Chi-Squared Test

Questions

Q1.

In an experiment, James flips a coin 3 times and records the number of heads. He carries out the experiment 100 times with his left hand and 100 times with his right hand.

	Number of heads			
	0	**1**	**2**	**3**
Left hand	7	29	42	22
Right hand	13	35	36	16

(a) Test, at the 5% level of significance, whether or not there is an association between the hand he flips the coin with and the number of heads.
You should state your hypotheses, the degrees of freedom and the critical value used for this test.

(7)

(b) Assuming the coin is unbiased, write down the distribution of the number of heads in 3 flips.

(1)

(c) Carry out a χ^2 test, at the 10% level of significance, to test whether or not the distribution you wrote down in part (b) is a suitable model for the number of heads obtained in the 200 trials of James' experiment.
You should state your hypotheses, the degrees of freedom and the critical value used for this test.

(7)

(Total for question = 15 marks)

Q2.

A factory produces pins.

An engineer selects 40 independent random samples of 6 pins produced at the factory and records the number of defective pins in each sample.

Number of defective pins	0	1	2	3	4	5	6
Observed frequency	19	11	7	2	0	1	0

(a) Show that the proportion of defective pins in the 40 samples is 0.15

(2)

The engineer suggests that the number of defective pins in a sample of 6 can be modelled using a binomial distribution. Using the information from the sample above, a test is to be carried out at the 10% significance level, to see whether the data are consistent with the engineer's suggested model.

The value of the test statistic for this test is 2.689

(b) Justifying the degrees of freedom used, carry out the test, at the 10% significance level, to see whether the data are consistent with the engineer's suggested model. State your hypotheses clearly.

(8)

The engineer later discovers that the previously recorded information was incorrect. The data should have been as follows.

Number of defective pins	0	1	2	3	4	5	6
Observed frequency	19	11	6	3	1	0	0

(c) Describe the effect this would have on the value of the test statistic that should be used for the hypothesis test.

Give reasons for your answer.

(3)

(Total for question = 13 marks)

Q3.

Charlie carried out a survey on the main type of investment people have.
The contingency table below shows the results of a survey of a random sample of people.

		Main type of investment		
		Bonds	Cash	Stocks
Age	25–44	a	$b-e$	e
	45–75	c	$d-59$	59

(a) Find an expression, in terms of *a*, *b*, *c* and *d*, for the difference between the observed and the expected value (O – E) for the group whose main type of investment is Bonds and are aged 45 – 75
Express your answer as a single fraction in its simplest form.

(4)

Given that $\sum \frac{(O-E)^2}{E} = 9.62$ for this information,

(b) test, at the 5% level of significance, whether or not there is evidence of an association between the age of a person and the main type of investment they have.
You should state your hypotheses, critical value and conclusion clearly.
You may assume that no cells need to be combined.

(3)

(Total for question = 7 marks)

Q4.

In a game, a coin is spun 5 times and the number of heads obtained is recorded.
Tao suggests playing the game 20 times and carrying out a chi-squared test to investigate whether the coin might be biased.

(a) Explain why playing the game only 20 times may cause problems when carrying out the test.

(1)

Chris decides to play the game 500 times. The results are as follows

Number of heads	0	1	2	3	4	5
Observed frequency	2	27	93	181	146	51

Chris decides to test whether or not the data can be modelled by a binomial distribution, with the probability of a head on each spin being 0.6

She calculates the expected frequencies, to 2 decimal places, as follows

Number of heads	0	1	2	3	4	5
Expected frequency	5.12	38.40	115.20	172.80	129.60	38.88

(b) State the number of degrees of freedom in Chris' test, giving a reason for your answer.

(1)

(c) Carry out the test at the 5% level of significance.
You should state your hypotheses, test statistic, critical value and conclusion clearly.

(5)

(d) Showing your working, find an alternative model which would better fit Chris' data.

(2)

(Total for question = 9 marks)

Q5.

A psychologist carries out a survey of the perceived body weight of 150 randomly chosen people. He asks them if they think they are underweight, about right or overweight. His results are summarised in the table below.

	Underweight	About right	Overweight
Male	20	22	30
Female	16	28	34

The psychologist calculates two of the expected frequencies, to 2 decimal places, for a test of independence between perceived body weight and gender. These results are shown in the table below.

	Underweight	About right	Overweight
Male	17.28		
Female	18.72		

(a) Complete the table of expected frequencies shown above.

(2)

(b) Test, at the 10% level of significance, whether or not perceived body weight is independent of gender. State your hypotheses clearly.

(7)

The psychologist now combines the male and female data to test whether or not body weight types are chosen equally.

(c) Find the smallest significance level, from the tables in the formula booklet, for which there is evidence of a preference.

(5)

(Total for question = 14 marks)

Q6.

Stuart is investigating a treatment for a disease that affects fruit trees. He has 400 fruit trees and applies the treatment to a random sample of these trees. The remainder of the trees have no treatment. He records the number of years, y, that each fruit tree remains free from this disease.

The results are summarised in the table below.

		Treatment	
		Applied	**Not applied**
Number of years free from this disease	$y < 1$	15	25
	$1 \leqslant y < 2$	35	61
	$2 \leqslant y$	124	140

The data are to be used to determine whether or not there is an association between the application of the treatment and the number of years that a fruit tree remains free from this disease.

(i) Applied and $y < 1$
(ii) Not applied and $1 \leq y < 2$

(2)

The value of $\sum \frac{(O-E)^2}{E}$ for the **other four** classes is 2.642 to 3 decimal places.

(b) Test, at the 5% level of significance, whether or not there is an association between the application of the treatment and the number of years a fruit tree remains free from this disease.

You should state your hypotheses, test statistic, critical value and conclusion clearly.

(5)

(Total for question = 7 marks)

Q7.

Liam and Simone are studying the distribution of oak trees in some woodland. They divided the woodland into 80 equal squares and recorded the number of oak trees in each square. The results are summarised in Table 1 below.

