t)$ must equal $(q + pt)^2$ for some p.

A baker produces loaves of bread that he claims weigh on average 800 g each. Many customers believe the average weight of his loaves is less than this. A food inspector visits the bakery and weighs a random sample of 10 loaves, with the following results, in grams:

783, 802, 804, 785, 810, 805, 789, 781, 800, 791.

Assume that these results are taken from a normal distribution.

13a. Determine unbiased estimates for the mean and variance of the distribution. *[3 marks]*

Markscheme

unbiased estimate of the mean: 795 (grams) ***A1***

unbiased estimate of the variance: 108 $(grams^2)$ ***(M1)A1***

[3 marks]

Examiners report

A successful question for many candidates. A few candidates did not read the question and adopted a 2-tailed test.

13b. In spite of these results the baker insists that his claim is correct. *[7 marks]*

Stating appropriate hypotheses, test the baker's claim at the 10 % level of significance.

Markscheme

null hypothesis

$H_0 : \mu = 800$ ***A1***

alternative hypothesis

$H_1 : \mu < 800$ ***A1***

using 1-tailed t-test ***(M1)***

EITHER

$p = 0.0812...$ ***A3***

OR

with 9 degrees of freedom ***(A1)***

$t_{calc} = \frac{\sqrt{10}(795-800)}{\sqrt{108}} = -1.521$ ***A1***

$t_{crit} = -1.383$ ***A1***

Note: Accept 2sf intermediate results.

THEN

so the baker's claim is rejected ***R1***

Note: Accept "reject
H_0 " provided
H_0 has been correctly stated.

Note: ***FT*** for the final ***R1***.

[7 marks]

Examiners report

A successful question for many candidates. A few candidates did not read the question and adopted a 2-tailed test.

The random variable X has a geometric distribution with parameter p .

14a. Show that
$\mathrm{P}(X \leqslant n) = 1 - (1-p)^n$, $n \in \mathbb{Z}^+$. *[3 marks]*

Markscheme

$\mathrm{P}(X \leqslant n) = \sum_{\mathrm{i}=1}^{n} \mathrm{P}(X = \mathrm{i}) = \sum_{\mathrm{i}=1}^{n} pq^{\mathrm{i}-1}$ ***M1A1***

$= p\frac{1-q^n}{1-q}$ ***A1***

$= 1 - (1-p)^n$ ***AG***

[3 marks]

Examiners report

In part (a) some candidates thought that the geometric distribution was continuous, so attempted to integrate the pdf! Others, less seriously, got the end points of the summation wrong.

In part (b) It was very disappointing that may candidates, who got an incorrect answer to part (a), persisted with their incorrect answer into this part.

14b. Deduce an expression for
$\mathrm{P}(m < X \leqslant n)$, m , $n \in \mathbb{Z}^+$ and $m < n$. *[1 mark]*

Markscheme

$(1-p)^m - (1-p)^n$ ***A1***

[1 mark]

Examiners report

In part (a) some candidates thought that the geometric distribution was continuous, so attempted to integrate the pdf! Others, less seriously, got the end points of the summation wrong.
In part (b) It was very disappointing that may candidates, who got an incorrect answer to part (a), persisted with their incorrect answer into this part.

14c. Given that $p = 0.2$, find the least value of n for which $\mathrm{P}(1 < X \leqslant n) > 0.5$, $n \in \mathbb{Z}^+$. *[2 marks]*

Markscheme

attempt to solve
$0.8 - (0.8)^n > 0.5$ ***M1***
obtain $n = 6$ ***A1***
[2 marks]

Examiners report

In part (a) some candidates thought that the geometric distribution was continuous, so attempted to integrate the pdf! Others, less seriously, got the end points of the summation wrong.
In part (b) It was very disappointing that may candidates, who got an incorrect answer to part (a), persisted with their incorrect answer into this part.

The n independent random variables

$X_1, X_2, \ldots, X_n$ all have the distribution

$\mathrm{N}(\mu, \sigma^2)$.

15a. Find the mean and the variance of *[8 marks]*

(i)

$X_1 + X_2$;

(ii)

$3X_1$;

(iii)

$X_1 + X_2 - X_3$;

(iv)

$\bar{X} = \frac{(X_1 + X_2 + \ldots + X_n)}{n}$.

Markscheme

(i)

$2\mu,\ 2\sigma^2$ ***A1A1***

(ii)

$3\mu,\ 9\sigma^2$ ***A1A1***

(iii)

$\mu,\ 3\sigma^2$ ***A1A1***

(iv)

$\mu,\ \frac{\sigma^2}{n}$ ***A1A1***

Note: If candidate clearly and correctly gives the standard deviations rather than the variances, give ***A1*** for 2 or 3 standard deviations and ***A1A1*** for 4 standard deviations.

[8 marks]

Examiners report

This was very well answered indeed with very many candidates gaining full marks including, pleasingly, part (b). Candidates who could not do question 2, struggled on the whole paper.

15b. Find *[3 marks]*

$\mathrm{E}(X_1^2)$ in terms of

μ and

σ .

Markscheme

$\mathrm{Var}(X_1) = \mathrm{E}(X_1^2) - (\mathrm{E}(X_1))^2$ ***(M1)***

$\sigma^2 = \mathrm{E}(X_1^2) - \mu^2$ ***(A1)***

$\mathrm{E}(X_1^2) = \sigma^2 + \mu^2$ ***A1***

[3 marks]

Examiners report

This was very well answered indeed with very many candidates gaining full marks including, pleasingly, part (b). Candidates who could not do question 2, struggled on the whole paper.

