Chi-squared tests 6D

1 a H_0: The data can be modelled by a Po(2) distribution.
H_1: The data cannot be modelled by Po(2) distribution.

The observed and expected results are shown in the table. The last two columns (for 5 and >5) have been combined to get all E values to be greater than 5.

x	0	1	2	3	4	$\geqslant 5$	Total
Observed (O_i)	12	23	24	24	12	5	100
Expected (E_i)	13.53	27.07	27.07	18.04	9.02	5.27	100
$\frac{(O_i - E_i)^2}{E_i}$	0.1730	0.6119	0.3882	1.9690	0.9843	0.0138	4.101

Note that the values found for the test statistic (X^2) will vary according to any rounding you do. The test statistic must be calculated with sufficient accuracy to ensure that the chi-squared test is correct. A range of answers that encompass sensible degrees of accuracy would be accepted in an examination.

The number of degrees of freedom $\nu = 6 - 1 = 5$ (there are six cells after combining the last two columns with a single constraint on the total – that the frequencies agree).
From the tables: $\chi^2_5(5\%) = 11.070$

As 4.101 is less than 11.070, there is insufficient evidence to reject H_0 at the 5% level. The data may be modelled by Po(2).

b If λ is calculated, then this becomes another constraint and there would be one less degree of the freedom; $\nu = 6 - 2 = 4$

2 H_0: The data can be modelled by a discrete uniform distribution.
H_1: The data cannot be modelled by a discrete uniform distribution.

The number of degrees of freedom $\nu = 5$ (six data cells with a single constraint on the total)
From the tables: $\chi^2_5(5\%) = 11.070$

$$\text{Expected frequency} = \frac{15+23+19+20+14+11}{6} = \frac{102}{6} = 17$$

$$\sum \frac{(O_i - E_i)^2}{E_i} = \frac{1}{17}\left((15-17)^2 + (23-17)^2 + (19-17)^2 + (20-17)^2 + (14-17)^2 + (11-17)^2\right)$$

$$= \frac{1}{17}(4 + 36 + 4 + 9 + 9 + 36) = \frac{98}{17} = 5.765$$

As 5.765 is less than 11.070, there is insufficient evidence to reject H_0 at the 5% level. The data may be modelled by a discrete uniform distribution.

3 a $\bar{X} = \dfrac{1\times13+2\times9+3\times13}{15+13+9+13} = \dfrac{70}{50} = 1.4$

b H_0: The data can be modelled by Po(1.4).
H_1: The data cannot be modelled by Po(1.4).

As $\lambda = 1.4$ the expected frequencies must be calculated using these equations:

$P(X=i) = \dfrac{e^{-1.4}1.4^i}{i!}$ and $E_i = 50P(X=i)$ as there are 50 observations in the data

Calculate the probability in the final column (for $x \geqslant 5$) by summing the other probabilities and subtracting from 1.

The observed and expected results are:

x	0	1	2	3	4	$\geqslant 5$	Total
Observed (O_i)	15	13	9	13	0	0	50
P($X=i$)	0.2466	0.3452	0.2417	0.1128	0.0395	0.0143	1
Expected (E_i)	12.330	17.262	12.083	5.639	1.974	0.712	50

The last 3 classes must be combined so *all* the E values are 5 or more. This gives:

x	0	1	2	$\geqslant 3$	Total
Observed (O_i)	15	13	9	13	50
P($X=i$)	0.2466	0.3452	0.2417	0.1666	1
Expected (E_i)	12.330	17.262	12.083	8.33	50
$\dfrac{(O_i-E_i)^2}{E_i}$	0.5782	1.0514	0.7875	2.6181	5.0352

The number of degrees of freedom $\nu = 2$ (four data cells with two constraints as λ is estimated by calculation)

From the tables: $\chi^2_2(10\%) = 4.605$

As 5.035 is greater than 4.605, H_0 should be rejected at the 10% level. The data cannot be modelled by Po(1.4).

4 a $\bar{r} = \dfrac{1\times 26 + 2\times 36 + 3\times 20 + 4\times 10 + 5\times 6 + 6\times 2}{26+36+20+10+6+2} = \dfrac{240}{100} = 2.4$

mean $\bar{r} = np$

So $2.4 = 6p \Rightarrow p = 0.4$

This is one constraint since a parameter p has been estimated by calculation.

b H_0: The data can be modelled by B(6, 0.4)
H_1: The data cannot be modelled by B(6, 0.4)

Find the expected frequencies by multiplying the total frequency 100 (this is a second constraint) by the probability $P(X = i)$ using the probability equation for a binomial random variable.

$$E(X=0) = 100\times P(X=0) = 100\times \binom{6}{0}\times 0.4^0\times 0.6^6 = 4.666$$

$$E(X=1) = 100\times P(X=1) = 100\times \binom{6}{1}\times 0.4^1\times 0.6^5 = 18.662$$