Number of oak trees in a square	0	1	2	3	4	5	6	7 or more
Frequency	1	4	21	23	13	11	7	0

Table 1

Liam believes that the oak trees were deliberately planted, with 6 oak trees per square and that a constant proportion *p* of the oak trees survived.

(a) Suggest the model Liam should use to describe the number of oak trees per square.

(2)

Liam decides to test whether or not his model is suitable and calculates the expected frequencies given in Table 2.

Number of oak trees in a square	0 or 1	2	3	4	5	6
Expected frequency	5.53	14.89	24.26	22.24	10.87	2.21

Table 2

(b) Showing your working clearly, complete the test using a 5% level of significance.

You should state your critical value and conclusion clearly.

(7)

Simone believes that a Poisson distribution could be used to model the number of oak trees per square. She calculates the expected frequencies given in Table 3.

Number of oak trees in a square	0 or 1	2	3	4	5	6 or more
Expected frequency	12.69	16.07	s	14.58	t	9.37

Table 3

(c) Find the value of *s* and the value of *t*, giving your answers to 2 decimal places.

(4)

(d) Write down hypotheses to test the suitability of Simone's model.

(1)

The test statistic for this test is 8.749

(e) Complete the test. Use a 5% level of significance and state your critical value and conclusion clearly.

(3)

(f) Using the results of these tests, explain whether the origin of this woodland is likely to be cultivated or wild.

(2)

(Total for question = 19 marks)

Q8.

A researcher is investigating the distribution of orchids in a field. He believes that the Poisson distribution with a mean of 1.75 may be a good model for the number of orchids in each square metre. He randomly selects 150 non-overlapping areas, each of one square metre, and counts the number of orchids present in each square.

The results are recorded in the table below.

Number of orchids in each square metre	0	1	2	3	4	5	6
Number of squares	30	42	35	26	11	6	0

He calculates the **expected** frequencies as follows

Number of orchids in each square metre	0	1	2	3	4	5	More than 5
Number of squares	26.07	45.62	39.91	23.28	10.19	3.57	r

(a) Find the value of r giving your answer to 2 decimal places.

(1)

The researcher will test, at the 5% level of significance, whether or not the data can be modelled by a Poisson distribution with mean 1.75

(b) State clearly the hypotheses required to test whether or not this Poisson distribution is a suitable model for these data.

(1)

The test statistic for this test is 2.0 and the number of degrees of freedom to be used is 4

(c) Explain fully why there are 4 degrees of freedom.

(2)

(d) Stating your critical value clearly, determine whether or not these data support the researcher's belief.

(2)

The researcher works in another field where the number of orchids in each square metre is known to have a Poisson distribution with mean 1.5

He randomly selects 200 non-overlapping areas, each of one square metre, in this second field, and counts the number of orchids present in each square.

(e) Using a Poisson approximation, show that the probability that he finds at least one square with exactly 6 orchids in it is 0.506 to 3 decimal places.

(4)

(Total for question = 10 marks)

Q9.

The discrete random variable X follows a Poisson distribution with mean 1.4

(a) Write down the value of

(i) $P(X = 1)$
(ii) $P(X \leq 4)$

(2)

The manager of a bank recorded the number of mortgages approved each week over a 40 week period.

Number of mortgages approved	0	1	2	3	4	5	6
Frequency	10	16	7	4	2	0	1

(b) Show that the mean number of mortgages approved over the 40 week period is 1.4

(1)

The bank manager believes that the Poisson distribution may be a good model for the number of mortgages approved each week.

She uses a Poisson distribution with a mean of 1.4 to calculate expected frequencies as follows.

Number of mortgages approved	0	1	2	3	4	5 or more
Expected frequency	9.86	r	9.67	4.51	1.58	s

(c) Find the value of r and the value of s giving your answers to 2 decimal places.

(2)

The bank manager will test, at the 5% level of significance, whether or not the data can be modelled by a Poisson distribution.

(d) Calculate the test statistic and state the conclusion for this test. State clearly the degrees of freedom and the hypotheses used in the test.

(6)

(Total for question = 11 marks)

Q10.

A leisure club offers a choice of one of three activities to its 150 members on a Tuesday evening. The manager believes that there may be an association between the choice of activity and the age of the member and collected the following data.

Age a years \ Activity	Badminton	Bowls	Snooker
$a < 20$	9	3	3
$20 \leqslant a < 40$	10	10	14
$40 \leqslant a < 50$	16	15	5
$50 \leqslant a < 60$	15	13	11
$a \geqslant 60$	4	19	3

(a) Write down suitable hypotheses for a test of the manager's belief.

(1)

The manager calculated expected frequencies to use in the test.

(b) Calculate the expected frequency of members aged 60 or over who choose snooker, used by the manager.

(1)

(c) Explain why there are 6 degrees of freedom used in this test.

(2)

The test statistic used to test the manager's belief is 19.583

(d) Using a 5% level of significance, complete the test of the manager's belief.

(2)

(Total for question = 6 marks)

Q11.

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)

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)

(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)

(Total for question = 14 marks)

Q12.

A spinner used for a game is designed to give scores with the following probabilities

Score	1	2	3	4	6
Probability	$\frac{3}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{2}{5}$	$\frac{1}{10}$

The spinner is spun 80 times and the results are as follows

Score	1	2	3	4	6
Frequency	15	4	12	41	8

Test, at the 10% level of significance, whether or not the spinner is giving scores as it is designed to do. Show your working and state your hypotheses clearly.

(Total for question = 7 marks)

Q13.

Abram carried out a survey of two treatments for a plant fungus. The contingency table below shows the results of a survey of a random sample of 125 plants with the fungus.

		Treatment		
		No action	Plant sprayed once	Plant sprayed every day
Outcome	Plant died within a month	15	16	25
	Plant survived for 1 – 6 months	8	25	10
	Plant survived beyond 6 months	7	14	5

Abram calculates expected frequencies to carry out a suitable test. Seven of these are given in the partly-completed table below.