μ

$\sigma^2 = \frac{1}{4}\ (\text{metres}^2)$

$\mu = 3$

$\mu < 3$

$\bar{x}$

$\bar{x} < A$

$\mu = 3$

μ

Markscheme

(i)

$H_0 : \mu = 3,\ H_1 : \mu < 3$

1 tailed z test as

σ^2 is known

under

$H_0,\ X \sim \mathrm{N}\left(3, \frac{1}{4}\right)$ so $\bar{X} \sim \mathrm{N}\left(3, \frac{\frac{1}{4}}{36}\right) = N\left(3, \frac{1}{144}\right)$ ***(M1)***

$z = \frac{\bar{x}-3}{\frac{1}{12}}$ is $\mathrm{N}(0, 1)$ ***(A1)***

$\mathrm{P}(z < -1.64485...) = 0.05$ ***(A1)***

so inequality is given by

$\frac{\bar{x}-3}{\frac{1}{12}} < -1.64485...$ giving $\bar{x} < 2.8629\ldots$ ***M1***

$\bar{x} < 2.863$ (4sf) ***A1***

Note: Candidates can get directly to the answer from

$\mathrm{N}\left(3, \frac{1}{144}\right)$ they do not have to go via z is $\mathrm{N}(0, 1)$. However they must give some explanation of what they have done; they cannot just write the answer down.

(ii) a Type I error is accepting

H_1 when

H_0 is true ***A1***

(iii) a Type II error is accepting

H_0 when

H_1 is true ***A1***

(iv) 0.05 ***A1***

Note: Accept anything that rounds to 0.050 if they do the conditional calculation.

(v)

$\bar{X} \sim \mathrm{N}\left(2.75, \frac{1}{144}\right)$ ***(M1)***

$\mathrm{P}(\bar{x} > 2.8629...) = 0.0877$ (3sf) ***(M1)A1***

Note: Accept any answer between 0.0875 and 0.0877 inclusive.

Note: Accept anything that rounded is between 0.087and 0.089 if there is evidence that the candidate has used tables.

[11 marks]

Examiners report

(a) There were many reasonable answers. In (i) not all candidates explained their method so that they could gain good partial marks even if they had the wrong final answer. A common mistake was to give an answer above 3. It was pleasing that almost all candidates had (ii) and (iii) correct, as this had caused problems in the past. In (iv) it was amusing to see a few candidates work out 5% using conditional probability rather than just write down the answer as asked.

(b) It was pleasing that almost all candidates realised that it was a t-test rather than a z-test.

There was good understanding on how to use the calculator in parts (ii) and (iii). The correct confidence interval to the desired accuracy was not always given.

The most common mistake in question 3 was forgetting to take into account the variance of the sample mean.

16b. The random variable Y represents the height of a wave on another surf beach. It is known that Y is normally distributed with unknown mean *[8 marks]*

μ (metres) and unknown variance

σ^2 (metres^2) . David wishes to test the claim made in a surf guide that

$\mu = 3$ against the alternative that

$\mu < 3$. He is also going to perform this test at the 5 % level. He measures the heights of 36 waves and finds that the sample mean,

$\bar{y} = 2.860$ and the unbiased estimate of the population variance,

$s_{n-1}^2 = 0.25$.

(i) State the name of the test that David should perform.

(ii) State the conclusion of David's test, justifying your answer by giving the p-value.

(iii) Using David's results, calculate the 90 % confidence interval for

μ , giving your answers to 4 significant figures.

Markscheme

(i) t-test ***A1***

(ii)

$\mathrm{H}_0 : \mu = 3, \ \mathrm{H}_1 : \mu < 3$

1 tailed t test as

σ^2 is unknown

$t = \frac{\bar{y}-3}{\frac{1}{12}}$ has the t-distribution with

$v = 35$ ***(M1)***

the p-value is $0.0509\ldots$ ***A2***

this is

> 0.05 ***R1***

so we accept that the mean wave height is 3 ***R1***

Note: Allow "Accept

H_0 " provided

H_0 has been stated.

Note: Accept ***FT*** on the p-value for the ***R1***s.

(iii)

$2.719 < \mu < 3.001$ (4 sf) ***A1A1***

Note:

$2.860 \pm 1.6896\ldots \times \frac{\frac{1}{2}}{6}$ would gain ***M1***.

Note: Award ***A1A0*** if answer are only given to 3sf.

[8 marks]

Examiners report

(a) There were many reasonable answers. In (i) not all candidates explained their method so that they could gain good partial marks even if they had the wrong final answer. A common mistake was to give an answer above 3. It was pleasing that almost all candidates had (ii) and (iii) correct, as this had caused problems in the past. In (iv) it was amusing to see a few candidates work out 5% using conditional probability rather than just write down the answer as asked.

(b) It was pleasing that almost all candidates realised that it was a t-test rather than a z-test.

There was good understanding on how to use the calculator in parts (ii) and (iii). The correct confidence interval to the desired accuracy was not always given.

The most common mistake in question 3 was forgetting to take into account the variance of the sample mean.

Jenny and her Dad frequently play a board game. Before she can start Jenny has to throw a "six" on an ordinary six-sided dice. Let the random variable X denote the number of times Jenny has to throw the dice in total until she obtains her first "six".

17a. If the dice is fair, write down the distribution of X , including the value of any parameter(s). *[1 mark]*

Markscheme

$X \sim \text{Geo}\left(\frac{1}{6}\right)$ or $\text{NB}\left(1, \frac{1}{6}\right)$ ***A1***

[1 mark]

Examiners report

This was well answered as the last question should be the most difficult. It seemed accessible to many candidates, if they realised what the distributions were. The goodness of fit test was well used in (c) with hardly any candidates mistakenly combining cells. Part (e) was made more complicated than it needed to be with calculator solutions when a bit of pure maths would have sufficed. Part (f) caused some problems but good candidates did not have too much difficulty.

17b. Write down $\text{E}(X)$ for the distribution in part (a). *[1 mark]*

Markscheme

$\text{E}(X) = 6$ ***A1***

[1 mark]

Examiners report

This was well answered as the last question should be the most difficult. It seemed accessible to many candidates, if they realised what the distributions were. The goodness of fit test was well used in (c) with hardly any candidates mistakenly combining cells. Part (e) was made more complicated than it needed to be with calculator solutions when a bit of pure maths would have sufficed. Part (f) caused some problems but good candidates did not have too much difficulty.