Combine to get $E \geqslant 5$

$$E(X=2) = 100\times P(X=2) = 100\times \binom{6}{2}\times 0.4^2\times 0.6^4 = 31.104$$

$$E(X=3) = 100\times P(X=3) = 100\times \binom{6}{3}\times 0.4^3\times 0.6^3 = 27.648$$

$$E(X=4) = 100\times P(X=4) = 100\times \binom{6}{4}\times 0.4^4\times 0.6^2 = 13.824$$

$$E(X=5) = 100\times P(X=5) = 100\times \binom{6}{5}\times 0.4^5\times 0.6^1 = 3.6864$$

$$E(X=6) = 100\times P(X=6) = 100\times \binom{6}{6}\times 0.4^6\times 0.6^0 = 0.4096$$

Combine to get $E \geqslant 5$

After combining the relevant cells, this gives:

x	$\leqslant 1$	2	3	$\geqslant 4$	Total
Observed (O_i)	26	36	20	18	100
Expected (E_i)	23.328	31.104	27.648	17.92	100
$\dfrac{(O_i - E_i)^2}{E_i}$	0.3061	0.7707	2.1156	0.0004	3.1927

The number of degrees of freedom $\nu = 2$ (four data cells with two constraints as p is estimated by calculation)

From the tables: $\chi^2_2(5\%) = 5.991$

As 3.19 is less than 5.991, there is insufficient evidence to reject H_0 at the 5% level. The data may be modelled by B(6, 0.4)

5 H_0: The rate of accidents is constant at the factories.
H_1: The rate of accidents isn't constant at the factories.

Total number of accidents $= 81$

Total number of employees $= 15$ (thousand)

Mean rate of accidents $= \frac{81}{15} = 5.4$ (per thousand)

The calculation of the mean rate is one constraint

Multiply the number of employees in each factory by the mean rate of accidents to get the expected frequencies of accidents. The observed and expected results are:

Factory	***A***	***B***	***C***	***D***	***E***	**Total**
Observed (O_i)	22	14	25	8	12	81
Expected (E_i)	21.6	16.2	27	5.4	10.8	81
$\frac{(O_i - E_i)^2}{E_i}$	0.0074	0.2988	0.1481	1.2519	0.1333	1.8395

There are 5 cells and 1 constraint, so the number of degrees of freedom is $5 - 1 = 4$

From the tables: $\chi^2_4(5\%) = 9.488$

As 1.84 is less than 9.488, there is insufficient evidence to reject H_0 at the 5% level. This supports the hypothesis that accidents occur at a constant rate at the factories.

6 H_0: The data can be modelled by a Poisson distribution.
H_1: The data cannot be modelled by Poisson distribution.

Total frequency $= 2+8+15+18+14+13+7+3 = 80$

$$\text{Mean} = \lambda = \frac{1\times 8+2\times 15+3\times 18+4\times 14+5\times 13+6\times 7+7\times 3}{80} = \frac{276}{80} = 3.45$$

Calculate the expected frequencies as follows:

$$E_0 = 80\times \text{P}(X=0) = 80\times \frac{\text{e}^{-3.45}\,3.45^0}{0!} = 2.540$$

$$E_1 = 80\times \text{P}(X=1) = 80\times \frac{\text{e}^{-3.45}\,3.45^1}{1!} = 8.762$$

Similarly $E_2 = 15.114$, $E_3 = 17.381$, $E_4 = 14.991$, $E_5 = 10.344$, $E_6 = 5.948$, $E_7 = 2.931$

$E_{i\geqslant 8} = 80-(E_0+E_1+\ldots+E_7) = 1.989$

To get values for E greater than 5, combine the first two cells and the last three cells:

x	$\leqslant 1$	2	3	4	5	$\geqslant 6$	Total
Observed (O_i)	10	15	18	14	13	10	80
Expected (E_i)	11.302	15.114	17.381	14.991	10.344	10.868	80
$\frac{(O_i-E_i)^2}{E_i}$	0.1500	0.0008	0.0220	0.0655	0.068 29	0.0693	0.990

The number of degrees of freedom $\nu = 4$ (six data cells with two constraints as λ is estimated by calculation)

From the tables: $\chi^2_4(5\%) = 9.488$

As 0.99 is less than 9.488, there is insufficient evidence to reject H_0 at the 5% level. The data may be modelled by a Poisson distribution.

7 a Breakdowns occur singly, independently and at random. They occur at a constant average rate.

b H_0: The data can be modelled by a Poisson distribution.
H_1: The data cannot be modelled by Poisson distribution.