		Treatment		
		No action	Plant sprayed once	Plant sprayed every day
Outcome	Plant died within a month			17.92
	Plant survived for 1 – 6 months	10.32	18.92	13.76
	Plant survived beyond 6 months	6.24	11.44	8.32

The value of $\sum \frac{(O-E)^2}{E}$ for the 7 given values is 8.29

Test at the 2.5% level of significance, whether or not there is an association between the treatment of the plants and their survival. State your hypotheses and conclusion clearly.

(7)

(Total for question = 7 marks)

Q14.

A university foreign language department carried out a survey of prospective students to find out which of three languages they were most interested in studying.

A random sample of 150 prospective students gave the following results.

		Language		
		French	**Spanish**	**Mandarin**
Gender	**Male**	23	22	20
	Female	38	32	15

A test is carried out at the 1% level of significance to determine whether or not there is an association between gender and choice of language.

(a) State the null hypothesis for this test.

(1)

(b) Show that the expected frequency for females choosing Spanish is 30.6

(1)

(c) Calculate the test statistic for this test, stating the expected frequencies you have used.

(3)

(d) State whether or not the null hypothesis is rejected. Justify your answer.

(2)

(e) Explain whether or not the null hypothesis would be rejected if the test was carried out at the 10% level of significance.

(1)

(Total for question = 8 marks)

<u>Mark Scheme</u> – Chi-Squared Test

Q1.

Question	Scheme	Marks	AOs
(a)	H_0: There is no association between the hand and the number of heads H_1: There is an association between the hand and the number of heads	B1	2.5
	$\begin{array}{c\|c\|c\|c\|c} & \mathbf{0} & \mathbf{1} & \mathbf{2} & \mathbf{3} \\ \hline \mathbf{E} & \frac{100\times 20}{200} = 10 & \frac{100\times 64}{200} = 32 & \frac{100\times 78}{200} = 39 & \frac{100\times 38}{200} = 19 \end{array}$	M1 A1	1.1b 1.1b
	$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(7-10)^2}{10} + \frac{(13-10)^2}{10} + \ldots + \frac{(16-19)^2}{19}$	M1	1.1b
	$= 3.7714\ldots$ awrt <u>**3.77**</u>	A1	1.1b
	Degrees of freedom $[= (4-1)\times(2-1)] = 3$ $\chi^2_{3,0.05} = 7.815$	M1	3.1b
	(Do not reject H_0.) There is not enough evidence to suggest an association between the hand flipping the coin and the number of heads.	A1	2.2b
		(7)	
(b)	B(3, 0.5)	B1	3.3
		(1)	
(c)	H_0: B(3, 0.5) is a suitable model H_1: B(3, 0.5) is not a suitable model	B1ft	3.4
	$\begin{array}{c\|c\|c\|c\|c} & \mathbf{0} & \mathbf{1} & \mathbf{2} & \mathbf{3} \\ \hline \mathbf{E} & 200\times P(X=0) = 25 & 200\times P(X=1) = 75 & 200\times P(X=2) = 75 & 200\times P(X=3) = 25 \end{array}$	M1 A1	2.1 1.1b
	$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(20-25)^2}{25} + \frac{(64-75)^2}{75} + \frac{(78-75)^2}{75} + \frac{(38-25)^2}{25}$	M1	1.1b
	$= 9.493\ldots$ awrt <u>**9.49**</u>	A1	1.1b
	[df = 3] $\chi^2_{3,0.1} = 6.251$	M1	3.1b
	(Reject H_0) B(3, 0.5) is <u>not a suitable model</u> for the number of heads.	A1	3.5a
		(7)	
			(15 marks)

<table>
<tr><th colspan="2">Notes</th></tr>
<tr><td>(a)</td><td>B1: For both hypotheses correct with at least one in context.
M1: For attempt at $\frac{\text{row total} \times \text{column total}}{\text{grand total}}$ (may be implied by one correct expected frequency). Working may be seen in table.
A1: All correct expected frequencies</td></tr>
<tr><td></td><td>M1: For applying $\sum \frac{(O-E)^2}{E}$ ft their values
A1: awrt 3.77
M1: For using degrees of freedom to set up χ^2 model
A1: Correct conclusion in context with all other marks scored.</td></tr>
<tr><td>(b)</td><td>B1: B(3, 0.5)

Allow a complete probability distribution with labels<table><tr><td>x</td><td>0</td><td>1</td><td>2</td><td>3</td></tr><tr><td>$P(X=x)$</td><td>0.125</td><td>0.375</td><td>0.375</td><td>0.125</td></tr></table></td></tr>
<tr><td>(c)</td><td>B1ft: For both hypotheses correct. Must have binomial and (3, 0.5) or ft their distribution in part (b)
M1: For attempt at expected frequencies using their distribution from part (b) (may be implied by one correct or correct ft expected frequency)
A1: All correct expected frequencies
M1: For applying $\sum \frac{(O-E)^2}{E}$ ft their values
A1: awrt 9.49
M1: For using degrees of freedom to set up χ^2 model
A1: Correct conclusion in context with all other marks scored.</td></tr>
</table>

Q2.

Qu.	Scheme	Marks	AOs
(a)	$p = \dfrac{(0)+11+14+6+(0)+5+(0)}{6 \times 40}$	M1	2.1
	$p = \underline{0.15}$ *	A1*cso	1.1b
		(2)	
(b)	X~B(6, 0.15) $\begin{array}{c\|c\|c\|c\|c\|c} x & 2 & 3 & 4 & 5 & 6 \\ \hline 40\times \mathrm{P}(X=x) & 7.04\ldots & 1.65\ldots & 0.219\ldots & 0.015\ldots & 0.00\ldots \end{array}$	M1	3.4
	Require 40 × P($X \geq k$) > 5 Exp. frequency for $X \geq 2 = 8.94\ldots$ / $X \geq 3 = 1.89\ldots$	M1	1.1b
	Combine last 5 cells / only 3 cells in total	A1	2.2a
	2 is subtracted (as there are 2 restrictions) and the proportion used from data (and 1 equal totals)	B1	2.4
	$3 - 2 = 1$ degree of freedom	A1	1.1b
	H_0: Binomial distribution is a suitable model H_1: Binomial distribution is not a suitable model	B1	3.4
	Critical value $\chi^2_{(1, 0.10)} = 2.705$ or 2.706	B1ft	1.1b
	Test statistic is not in the critical region, insufficient evidence to reject H_0 $(2.689 < 2.705/6)$ Data are consistent with binomial/engineer's/suggested model.	B1ft	3.5a
		(8)	
(c)	The total amount/proportion of defective pins remains the same.	M1	2.4
	The cells for $X \geq 2$ are still combined in the test.	M1	1.1b
	So there is no change to the value of the test statistic.	A1	2.2a
		(3)	
			(13 marks)