17c. Before Jenny's Dad can start, he has to throw two "sixes" using a fair, ordinary six-sided dice. Let the random variable Y denote the total number of times Jenny's Dad has to throw the dice until he obtains his second "six". *[1 mark]*

Write down the distribution of Y , including the value of any parameter(s).

Markscheme

Y is

$\text{NB}\left(2, \frac{1}{6}\right)$ ***A1***

[1 mark]

Examiners report

This was well answered as the last question should be the most difficult. It seemed accessible to many candidates, if they realised what the distributions were. The goodness of fit test was well used in (c) with hardly any candidates mistakenly combining cells. Part (e) was made more complicated than it needed to be with calculator solutions when a bit of pure maths would have sufficed. Part (f) caused some problems but good candidates did not have too much difficulty.

17d. Before Jenny's Dad can start, he has to throw two "sixes" using a fair, ordinary six-sided dice. Let the random variable Y denote the total number of times Jenny's Dad has to throw the dice until he obtains his second "six". *[1 mark]*

Find the value of y such that

$\text{P}(Y = y) = \frac{1}{36}$.

Markscheme

$\text{P}(Y = y) = \frac{1}{36}$ gives $y = 2$ ***A1***

(as all other probabilities would have a factor of 5 in the numerator)

[1 mark]

Examiners report

This was well answered as the last question should be the most difficult. It seemed accessible to many candidates, if they realised what the distributions were. The goodness of fit test was well used in (c) with hardly any candidates mistakenly combining cells. Part (e) was made more complicated than it needed to be with calculator solutions when a bit of pure maths would have sufficed. Part (f) caused some problems but good candidates did not have too much difficulty.

17e. Before Jenny's Dad can start, he has to throw two "sixes" using a fair, ordinary six-sided dice. Let the random variable Y denote the total number of times Jenny's Dad has to throw the dice until he obtains his second "six". *[2 marks]*

Find

$\text{P}(Y \leqslant 6)$.

Markscheme

$\text{P}(Y \leqslant 6) = \left(\frac{1}{6}\right)^2 + 2\left(\frac{5}{6}\right)\left(\frac{1}{6}\right)^2 + 3\left(\frac{5}{6}\right)^2\left(\frac{1}{6}\right)^2 + 4\left(\frac{5}{6}\right)^3\left(\frac{1}{6}\right)^2 + 5\left(\frac{5}{6}\right)^4\left(\frac{1}{6}\right)^2$ ***(M1)***

$= 0.263$ ***A1***

[2 marks]

Examiners report

This was well answered as the last question should be the most difficult. It seemed accessible to many candidates, if they realised what the distributions were. The goodness of fit test was well used in (c) with hardly any candidates mistakenly combining cells. Part (e) was made more complicated than it needed to be with calculator solutions when a bit of pure maths would have sufficed. Part (f) caused some problems but good candidates did not have too much difficulty.

The random variable X is normally distributed with unknown mean

μ and unknown variance

σ^2. A random sample of 20 observations on X gave the following results.

$$\sum x = 280, \ \sum x^2 = 3977.57$$

18a. Find unbiased estimates of *[3 marks]*

μ and

σ^2.

Markscheme

$\bar{x} = 14$ ***A1***

$s_{n-1}^2 = \frac{3977.57}{19} - \frac{280^2}{380}$ ***(M1)***

$= 3.03$ ***A1***

[3 marks]

Note: Accept any notation for these estimates including

μ and

σ^2.

Note: Award ***M0A0*** for division by 20.

Examiners report

In (a), most candidates estimated the mean correctly although many candidates failed to obtain a correct unbiased estimate for the variance. The most common error was to divide
$\sum x^2$ by
20 instead of
19. For some candidates, this was not a costly error since we followed through their variance into (b) and (c).

18b. Determine a 95 % confidence interval for *[3 marks]*

μ.

Markscheme

the 95% confidence limits are

$\bar{x} \pm t\sqrt{\frac{s_{n-1}^2}{n}}$ ***(M1)***

Note: Award ***M0*** for use of *z*.

ie,

$14 \pm 2.093\sqrt{\frac{3.03}{20}}$ ***(A1)***

Note:*FT* their mean and variance from (a).

giving [13.2, 14.8] ***A1***

Note: Accept any answers which round to 13.2 and 14.8.

[3 marks]

Examiners report

In (b) and (c), since the variance was estimated, the confidence interval and test should have been carried out using the t-distribution. It was extremely disappointing to note that many candidates found a Z-interval and used a Z-test and no marks were awarded for doing this. Candidates should be aware that having to estimate the variance is a signpost pointing towards the t-distribution.

18c. Given the hypotheses *[4 marks]*

$$\mathrm{H}_0 : \mu = 15;\ \mathrm{H}_1 : \mu \neq 15,$$

find the *p*-value of the above results and state your conclusion at the 1 % significance level.

Markscheme

Use of t-statistic

$\left(= \frac{14-15}{\sqrt{\frac{3.03}{20}}}\right)$ ***(M1)***

Note: ***FT*** their mean and variance from (a).

Note: Award ***M0*** for use of *z*.

Note: Accept

$\frac{15-14}{\sqrt{\frac{3.03}{20}}}$.

$= -2.569\ldots$ ***(A1)***

Note: Accept

$2.569\ldots$

p - value $= 0.009392\ldots \times 2 = 0.0188$ ***A1***

Note: Accept any answer that rounds to 0.019.

Note: Award ***(M1)(A1)A0*** for any answer that rounds to 0.0094.

insufficient evidence to reject

H_0 (or equivalent, *eg* accept

H_0 or reject

H_1) ***R1***

Note: ***FT*** on their *p*-value.

[4 marks]

Examiners report

In (b) and (c), since the variance was estimated, the confidence interval and test should have been carried out using the t-distribution. It was extremely disappointing to note that many candidates found a Z-interval and used a Z-test and no marks were awarded for doing this. Candidates should be aware that having to estimate the variance is a signpost pointing towards the t-distribution.