Notes	
(a)	**M1:** Correct expression for p (may be seen in stages). Allow $\frac{36}{240}$ but not $\frac{6}{40}$ on its own **A1*cso:** $p = 0.15$ stated and no incorrect working seen
(b)	**M1:** Attempting to find expected frequencies, at least 2 correct trunc. or rounded 1dp **M1:** Recognising need to combine cells (Sight of awrt 8.94 implies M1M1) **A1:** Combining cells for $X \geq 2$ (to make 3 cells) **B1:** Justifying why 2 is subtracted with p being calculated from data **A1:** 1 degree of freedom **B1:** Correct hypotheses (0.15 must **not** be included) Allow engineer's model. **B1ft:** Correct critical value (ft their df) May see $\chi^2_{(2, 0.10)} = 4.605$ or $\chi^2_{(3, 0.10)} = 6.251$ **B1ft:** Correct inference (ft comparison of their CV with 2.689). Condone $p = 0.15$ included here. Do not allow contradictory statements to score here. Hypotheses must be correct way round.
(c)	**M1**: Determining the number (N=36)/proportion (p=0.15) of defective pins has not changed. e.g. 11 + 12 + 9 + 4 = 36. But not 7 + 2 + 1 = 6 + 3 + 1 **M1**: Understanding the cells for $X \geq 2$ are still combined in the test **A1**: (dep on both M1s) Concluding that there is no change to the value of the test statistic.

Q3.

Question	Scheme	Marks	AOs
(a)	$E = \dfrac{(c+d)(a+c)}{a+b+c+d}$	B1	1.1b
	$O - E = c - "\dfrac{(c+d)(a+c)}{a+b+c+d}"$	M1	1.1b
	$O - E = \dfrac{ca+cb+c^2+cd-ac-c^2-ad-dc}{a+b+c+d}$	dM1	1.1b
	$O - E = \dfrac{cb-ad}{a+b+c+d}$	A1	1.1b
		(4)	
(b)	H_0: There is no association between the age of a person and the main type of investment they have. H_1: There is an association between the age of a person and the main type of investment they have.	B1	3.4
	Degrees of freedom $= (3-1)(2-1) = 2$ $\chi^2_{2,0.05} = 5.991$	M1	3.1b
	Reject H_0. There is evidence that there is an association between the age of a person and the main type of investment they have.	A1	2.2b
		(3)	
			(7 marks)

Notes:		
(a)	B1:	For correct expected value
	M1:	For finding c – their expected value
	dM1:	Dependent on previous method being awarded. For correctly gaining a single fraction
	A1:	Correct answer only
(b)	B1:	For correct hypotheses with at least one in context. Allow independent and not independent. Do not accept correlation
	M1:	For using degrees of freedom to set up χ^2 model critical value, implied by CV 5.991 or better
	A1:	Correct conclusion including the words **age** and **investment**. Do not allow contradicting statements. Do not award if hypotheses are the wrong way round or there are no hypotheses.

Q4.

Question	Scheme	Marks	AOs
(a)	Not all the expected frequencies are likely to be over 5 Or the sample size is too small.	B1	3.5b
		(1)	
(b)	5 degrees of freedom since the parameter is not estimated from the data [and the totals agree]	B1	2.4
		(1)	
(c)	H_0: B(5,0.6) is a suitable model H_1: B(5,0.6) is not a suitable model	B1	3.4
	$\sum\frac{(O-E)^2}{E}=\frac{(2-5.12)^2}{5.12}+...+\frac{(51-38.88)^2}{38.88}$	M1	2.1
	$= 15.8063...$ awrt 16	A1	1.1b
	[15.8 >] $\chi^2_{5,(0.05)}=11.070$	B1ft	1.1b
	B(5,0.6) is not a suitable model [for the number of heads spun]	A1ft	3.5a
		(5)	
(d)	$\frac{[0\times2]+(1\times27)+(2\times93)+(3\times181)+(4\times146)+(5\times51)}{500}$ [= 3.19]	M1	3.3
	B([5], $p=\frac{3.19}{5}=0.638$)	A1	1.1b
		(2)	
Notes:			**(9 marks)**

(a)	B1:	For recognising the limitations of using a chi squared model on small sample sizes eg 20 is not large, not enough data, sample needs to be larger, you may need to combine cells.
(b)	B1 :	For 5 [dof] and a correct reason indicating parameter(probability) is not estimated. Condone missing comment about totals
(c)	B1:	Both hypotheses correct Must have B(5,0.6) or binomial with number (n) = 5 and probability(p) = 0.6 (in at least 1) and be attached to H_0 and H_1 the right way round.
	M1:	Attempting to find the test statistic $\sum\frac{(O-E)^2}{E}$ (at least two correct expressions, fractions or decimals) or $\chi^2=\sum\frac{O^2}{E}=\frac{(2)^2}{"5.12"}+...+\frac{51^2}{38.88}-500$ (at least two correct expressions, fractions or decimals plus the − 500) Implied by awrt 15.8
	A1:	Awrt16
	B1ft:	Allow 11.07 or awrt 11.070 For correct CV, ft their answer to (b) NB dof 3 is 7.815 dof 4 is 9.488
	A1ft:	Ft "their 11.070" and their CV or p value. A correct conclusion independent of the hypotheses ie [If they should reject H_0 then they need "is not a suitable model.If they should accept H_0 then they need "is suitable"…] Allow Binomial is not a suitable model eg condone B(500, 0.6) is not a suitable model. Do not accept contradictory statements
		NB If p value [0.007419] given instead of CV they could get B1M1A1B0A1unless they give the CV as well
(d)	M1:	For a correct method using the data to improve the model. Implied by 3.19
	A1:	Correct model. Condone use of any value of n Accept Binomial with p = 0.638

Q5.

Question Number	Scheme	Notes	Marks
(a)	$\frac{72\times50}{150}=24,\ \frac{78\times50}{150}=26$	For one correct $\frac{\text{Row Total}\times\text{Column Total}}{\text{Grand Total}}$; can be implied by correct answers.	M1
	$\frac{72\times64}{150}=30.72,\ \frac{78\times64}{150}=33.28$	24, 26, 30.72, 33.28 only.	A1
			(2)
(b)	$\mathrm{H_0}$: Perceived (body) weight is independent of gender (no association) $\mathrm{H_1}$: Perceived (body) weight is not independent of gender (association)	Both hypotheses required. Must mention "Perceived", "weight" and "gender" at least once. Use of "relationship" or "correlation" or "connection" or "link" award B0.	B1
	O / E / $\frac{(O-E)^2}{E}$ / $\frac{O^2}{E}$: 20 / 17.28 / 0.428148 / 23.14815 22 / 24 / 0.166667 / 20.16667 30 / 30.72 / 0.016875 / 29.29688 16 / 18.72 / 0.395214 / 13.67521 28 / 26 / 0.153846 / 30.15385 34 / 33.28 / 0.015577 / 34.73558 150 / 150 / 1.176327 / 151.1763	M1 for at least 2 correct terms (as in 3rd or 4th column) or correct expressions. A1 for all correct. Accept 2sf accuracy. Allow truncation e.g. 1.17…	M1A1
	$\sum\frac{(O-E)^2}{E}$ or $\sum\frac{O^2}{E}-150=1.18$	Awrt 1.18-1.19	A1
	$\nu=(3-1)(2-1)=2,\ \chi_2^2(10\%)=4.605$	2 can be implied by 4.605 seen	B1B1ft
	(Accept $\mathrm{H_0}$) Perceived (body) **weight** is independent of **gender** (no association)	A correct comment in context - must mention "weight" and "gender". Condone "relationship" or "connection" here but **not** "correlation". Follow through from their test stat and cv, but **hypotheses must be correct.**	A1ft
			(7)

(c)	O / E / $\frac{(O-E)^2}{E}$ / $\frac{O^2}{E}$: 36 / 50 / 3.92 / 25.92 50 / 50 / 0 / 50 64 / 50 / 3.92 / 81.92 150 / 150 / 7.84 / 157.84	B1 for $E_i=50$, could be implied. M1 for combining values and for attempting $\frac{(O-E)^2}{E}$ or $\frac{O^2}{E}$ with at least 2 correct expressions or values. A1 for all correct, can be implied by correct answer below.	B1 M1A1
	$\sum\frac{(O-E)^2}{E}$ or $\sum\frac{O^2}{E}-150=7.84$	Awrt 7.84	A1
	$\nu=2,\ \chi_2^2(2.5\%)=7.378$	0.025 or 2.5%	A1
			(5)
			Total 14

Q6.

Question	Scheme	Marks	AOs
(a)	(i) $\frac{40\times174}{400}$ (ii) $\frac{96\times226}{400}$	M1	1.1b
	$=17.4$ $=54.24$	A1	1.1b
		(2)	
(b)	H_0: There is no association between the **application** of the **treatment** and the number of years that a fruit tree remains free from this disease. H_1: There is an association between the application of the treatment and the number of years that a fruit tree remains free from this disease.	B1	3.4
	$\sum\frac{(O-E)^2}{E}=\frac{(15-\text{"}17.4\text{"})^2}{\text{"}17.4\text{"}}+\frac{(61-\text{"}54.24\text{"})^2}{\text{"}54.24\text{"}}+2.642$	M1	1.1b
	$=3.815\ldots$ awrt 3.82	A1	1.1b
	[3.82 <] $\chi^2_{2,(0.05)}=5.991$	B1	3.1b
	There is no evidence of association between the **application** of the **treatment** and the number of years that a fruit tree remains free from this disease.	A1ft	2.2b
		(5)	
			(7 marks)

Notes:		
(a)	M1	A correct method to work out either expected frequencies – or 1 correct
	A1	17.4 and 54.24 (accept 54.2)
(b)	B1:	For both hypotheses in terms of "association" or independence" Must mention application/treatment and years in at least one and be connected correctly to H_0 and H_1 [Use of link, relationship or connection. is B0 but allow for last A1ft]
	M1:	A correct method to find the total χ^2 value. ft their values from (a) If no method shown at least 1 of the two missing χ^2 contributions must be correct $(0.331\ldots\left(\frac{48}{145}\right)$ and $0.8425\ldots$ allow 2sf). Implied by awrt 3.82
	A1:	awrt 3.82 or awrt 3.83
	B1:	Using the degrees of freedom to find the χ^2 CV for the appropriate model. awrt 5.991 allow 5.9915
	A1ft:	Ft "their 3.82" and their CV or p-value. Correct conclusion in context. (application or treatment and years) This is independent of hypotheses ie if they should accept H_0 then they need eg there is no association between …. If they should reject H_0 then they need there is an association"… Allow relationship, link, connection for association BUT do not accept correlation or contradictory statements
		NB If p-value [0.148388] given instead of CV could get B1M1A1B0A1 unless they give the CV as well

Q7.