The number of machine breakdowns occurring in a day in a certain factory may be assumed to follow a Poisson distribution with mean

μ. The value of

μ is known, from past experience, to be 1.2. In an attempt to reduce the value of

μ, all the machines are fitted with new control units. To investigate whether or not this reduces the value of

μ, the total number of breakdowns, x, occurring during a 30-day period following the installation of these new units is recorded.

19a. State suitable hypotheses for this investigation. *[1 mark]*

Markscheme

$\mathrm{H_0}: \mu = 1.2$;

$\mathrm{H_1}: \mu < 1.2$ ***A1***

Note: Accept "

$\mathrm{H_0}:$ (

30-day) mean

$= 36$;

$\mathrm{H_1}:$ (

30-day) mean

$= 36$ ".

[1 mark]

Examiners report

This question was well answered by many candidates. The most common error was to attempt to use a normal approximation to find approximate probabilities instead of the Poisson distribution to find the exact probabilities. Some candidates appeared not to be familiar with the term 'Type II error probability' which made (b)(ii) inaccessible. Another fairly common error was to believe that the complement of
$x \leqslant 25$ is
$x \geqslant 25$.

19b. It is decided to define the critical region by *[8 marks]*

$x \leqslant 25$.

(i) Calculate the significance level.

(ii) Assuming that the value of

μ was actually reduced to 0.75, determine the probability of a Type II error.

Markscheme

(i) let X denote the number of breakdowns in 30 days

then under
H_0 ,
$E(X) = 36$ ***(A1)***

sig level $= \mathrm{P}(X \leqslant 25|\text{mean} = 36)$ ***(M1)(A1)***

= 0.0345 (3.45%) ***A1***

Note: Accept any answer that rounds to 0.035 (3.5%) .

Note: Do not accept the use of a normal approximation.

(ii) under
H_1,
$E(X) = 22.5$ ***(A1)***

$P(\text{Type II error}) = P(X \geqslant 26|\text{mean} = 22.5)$ ***(M1)(A1)***
= 0.257 ***A1***

Note: Accept any answer that rounds to 0.26.

Note: Do not accept the use of a normal approximation.

[8 marks]

Examiners report

This question was well answered by many candidates. The most common error was to attempt to use a normal approximation to find approximate probabilities instead of the Poisson distribution to find the exact probabilities. Some candidates appeared not to be familiar with the term 'Type II error probability' which made (b)(ii) inaccessible. Another fairly common error was to believe that the complement of
$x \leqslant 25$ is
$x \geqslant 25$.

The continuous random variable X has probability density function f given by

$$f(x) = \begin{cases} \frac{3x^2+2x}{10}, & \text{for } 1 \leqslant x \leqslant 2 \\ 0, & \text{otherwise.} \end{cases}$$

20a. (i) Determine an expression for *[6 marks]*

$F(x)$, valid for

$1 \leqslant x \leqslant 2$, where F denotes the cumulative distribution function of X.

(ii) Hence, or otherwise, determine the median of X.

Markscheme

(i)

$F(x) = \int_1^x \frac{3u^2+2u}{10}\mathrm{d}u$ ***(M1)***

$= \left[\frac{u^3+u^2}{10}\right]_1^x$ ***A1***

Note: Do not penalise missing or wrong limits at this stage.

Accept the use of *x* in the integrand.

$= \frac{x^3+x^2-2}{10}$ ***A1***

(ii) the median m satisfies the equation

$F(m) = \frac{1}{2}$ so ***(M1)***

$m^3 + m^2 - 7 = 0$ ***(A1)***

Note: Do not ***FT*** from an incorrect

$F(x)$.

$m = 1.63$ ***A1***

Note: Accept any answer that rounds to 1.6.

[6 marks]

Examiners report

Solutions to (a)(i) were disappointing in general, suggesting that many candidates are unfamiliar with the concept of the cumulative distribution function. Many candidates knew that it was something to do with the integral of the probability density function but some thought it was

$\int_1^2 f(x)\mathrm{d}x$ which they then evaluated as

1 while others thought it was just

$\int f(x)\mathrm{d}x = \frac{(x^2+x^3)}{10}$ which is not, in general, a valid method. However, most candidates solved (a)(ii) correctly, usually by integrating the probability density function from

1 to

m.

20b. (i) State the central limit theorem. *[8 marks]*

(ii) A random sample of 150 observations is taken from the distribution of X and

$\bar{X}$ denotes the sample mean. Use the central limit theorem to find, approximately, the probability that

$\bar{X}$ is greater than 1.6.

Markscheme

(i) the mean of a large sample from any distribution is approximately

normal ***A1***

Note: This is the minimum acceptable explanation.

(ii) we require the mean

μ and variance

σ^2 of X

$\mu = \int_1^2 \left(\frac{3x^3+2x^2}{10}\right) \mathrm{d}x$ ***(M1)***

$= \frac{191}{120}$ $(1.591666\ldots)$ ***A1***

$\sigma^2 = \int_1^2 \left(\frac{3x^4+2x^3}{10}\right) \mathrm{d}x - \mu^2$ ***(M1)***

$= 0.07659722\ldots$ ***A1***

the central limit theorem states that

$\bar{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right)$, *i.e.*

$N(1.591666\ldots, 0.0005106481\ldots)$ ***M1A1***

$\mathrm{P}(\bar{X} > 1.6) = 0.356$ ***A1***

Note: Accept any answer that rounds to 0.36.

[8 marks]

Examiners report

In (b)(i), the statement of the central limit theorem was often quite dreadful. The term ‘sample mean’ was often not mentioned and a common misconception appears to be that the actual distribution rather than the sample mean tends to normality as the sample size increases. Solutions to (b)(ii) often failed to go beyond finding the mean and variance of X. In calculating the variance, some candidates rounded the mean from $1.5916666..$ to 1.59 which resulted in an incorrect value for the variance. It is important to note that calculating a variance usually involves a small difference of two large numbers so that full accuracy must be maintained.

When Ben shoots an arrow, he hits the target with probability 0.4. Successive shots are independent.

21a. Find the probability that *[6 marks]*

(i) he hits the target exactly 4 times in his first 8 shots;

(ii) he hits the target for the 4^{th} time with his 8^{th} shot.

Markscheme

(i) the number of hits,

$X \sim \mathrm{B}(8, 0.4)$ ***(A1)***

$P(X = 4) = \binom{8}{4} \times 0.4^4 \times 0.6^4$ ***(M1)***

= 0.232 ***A1***

Note: Accept any answer that rounds to 0.23.

(ii) let the

4^{th} hit occur on the

Y^{th} shot so that

$Y \sim \mathrm{NB}(4, 0.4)$ ***(A1)***

$P(Y = 8) = \binom{7}{3} \times 0.4^4 \times 0.6^4$ ***(M1)***

= 0.116 ***A1***

Note: Accept any answer that rounds to 0.12.

[6 marks]

Examiners report

Part (a) was well answered in general although some candidates were unable to distinguish between the binomial and negative binomial distributions.

21b. Ben hits the target for the *[9 marks]*

10^{th} time with his

X^{th} shot.

(i) Determine the expected value of the random variable X.

(ii) Write down an expression for

$\mathrm{P}(X = x)$ and show that

$$\frac{\mathrm{P}(X = x)}{\mathrm{P}(X = x - 1)} = \frac{3(x - 1)}{5(x - 10)}.$$

(iii) Hence, or otherwise, find the most likely value of X.

Markscheme

(i)

$X \sim \mathrm{NB}(10,\, 0.4)$ ***(M1)***

$\mathrm{E}(X) = \frac{10}{0.4} = 25$ ***A1***

(ii) let

P_x denote

$\mathrm{P}(X = x)$

$P_x = \binom{x-1}{9} \times 0.4^{10} \times 0.6^{x-10}$ ***A1***

$\frac{P_x}{P_{x-1}} = \frac{\binom{x-1}{9} \times 0.4^{10} \times 0.6^{x-10}}{\binom{x-2}{9} \times 0.4^{10} \times 0.6^{x-11}}$ ***M1A1***

$= \frac{(x-1)!}{9!(x-10)!} \times \frac{9!(x-11)! \times 0.6}{(x-2)!}$ ***A1***

Note: Award ***A1*** for correct evaluation of combinatorial terms.

$= \frac{3(x-1)}{5(x-10)}$ ***AG***

(iii)

$\mathrm{P}_x > \mathrm{P}_{x-1}$ as long as

$3x - 3 > 5x - 50$ ***(M1)***

i.e.

$x < 23.5$ ***(A1)***

the most likely value is 23 ***A1***

Note: Allow solutions based on creating a table of values of

P_x.

[9 marks]

Examiners report

In (b)(ii), most candidates knew what to do but algebraic errors were not uncommon. Candidates often used equal instead of inequality signs and this was accepted if it led to
$x = 23.5$. The difficulty for these candidates was whether to choose
23 or
24 for the final answer and some made the wrong choice. Some candidates failed to see the relevance of the result in (b)(ii) to finding the most likely value of
X and chose an ‘otherwise’ method, usually by creating a table of probabilities and selecting the largest.

22. (a) Consider the random variable X for which $\mathrm{E}(X) = a\lambda + b$, where a and b are constants and λ is a parameter. Show that $\frac{X-b}{a}$ is an unbiased estimator for λ. *[14 marks]*

(b) The continuous random variable Y has probability density function

$$f(y) = \begin{cases} \frac{2}{9}(3 + y - \lambda), & \text{for } \lambda - 3 \le y \le \lambda \\ 0, & \text{otherwise} \end{cases}$$

where λ is a parameter.

(i) Verify that $f(y)$ is a probability density function for all values of λ.

(ii) Determine $\mathrm{E}(Y)$.

(iii) Write down an unbiased estimator for λ.

Markscheme

(a)

$\mathrm{E}\left(\frac{X-b}{a}\right) = \frac{a\lambda+b-b}{a}$ ***M1A1***

$= \lambda$ ***A1***

(Therefore
$\frac{X-b}{a}$ is an unbiased estimator for
λ) ***AG***

[3 marks]

(b) (i)

$f(y) \geqslant 0$ ***R1***

Note: Only award ***R1*** if this statement is made explicitly.

recognition or showing that integral of f is 1 (seen anywhere) ***R1***

EITHER

$\int_{\lambda-3}^{\lambda} \frac{2}{9}(3+y-\lambda)\mathrm{d}y$ ***M1***

$= \frac{2}{9}\left[(3-\lambda)y+\frac{1}{2}y^2\right]_{\lambda-3}^{\lambda}$ ***A1***

$= \frac{2}{9}\left(\lambda(3-\lambda)+\frac{1}{2}\lambda^2-(3-\lambda)(\lambda-3)-\frac{1}{2}(\lambda-3)^2\right)$ or equivalent ***A1***

$= 1$

OR

the graph of the probability density is a triangle with base length 3 and height
$\frac{2}{3}$ ***M1A1***

its area is therefore
$\frac{1}{2}\times 3\times\frac{2}{3}$ ***A1***

$= 1$

(ii)

$\mathrm{E}(Y) = \int_{\lambda-3}^{\lambda} \frac{2}{9}y(3+y-\lambda)\mathrm{d}y$ ***M1***

$= \frac{2}{9}\left[(3-\lambda)\frac{1}{2}y^2+\frac{1}{3}y^3\right]_{\lambda-3}^{\lambda}$ ***A1***

$= \frac{2}{9}\left((3-\lambda)\frac{1}{2}\left(\lambda^2-(\lambda-3)^2\right)+\frac{1}{3}\left(\lambda^3-(\lambda-3)^3\right)\right)$ ***M1***

$= \lambda - 1$ ***A1A1***

Note: Award 3 marks for noting that the mean is
$\frac{2}{3}$rds the way along the base and then ***A1A1*** for
$\lambda - 1$.

Note: Award ***A1*** for
λ and ***A1*** for –1.

(iii) unbiased estimator:
$Y+1$ ***A1***

Note: Accept

$\bar{Y}+1$.

Follow through their

$\mathrm{E}(Y)$ if linear.