Qu	Scheme	Marks	AO
(a)	[T = no. of oak trees in a square] $T \sim$ Binomial	M1	3.3
	$T \sim B(6, p)$	A1	1.1b
		(2)	
(b)	Expected frequency for 6 is less than 5 so pool: new $E_i = 13.08$	M1	2.1
	$\frac{(O_i - E_i)^2}{E_i}$: 0.051, 2.51, 0.0654, 3.84, 1.85; $\frac{O_i^2}{E_i}$: 4.521, 29.617, 21.805, 7.599, 24.771; $\sum \frac{(O_i-E_i)^2}{E_i} = 8.313$	M1,A1	1.1b x2
	p needed estimating ($\hat{p} = 0.55$) so $\nu = 5 - 2 = 3$; cv 7.815	B1,B1ft	1.1b x2
	Significant result, so Liam's model is not suitable	M1,A1	1.1b2.2b
		(7)	
(c)	[R = no. of oak trees in a square for Simone's model] $R \sim Po(3.3)$	M1	3.3
	Correct expression for s or t using Poisson	M1	3.4
	$s =$ **17.67** and $t =$ **9.62**	A1,A1	1.1b x2
		(4)	
(d)	H_0: Poisson is a good fit (for no. of oak trees per square) H_1: Poisson is not a good fit (for no. of oak trees per square)	B1	2.5
		(1)	
(e)	No pooling needed so degrees of freedom is $6 - 2 = 4$	B1	1.1b
	Critical value is 9.488 (accept 9.49)	B1	1.1a
	Not significant so Poisson (or Simone's) model is suitable	B1	2.2b
		(3)	
(f)	Poisson model has better fit so suggests that oak trees occur at random Or binomial suggests deliberately planted or cultivated	B1	2.2b
	Therefore the forest is likely to be wild not cultivated	B1	3.5a
		(2)	
		(19 marks)	

	Notes
(a)	M1 for choosing binomial A1 for B(6, p) can be in words and allow B(6, 0.55)
(b)	1st M1 for pooling last 2 classes ($E_i = 13.08$ but accept 13.1) 2nd M1 for at least 3 correct values or expressions. Either row to at least 2 sf 1st A1 for awrt 8.31 (8.31 gets 3/3) [NB no pooling gives awrt 16.8458.. and implies M0M1A0] 1st B1 for 3 degrees of freedom 2nd B1ft for critical value of 7.815 (e.g. $\nu = 4$ use 9.488) 3rd M1 for a correct conclusion (non-contextual ignore any contradictory contextual comments for this mark) based on their cv and their test statistic This mark can be implied by a fully correct solution ending with correct contextual conclusion 2nd A1 for correct conclusion in context with **all other marks scored**
(c)	1st M1 for selecting a correct model Po(3.3) [Allow Po(awrt 3.3)] 2nd M1 for use of the model with an expression or correct value for s or t 1st A1 for one correct 2nd A1 for both correct (allow awrt 2dp)
(d)	B1 for correct hypotheses must mention Poisson: use of Po(3.3) is B0
(e)	1st B1 for correct degrees of freedom $\nu = 4$ only 2nd B1 for selecting correct critical value (9.488 only) 3rd B1 for not significant conclusion based on 8.749 vs their cv (condone use of Po(3.3) here)
(f)	1st B1 for choosing Poisson as better or stating Poisson implies wild or bino'l implies cultivated 2nd B1 (dep on rejecting bin and accepting Poisson) for clearly stating woodland is wild If the tests give the same results then 2nd B0 automatically

Q8.

Question	Scheme	Marks	AOs
(a)	<u>1.36 or 1.37</u>	B1	1.1b
		(1)	
(b)	H_0: Po(1.75) is a suitable model H_1: Po(1.75) is not a suitable model	B1	3.4
		(1)	
(c)	Cells are combined for expected **frequencies** < **5** so **combine** the last **3 cells**	B1	2.4
	subtract 1 since **totals** agree	B1	2.4
		(2)	
(d)	$\chi_4^2 = 9.488$	B1	1.1b
	therefore, the **researcher's belief** is supported or evidence that **Po**(1.75) is a good model for the **number of orchids** in each square metre	B1ft	3.5a
		(2)	
(e)	P(exactly 6 orchids) = awrt 0.00353	B1	1.1b
	$X \sim \text{B}(200, \text{"}0.00353\text{"})$ mean $= 200 \times \text{"}0.00353\text{"} =$ awrt 0.706	M1	3.3
	$Y \sim \text{Po}(\text{"}0.706\text{"})$ $1 - \text{P}(Y = 0) = 1 - e^{-\text{"}0.706\text{"}}$	M1	3.4
	$= 0.506^*$	A1*	2.1
		(4)	
			(10 marks)

Notes
(a) **B1:** accept 1.36 or 1.37
(b) **B1:** For both hypotheses correct. Must have Po(1.75) or Poisson with mean 1.75 and be attached to H_0 and H_1 the right way round.
(c) **B1:** Explaining why there are 5 classes. Must mention **combine** the **3** cells when **frequencies** < **5** or to **combine** the 3 cells to make **frequency** > **5** **B1:** Explaining why 1 is subtracted. Must say/show **1 is subtracted and** Totals agree **or** Total frequency must be 150 **or** only need 4 pieces of data to find the other **or** λ is known or 1.75 is given. **NB** B0 for "only 1 constraint" on its own.
(d) **B1:** awrt 9.49 **B1ft:** ft their critical value only. For drawing the correct conclusion – condone missing 1.75. If hypotheses are the wrong way round or there are no hypotheses in (b) award B0
(e) **B1:** awrt 0.00353. May be implied by awrt 0.706 for mean. **M1:** Selecting the model $\text{B}(200, \text{"their P(exactly 6 orchids)"})$ and using np $(0 < p < 1)$ to find the mean. May be implied by awrt 0.706 **M1: Using** the model Po(their np) **and** using or writing $1 - \text{P}(Y = 0)$ or $1 - \text{P}(Y \le 0)$ or $1 - e^{-\text{"}0.706\text{"}}$ **A1*:** only award if the previous 3 marks have been awarded. **and 0.506** stated.

Q9.