[11 marks]

Total [14 marks]

Examiners report

[N/A]

23. Consider the random variable

[16 marks]

$X \sim \mathrm{Geo}(p)$.

(a) State

$\mathrm{P}(X<4)$.

(b) Show that the probability generating function for X is given by

$G_X(t)=\frac{pt}{1-qt}$, where

$q=1-p$.

Let the random variable

$Y=2X$.

(c) (i) Show that the probability generating function for Y is given by

$G_Y(t)=G_X(t^2)$.

(ii) By considering

$G_Y'(1)$, show that

$\mathrm{E}(Y)=2\mathrm{E}(X)$.

Let the random variable

$W=2X+1$.

(d) (i) Find the probability generating function for W in terms of the probability generating function of Y.

(ii) Hence, show that

$\mathrm{E}(W)=2\mathrm{E}(X)+1$.

Markscheme

(a) use of

$\mathrm{P}(X = n) = pq^{n-1}$ $(q = 1 - p)$ ***(M1)***

$\mathrm{P}(X < 4) = p + pq + pq^2$ $\left(= 1 - q^3\right)$ $\left(= 1 - (1-p)^3\right)$ $\left(= 3p - 3p^2 + p^3\right)$ ***A1***

[2 marks]

(b)

$G_X(t) = \mathrm{P}(X = 1)t + \mathrm{P}(X = 2)t^2 + \ldots$ ***(M1)***

$= pt + pqt^2 + pq^2t^3 + \ldots$ ***A1***

summing an infinite geometric series ***M1***

$= \frac{pt}{1-qt}$ ***AG***

[3 marks]

(c) (i) **EITHER**

$G_Y(t) = \mathrm{P}(Y = 1)t + \mathrm{P}(Y = 2)t^2 + \ldots$ ***A1***

$= 0 \times t + \mathrm{P}(X = 1)t^2 + 0 \times t^3 + \mathrm{P}(X = 2)t^4 + \ldots$ ***M1A1***

$= G_X(t^2)$ ***AG***

OR

$G_Y(t) = E(t^Y) = E(t^{2X})$ ***M1A1***

$= E\left((t^2)^X\right)$ ***A1***

$= G_X(t^2)$ ***AG***

(ii)

$\mathrm{E}(Y) = G'_Y(1)$ ***A1***

EITHER

$= 2tG'_X(t^2)$ evaluated at

$t = 1$ ***M1A1***

$= 2\mathrm{E}(X)$ ***AG***

OR

$= \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{pt^2}{(1-qt^2)}\right) = \frac{2pt(1-qt^2)+2pqt^3}{(1-qt^2)^2}$ evaluated at

$t = 1$ ***A1***

$= 2 \times \frac{p(1-qt)+pqt}{(1-qt)^2}$ evaluated at

$t = 1$ (or $\frac{2}{p}$) ***A1***

$= 2\mathrm{E}(X)$ ***AG***

[6 marks]

(d) (i)

$G_W(t) = tG_Y(t)$ (or equivalent) ***A2***

(ii) attempt to evaluate

$G'_W(t)$ ***M1***

EITHER

obtain

$1 \times G_Y(t) + t \times G'_Y(t)$ ***A1***

substitute

$t = 1$ to obtain

$1 \times 1 + 1 \times G'_Y(1)$ ***A1***

OR

$= \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{pt^3}{(1-qt^2)}\right) = \frac{3pt^2(1-qt^2)+2pqt^4}{(1-qt^2)^2}$ ***A1***

substitute

$t = 1$ to obtain

$1 + \frac{2}{p}$ ***A1***

$= 1 + 2\mathrm{E}(X)$ ***AG***

[5 marks]

Total [16 marks]

Examiners report

[N/A]

A traffic radar records the speed,
v kilometres per hour (
$\mathrm{km\,h^{-1}}$), of cars on a section of a road.
The following table shows a summary of the results for a random sample of 1000 cars whose speeds were recorded on a given day.

Speed	Number of cars
$50 \le v < 60$	5
$60 \le v < 70$	13
$70 \le v < 80$	52
$80 \le v < 90$	68
$90 \le v < 100$	98
$100 \le v < 110$	105
$110 \le v < 120$	289
$120 \le v < 130$	142
$130 \le v < 140$	197
$140 \le v < 150$	31

24a. Using the data in the table, *[4 marks]*

(i) show that an estimate of the mean speed of the sample is 113.21
$\mathrm{km\,h^{-1}}$;

(ii) find an estimate of the variance of the speed of the cars on this section of the road.

Markscheme

(i)

$\bar{v} = \frac{1}{1000}(55 \times 5 + 65 \times 13 + \ldots + 145 \times 31)$ ***A1M1***

Note: ***A1*** for mid-points, ***M1*** for use of the formula.

$= \frac{113\,210}{1000} = 113.21$ ***AG***

(ii)

$s^2 = \frac{(55-113.21)^2 \times 5 + (65-113.21)^2 \times 13 + \ldots + (145-113.21)^2 \times 31}{999}$ ***(M1)***

$= \frac{362\,295.9}{999} = 362.6585\ldots = 363$ ***A1***

Note: Award ***A1*** if answer rounds to 362 or 363.

Note: Condone division by 1000.

[4 marks]

Examiners report

In (a)(i), the candidates were required to show that the estimate of the mean is 113.21 so that those who stated simply 'Using my GDC, mean = 113.21' were given no credit. Candidates were expected to indicate that the interval midpoints were used and to show the appropriate formula. In (a)(ii), division by either 999 or 1000 was accepted, partly because of the large sample size and partly because the question did not ask for an unbiased estimate of variance.

24b. Find the 95% confidence interval, I, for the mean speed. *[2 marks]*

Markscheme

$\bar{v} \pm \frac{t_{0.025} \times s}{\sqrt{n}}$ ***(M1)***

hence the confidence interval

$I = [112.028,\ 114.392]$ ***A1***

Note: Accept answers which round to 112 and 114.

Note: Condone the use of $z_{0.025}$ for $t_{0.025}$ and σ for s.