Question	Scheme	Marks	AOs
(a)(i)	$P(X = 1) = 0.34523...$ awrt **0.345**	B1	1.1b
(a)(ii)	$P(X \leqslant 4) = 0.98575...$ awrt **0.986**	B1	1.1b
		(2)	
(b)	$\frac{(0\times10)+1\times16+2\times7+3\times4+4\times2+(5\times0)+6\times1}{40} = 1.4^*$	B1*cso	1.1b
		(1)	
(c)	$r = 40 \times$ '0.34523…' $s = 40 \times$ '1 − 0.986…'	M1	3.4
	$r =$ **13.81** $s =$ **0.57**	A1ft	1.1b
		(2)	
(d)	H_0: The Poisson distribution is a suitable model H_1: The Poisson distribution is not a suitable model	B1	3.4
	[Cells are combined when expected frequencies < 5] So combine the last 3 cells	M1	2.1
	$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(10-9.86)^2}{9.86} + ... + \frac{(7-(4.51+1.58+0.57))^2}{(4.51+1.58+0.57)}$	M1	1.1b
	awrt **1.1**	A1	1.1b
	Degrees of freedom = 4 − 1 − 1 = 2	B1	3.1b
	(Do not reject H_0 since $1.10 < \chi^2_{2,(0.05)} = 5.991$).The number of mortgages approved each week follows a Poisson distribution.	A1	3.5a
		(6)	
			(11 marks)

Notes	
(a)(i) **(a)(ii)**	B1 awrt 0.345 B1 awrt 0.986
(b)	B1* for a fully correct calculation leading to given answer with no errors seen
(c)	M1 for attempt at r or s (may be implied by correct answers) A1ft for both values correct (follow through their answers to part (a))
(d)	1st B1 for both hypotheses correct (lambda should not be defined so correct use of the model) 1st M1 for understanding the need to combine cells before calculating the test statistic (may be implied) 2nd M1 for attempt to find the test statistic using $\chi^2 = \sum \frac{(O-E)^2}{E}$ 1st A1 awrt 1.1 2nd B1 for realising that there are 2 degrees of freedom leading to a critical value of $\chi_2^2(0.05) = 5.991$ 2nd A1 concluding that a Poisson model is suitable for the number of mortgages approved each week

Q10.

	Scheme	Marks	AO
(a)	H_0 : There is no association between age and activity H_1 : There is an association between age and activity	B1 (1)	1.2
(b)	$\frac{26\times36}{150}=6.24$	B1 (1)	1.1b
(c)	Since expected value in $a<20$ and snooker $=3.6<5$ we amalgamate two rows	B1	2.4
	Table is now 4×3 so degrees of freedom is $(4-1)\times(3-1)=6$ (*)	B1* (2)	1.1b
(d)	Critical value $\chi_6^2(5\%)=12.592$	B1	1.1b
	[Significant result]: so there is evidence to support manager's belief	B1ft (2)	2.2b
		(6 marks)	

Notes	
(a)	B1 for both hypotheses in terms of "association" or "independence". Must mention age and activity (or sport). [Use of "relationship" or "link" here is B0 but allow for last B1ft]
(b)	B1 for 6.24
(c)	1st B1 for a reason to get a 4 x 3 table based on amalgamation of rows Must mention $a<20$ and snooker and see 3.6 and be combining <u>rows</u> (not columns) 2nd B1* for 6 degrees of freedom clearly coming from a 4 x 3 table formed from amalgamation of <u>rows</u>. [$8-2=6$ is B0]
(d)	1st B1 for correct critical value (allow 12.6 or 12.59 or awrt 12.592) NB p-value = 0.0032839… so allow awrt 0.00328 2nd B1ft for a correct comparison and conclusion (ft their cv) [<u>Independent of hypotheses</u>] e.g. there is an "association" or "relationship" or "link" between age and activity is OK BUT there is a "correlation" between age and activity is B0 Do not accept contradictory contextual statements e.g. "manager's belief supported , there is no association between age and activity"

Q11.

Question	Scheme	Marks	AOs
(a)	Expected value for 2 $= 150 \times P(X = 2)$	M1	3.4
	$= 28.3015...$	A1	1.1b
	Expected value for 4 or more $= 150 - (53.8 + 56.6 + 28.3 + 8.9)$ $= 2.4$	A1ft	1.1b
	H_0: Bin(20, 0.05) is a suitable model H_1: Bin(20, 0.05) is not a suitable model	B1	2.5
	Combining last two groups <table><tr><td></td><td>$\geqslant$ **3**</td></tr><tr><td>**Observed frequency**</td><td>19</td></tr><tr><td>**Expected frequency**</td><td>11.3</td></tr></table>	M1	2.1
	$\nu = 4 - 1 = 3$	B1	1.1b
	Critical value, $\chi_3^2(0.05) = 7.815$	B1	1.1a
	Test statistic $= \dfrac{(43-53.8)^2}{53.8} + \dfrac{(62-56.6)^2}{56.6} + ...$	M1	1.1b
	$= 8.117$	A1	1.1b
	In critical region, sufficient evidence to reject H_0, accept H_1 Significant evidence at 5% level to reject the manager's model	A1	3.5a
		(10)	
(b)	$\nu = 4 - 2 = 2$		
	4 classes due to pooling	B1	2.4
	2 restrictions (equal total and mean/proportion)	B1	2.4
		(2)	
(c)	H_0: Binomial distribution is a good model H_1: Binomial distribution is not a good model	B1	3.4
	Critical value, $\chi_2^2(0.05) = 5.991$ 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	3.5a
		(2)	
		(14 marks)	

Notes	
(a)	**M1:** Using the binomial model $150 \times p^2 \times (1-p)^{18}$ may be implied by 28.3 **A1:** awrt 28.3 **A1:** awrt 2.4 or ft their "28.3" **B1:** Both hypotheses correct using the correct notation or written out in full. **M1:** For recognising the need to combine groups **B1:** Number of degrees of freedom = 3 may be implied by a correct CV **B1:** awrt 7.82 **M1:** Attempting to find $\sum \dfrac{O_i - E_i}{E_i}^2$ or $\sum \dfrac{O_i^2}{E_i} - N$ may be implied by awrt 8.12 **A1:** awrt 8.12 **A1:** Evaluating the outcome of a model by drawing a correct inference in context
(b)	**B1:** Explaining why there are 4 classes **B1:** Explanation of why 2 is subtracted
(c)	**B1:** Correct hypotheses for the refined model **B1:** The CV awrt 5.99 and drawing the correct inference for the refined model

Q12.