[2 marks]

Examiners report

24c. Let J be the 90% confidence interval for the mean speed. *[2 marks]*

Without calculating J, explain why $J \subset I$.

Markscheme

less confidence implies narrower interval ***R2***

Note: Accept equivalent statements or arguments having a meaningful diagram and/or relevant percentiles.

hence the confidence interval I at the 95% level contains the confidence interval J at the 90% level ***AG***

[2 marks]

Examiners report

Solutions to (c) were often badly written, often quite difficult to understand exactly what was being stated.

Jenny tosses seven coins simultaneously and counts the number of tails obtained. She repeats the experiment 750 times. The following frequency table shows her results.

Number of tails	Frequency
0	6
1	19
2	141
3	218
4	203
5	117
6	38
7	8

25a. Explain what can be done with this data to decrease the probability of making a type I error. *[2 marks]*

Markscheme

reduce the significance level (or equivalent statement) ***R2***

[2 marks]

Examiners report

It was disappointing to see that some candidates wrote incorrect hypotheses, eg ‘
H_0: Data are binomial;
H_1: Data are not binomial’ without specifying any parameters. Part (b) caused unexpected problems for many candidates who misunderstood the question and gave ‘increase the number of trials’ as their answer.

25b. (i) State the meaning of a type II error. *[2 marks]*

(ii) Write down how to proceed if it is required to decrease the probability of making both a type I and type II error.

Markscheme

(i) accepting

H_0 (or failing to reject

H_0) when it is false (or equivalent) ***A1***

(ii) increase the number of trials ***A1***

[2 marks]

Examiners report

26. Francisco and his friends want to test whether performance in running 400 metres improves if they follow a particular training schedule. The competitors are tested before and after the training schedule. *[10 marks]*

The times taken to run 400 metres, in seconds, before and after training are shown in the following table.

Competitor	**A**	**B**	**C**	**D**	**E**
Time before training	75	74	60	69	69
Time after training	73	69	55	72	65

Apply an appropriate test at the 1% significance level to decide whether the training schedule improves competitors' times, stating clearly the null and alternative hypotheses. (It may be assumed that the distributions of the times before and after training are normal.)

Markscheme

H_0: the training schedule does not help improve times (or $\mu = 0$) ***A1***

H_1: the training schedule does help improve times (or $\mu > 0$) ***A1***

Note: Subsequent marks can be awarded even if the hypotheses are not stated.

(Assuming difference of times is normally distributed.)

let d time before training $-$ time after training ***(M1)***

Competitor	**A**	**B**	**C**	**D**	**E**
Time before training (in seconds)	75	74	60	69	69
Time after training (in seconds)	73	69	55	72	65
Difference *d*	2	5	5	–3	4

EITHER

$n = 5,\ \sum d = 13,\ \sum d^2 = 79 \Rightarrow s_{n-1}^2 = \frac{1}{4}\left(79 - \frac{169}{5}\right) = 11.3$ ***(M1)***

(small sample) so use a one-sided t-test ***(M1)***

Note: The “one-sided” t-test may have been seen above when stating H_1.

$t = \frac{2.6}{\sqrt{\frac{11.3}{5}}} = 1.7\ldots$ ***A1***

$v = 4,$ ***A1***

at the 1% level the critical value is 3.7 ***A1***

since $3.7 > 1.7\ldots$ H_0 is accepted (insufficient evidence to reject H_0) ***R1***

Note: Follow through their t-value.

OR

(small sample) so use a one-sided t-test ***(M1)***

$p = 0.079\ldots$ ***A4***

since $0.079\ldots > 0.01$ H_0 is accepted (insufficient evidence to reject H_0) ***R1***

Note: Follow through their p-value.

Note: Accept $d =$ time after training $-$ time before training throughout.

[10 marks]

Examiners report

It was again disappointing to see many candidates giving incorrect hypotheses. A common error was to give the hypotheses the wrong way around. Candidates should be aware that in this type of problem the null hypothesis always represents the status quo. Also, some candidates defined ‘
$d = \text{time before } - \text{ time after}$’ and then gave the hypotheses incorrectly as
$\mathrm{H}_0 : d = 0$ or
$\bar{d} = 0$; $\mathrm{H}_1 : d > 0$ or
$\bar{d} > 0$. It is important to note that the parameter being tested here is
$E(d)$ or
μ_d although
μ was accepted.

27. Let *[14 marks]*
X and
Y be independent random variables with
$X \sim P_o\ (3)$ and
$Y \sim P_o\ (2)$.
Let
$S = 2X + 3Y$.

(a) Find the mean and variance of
S.

(b) Hence state with a reason whether or not
S follows a Poisson distribution.

Let
$T = X + Y$.

(c) Find
$\mathrm{P}(T = 3)$.

(d) Show that
$\mathrm{P}(T = t) = \sum\limits_{r=0}^{t} \mathrm{P}(X = r)\mathrm{P}(Y = t - r)$.

(e) Hence show that
T follows a Poisson distribution with mean 5.

Markscheme

(a)

$\mathrm{E}(S) = 2\mathrm{E}(X) + 3\mathrm{E}(Y) = 6 + 6 = 12$ ***A1***

$\mathrm{Var}(S) = 4\mathrm{Var}(X) + 9\mathrm{Var}(Y) = 12 + 18 = 30$ ***A1***

[2 marks]

(b)

S does not have a Poisson distribution ***A1***

because

$\mathrm{Var}(S) \neq \mathrm{E}(S)$ ***R1***

Note: Follow through their
$\mathrm{E}(S)$ and
$\mathrm{Var}(S)$ if different.

[2 marks]

(c) **EITHER**

$\mathrm{P}(T=3) = \mathrm{P}\left((X,\, Y) = (3,\, 0)\right) + \mathrm{P}\left((X,\, Y) = (2,\, 1)\right) +$

$+\mathrm{P}\left((X,\, Y) = (1,\, 2)\right) + \mathrm{P}\left((X,\, Y) = (0,\, 3)\right)$ ***(M1)***

$= \mathrm{P}(X=3)\mathrm{P}(Y=0) + \mathrm{P}(X=2)\mathrm{P}(Y=1) +$

$+\mathrm{P}(X=1)\mathrm{P}(Y=2) + \mathrm{P}(X=0)\mathrm{P}(Y=3)$ ***(M1)***

$= \frac{125e^{-5}}{6}\ (= 0.140)$ ***A2***

Note: Accept answers which round to 0.14.