	Scheme	Marks	AO
	H_0: Spinner is working as designed (o.e.) H_1: Spinner is not working as designed (o.e.)	B1	1.2
	(table below)	M1 A1	3.4 1.1b
		M1	1.1b
	$\sum \frac{(O_i - E_i)^2}{E_i} = 3.375 + 2 + 2 + 2.53125 + 0 = 9.90625$ or $\sum \frac{O_i^2}{E_i} - N = 9.375 + 2 + 18 + 52.53125 + 8 - 80 = 9.90625$	A1	1.1b
	$\nu = 5 - 1 = 4$ so $\chi_4^2(10\%)$ cv $= 7.779$ or better	B1	3.4
	Result is significant so there is evidence that the spinner is not operating as designed	A1cso	3.5a
		(7 marks)	

E_i	24	8	8	32	8
O_i	15	4	12	41	8
$\frac{(O_i - E_i)^2}{E_i}$	$\frac{81}{24}$	$\frac{16}{8}$	$\frac{16}{8}$	$\frac{81}{32}$	0
$\frac{O_i^2}{E_i}$	$\frac{225}{24}$	$\frac{16}{8}$	$\frac{144}{8}$	$\frac{1681}{32}$	$\frac{64}{8}$

Notes

1st B1 for both hypotheses given in suitable context
1st M1 for using the model to find at least 2 correct expected frequencies
1st A1 for all correct E_i
2nd M1 for attempt to find test statistic (at least two correct expressions, fractions or decimals)
2nd A1 for a correct test statistic (awrt 9.91) [accept $\frac{317}{32}$]
2nd B1 for correct critical value (allow 7.78)
NB p- value = 0.042036…so allow awrt 0.042
3rd A1cso dep on <u>all</u> previous marks for a correct conclusion in context
(can be in terms of model or spinner's design) Must mention spinner **and** scores <u>or</u> design
Accept "spinner is not accurate"

Q13.

Question	Scheme	Marks	AOs
	H_0: There is no association between the treatment of the plants and their survival/outcome. H_1: There is an association between the treatment of the plants and their survival/outcome	B1	3.4
	(see table below)	M1 A1	1.1b 1.1b
	$\chi^2 = \sum\frac{(O-E)^2}{E} = \frac{(15-"13.44")^2}{"13.44"} + \frac{(16-"24.64")^2}{"24.64"} + 8.29$	M1	1.1b
	awrt <u>11.5</u>	A1	1.1b
	Degrees of freedom (3 −1) (3 −1) = 4 $\chi^2_{4,0.025} = 11.143$	M1	3.1b
	Reject H_0 There is an association between the treatment of the plants and their survival/outcome	dA1ft	2.2b
			(7 marks)

	No action	Plant sprayed once	Plant sprayed every day
Plant died within a month	**13.44**	**24.64**	17.92
Plant survived for 1 – 6 months	10.32	18.92	13.76
Plant survived beyond 6 months	6.24	11.44	8.32

Notes

B1: For correct hypotheses at least one in context. Allow independent and not independent. Do not accept correlation.

M1: For attempt at $\frac{(\text{Row Total})(\text{Column Total})}{(\text{Grand Total})}$ to find expected frequencies. (they may put numbers in table)

A1: awrt 13.44 and 24.64 This may be implied by a correct value of χ^2

M1: For applying $\sum\frac{(O-E)^2}{E}$ ft their expected values. If no method shown at least 1 of the two missing χ^2 contributions must be correct – you may need to check this (correct ones are 0.181… and 3.0296… allow 2sf) (condone missing 8.29)

A1: awrt 11.5

M1: For using degrees of freedom to set up χ^2 model critical value, implied by CV 11.143 or better

dA1ft: dependent on the 2nd and 3rd M marks. Correct conclusion ft their $\sum\frac{(O-E)^2}{E}$ there is an **association** between the **treatment** of the plants and their **survival**/outcome: - do not allow contradicting statements. Do not award if hypotheses are the wrong way round or there are no hypotheses.

Q14.

Question	Scheme	Marks	AOs
(a)	H_0: There is no association between language and gender.	B1	1.2
		(1)	
(b)	$\frac{54\times85}{150}=30.6$ *	B1*cso	1.1b
		(1)	
(c)	Expected frequencies: Language — French / Spanish / Mandarin; Gender Male: 26.43… / 23.4 / 15.16…; Female: 34.56… / [30.6] / 19.83…	M1	2.1
	$\chi^2=\sum\frac{(O-E)^2}{E}=\frac{(23-26.43)^2}{26.43}+\ldots+\frac{(15-19.83)^2}{19.83}$	M1	1.1b
	awrt **<u>3.6/3.7</u>**	A1	1.1b
		(3)	
(d)	Degrees of freedom $(3-1)(2-1)\to$ Critical value $\chi^2_{2,0.01}=9.210$	M1	3.1b
	As $\sum\frac{(O-E)^2}{E}<9.210$, the null hypothesis is not rejected.	A1	2.2b
		(2)	
(e)	Still not rejected since $\sum\frac{(O-E)^2}{E}<\chi^2_{2,0.1}=4.605$	B1	2.4
		(1)	
			(8 marks)

Expected frequencies table for (c):

Expected frequencies		Language		
		French	Spanish	Mandarin
Gender	Male	26.43…	23.4	15.16…
	Female	34.56…	[30.6]	19.83…

Notes	
(a)	B1 for correct hypothesis in context
(b)	B1* for a correct calculation leading to the given answer and no errors seen
(c)	M1 for attempt at $\frac{\text{(Row Total)(Column Total)}}{\text{(Grand Total)}}$ to find expected frequencies M1 for applying $\sum\frac{(O-E)^2}{E}$ A1 awrt 3.6 or 3.7
(d)	M1 for using degrees of freedom to set up a χ^2 model critical value A1 for correct comparison and conclusion
(e)	B1 for correct conclusion with supporting reason