OR

T is

$\mathrm{P}_o(2+3) = \mathrm{P}_o(5)$ ***(M1)(A1)***

$\mathrm{P}(T=3) = \frac{125e^{-5}}{6}\ (= 0.140)$ ***A2***

Note: Accept answers which round to 0.14.

[4 marks]

(d)

$\mathrm{P}(T=t) = \mathrm{P}\left((X,\, Y) = (0,\, t)\right) + \mathrm{P}\left((X,\, Y) = (1,\, t-1)\right) + \ldots \mathrm{P}\left((X,\, Y) = (t,\, 0)\right)$ ***(M1)***

$= \mathrm{P}(X=0)\mathrm{P}(Y=t) + \mathrm{P}(X=1)\mathrm{P}(Y=t-1) + \ldots + \mathrm{P}(X=t)\mathrm{P}(Y=0)$ ***A1***

$= \sum\limits_{r=0}^{t} \mathrm{P}(X=r)\mathrm{P}(Y=t-r)$ ***AG***

[2 marks]

(e)

$\mathrm{P}(T=t) = \sum\limits_{r=0}^{t} \mathrm{P}(X=r)\mathrm{P}(Y=t-r)$

$= \sum\limits_{r=0}^{t} \frac{e^{-3}3^r}{r!} \times \frac{e^{-2}2^{t-r}}{(t-r)!}$ ***M1A1***

$= \frac{e^{-5}}{t!} \sum\limits_{r=0}^{t} \frac{t!}{r!(t-r)!} \times 3^r 2^{t-r}$ ***M1***

$= \frac{e^{-5}}{t!}(3+2)^t$ ***A1***

$\left(= \frac{e^{-5}5^t}{t!}\right)$

hence

T follows a Poisson distribution with mean 5 ***AG***

[4 marks]

Examiners report

Parts (a) and (b) were well answered by most candidates. The most common error in (a) was to calculate $E(2X + 3Y)$ correctly as 12 and then state that, because the sum is Poisson, the variance is also 12. Many of these candidates then stated in (b) that the sum is Poisson because the mean and variance are equal, without apparently realising the circularity of their argument. Although (c) was intended as a possible hint for solving (d) and (e), many candidates simply noted that $X + Y$ is $\mathrm{P}_o(5)$ which led immediately to the correct answer. Some candidates tended to merge (d) and (e), often unsuccessfully, while very few candidates completed (e) correctly where the need to insert $t!$ in the numerator and denominator was not usually spotted.

The discrete random variable X has the following probability distribution, where $0 < \theta < \frac{1}{3}$.

x	1	2	3
$\mathrm{P}(X = x)$	θ	2θ	$1 - 3\theta$

28a. Determine $\mathrm{E}(X)$ and show that $\mathrm{Var}(X) = 6\theta - 16\theta^2$. *[4 marks]*

Markscheme

$\mathrm{E}(X) = 1 \times \theta + 2 \times 2\theta + 3(1 - 3\theta) = 3 - 4\theta$ ***M1A1***

$\mathrm{Var}(X) = 1 \times \theta + 4 \times 2\theta + 9(1 - 3\theta) - (3 - 4\theta)^2$ ***M1A1***

$= 6\theta - 16\theta^2$ ***AG***

[4 marks]

Examiners report

[N/A]

28b. In order to estimate [10 marks]

θ, a random sample of n observations is obtained from the distribution of X .

(i) Given that

$\bar{X}$ denotes the mean of this sample, show that

$$\hat{\theta}_1 = \frac{3 - \bar{X}}{4}$$

is an unbiased estimator for

θ and write down an expression for the variance of

$\hat{\theta}_1$ in terms of n and

θ.

(ii) Let Y denote the number of observations that are equal to 1 in the sample. Show that Y has the binomial distribution

$\mathrm{B}(n,\ \theta)$ and deduce that

$\hat{\theta}_2 = \frac{Y}{n}$ is another unbiased estimator for

θ. Obtain an expression for the variance of

$\hat{\theta}_2$.

(iii) Show that

$\mathrm{Var}(\hat{\theta}_1) < \mathrm{Var}(\hat{\theta}_2)$ and state, with a reason, which is the more efficient estimator,

$\hat{\theta}_1$ or

$\hat{\theta}_2$.

Markscheme

(i)

$\mathrm{E}(\hat{\theta}_1) = \frac{3-\mathrm{E}(\bar{X})}{4} = \frac{3-(3-4\theta)}{4} = \theta$ ***M1A1***

so

$\hat{\theta}_1$ is an unbiased estimator of

θ ***AG***

$\mathrm{Var}(\hat{\theta}_1) = \frac{6\theta-16\theta^2}{16n}$ ***A1***

(ii) each of the n observed values has a probability

θ of having the value 1 ***R1***

so

$Y \sim \mathrm{B}(n,\ \theta)$ ***AG***

$\mathrm{E}(\hat{\theta}_2) = \frac{\mathrm{E}(Y)}{n} = \frac{n\theta}{n} = \theta$ ***A1***

$\mathrm{Var}(\hat{\theta}_2) = \frac{n\theta(1-\theta)}{n^2} = \frac{\theta(1-\theta)}{n}$ ***M1A1***

(iii)

$\mathrm{Var}(\hat{\theta}_1) - \mathrm{Var}(\hat{\theta}_2) = \frac{6\theta-16\theta^2-16\theta+16\theta^2}{16n}$ ***M1***

$= \frac{-10\theta}{16n} < 0$ ***A1***

$\hat{\theta}_1$ is the more efficient estimator since it has the smaller variance ***R1***

[10 marks]

Examiners report

[N/A]