Total frequency $= 50 + 24 + 12 + 9 + 5 = 100$

$$\text{Mean} = \lambda = \frac{1\times 24 + 2\times 12 + 3\times 9 + 4\times 5}{100} = \frac{95}{100} = 0.95$$

Calculate the expected frequencies as follows:

$$E_0 = 100 \times \text{P}(X = 0) = 100 \times \frac{\text{e}^{-0.95}\,0.95^0}{0!} = 38.674$$

$$E_1 = 100 \times \text{P}(X = 1) = 100 \times \frac{\text{e}^{-0.95}\,0.95^1}{1!} = 36.740$$

Similarly $E_2 = 17.452,\ E_3 = 5.526,\ E_4 = 1.3125$

There is no need to go further, as further terms are extremely small. Find $E_{i\geq 3}$ to get all the E values to be 5 or more.

x	0	1	2	$\geqslant 3$	Total
Observed (O_i)	50	24	12	14	100
Expected (E_i)	38.674	36.740	17.452	7.134	100
$\frac{(O_i - E_i)^2}{E_i}$	3.317	4.4188	1.703	6.608	16.05

The number of degrees of freedom $\nu = 2$ (four data cells with two constraints as λ is estimated by calculation)

From the tables: $\chi^2_2(5\%) = 5.991$

As 16.05 is greater than 5.991, reject H_0 at the 5% level. The data cannot be modelled by Po(0.95)

8 H_0: The prizes are uniformly distributed.
H_1: The prizes are not uniformly distributed.

Total frequency $= 505$, so expected frequency for each class $= \frac{505}{10} = 50.5$

$$\text{Test statistic } (X^2) = \sum \frac{(O_i - E_i)^2}{E_i} = \frac{(56-50.5)^2}{50.5} + \frac{(49-50.5)^2}{50.5} + \ldots + \frac{(50-50.5)^2}{50.5} = 10.74$$

The number of degrees of freedom $\nu = 9$ (ten data cells with a single constraint on the total)

From the tables: $\chi^2_9(5\%) = 16.919$

As 10.74 is less than 16.919, there is insufficient evidence to reject H_0 at the 5% level. There is no reason to doubt that the prizes are distributed uniformly.

9 a The expected number of litters is modelled by B(8, 0.5)

As total frequency $= 200$

$$R = 200 \times \mathrm{P}(X = 3) = 200 \times \left(\binom{8}{3} \times 0.5^3 \times 0.5^5 \right) = 43.75$$

$$S = 200 \times \mathrm{P}(X = 4) = 200 \times \left(\binom{8}{4} \times 0.5^4 \times 0.5^4 \right) = 54.69$$

$$T = 200 \times \mathrm{P}(X = 5) = 200 \times \left(\binom{8}{5} \times 0.5^5 \times 0.5^3 \right) = 43.75$$

b H_0: The data can be modelled by B(8, 0.5)
H_1: The data cannot be modelled by B(8, 0.5)

To get values for E greater than 5, combine the first two cells and the last two cells:

No of females	**$\leqslant 1$**	**2**	**3**	**4**	**5**	**6**	**$\geqslant 7$**	**Totals**
Observed (O_i)	10	27	46	49	35	26	7	200
Expected (E_i)	7.03	21.88	43.75	54.69	43.75	21.88	7.03	200
$\frac{(O_i - E_i)^2}{E_i}$	1.255	1.198	0.116	0.592	1.75	0.776	0.0001	5.69

The number of degrees of freedom $\nu = 6$ (seven data cells with one constraint; note that p is not estimated by calculation but given in the question)

From the tables: $\chi^2_6(5\%) = 12.592$

As 5.69 is less than 12.592, there is insufficient evidence to reject H_0 at the 5% level. The data may be modelled by B(8, 0.5)

c If p is estimated by calculation, this would give an extra constraint. This would reduce the degrees of freedom by 1 so $\nu = 5$

The critical value would become $\chi^2_5(5\%) = 11.070$. However, the test statistic ($X^2 = 5.69$) would still be less than this critical value so the conclusion would remain the same: there is insufficient evidence to reject H_0

10 a $$\text{Mean} = \frac{0 \times 33 + 1 \times 55 + 2 \times 80 + 3 \times 56 + 4 \times 56 + 5 \times 11 + 6 \times 5 + 7 \times 4}{33 + 55 + 80 + 56 + 56 + 11 + 5 + 4} = \frac{718}{300} = 2.4$$

$$\text{Variance} = \frac{\sum fx^2}{n} - \left(\frac{\sum fx}{n} \right)^2$$

$$= \frac{0 \times 33 + 1 \times 55 + 4 \times 80 + 9 \times 56 + 16 \times 56 + 25 \times 11 + 36 \times 5 + 49 \times 4}{300} - (2.4)^2$$

$$= \frac{2426}{300} - 5.76 = 2.33 \text{ (2 d.p.)}$$

b The fact that the sample mean is close to the variance supports the use of a Poisson distribution.

10 c $s = E_0 = 300 \times \mathrm{P}(X = 0) = 300 \times \dfrac{\mathrm{e}^{-2.4}\,2.4^0}{0!} = 27.2$ (1 d.p.)

$t = E_2 = 300 \times \mathrm{P}(X = 2) = 300 \times \dfrac{\mathrm{e}^{-2.4}\,2.4^2}{2!} = 78.4$ (1 d.p.)

d H_0: The data can be modelled by a Po(2.4) distribution.
H_1: The data cannot be modelled by Po(2.4) distribution.

e Expected frequency for '7 or more goals'

$$E_{i\geq 7} = 300 - (E_0 + E_1 + E_2 + E_3 + E_4 + E_5 + E_6)$$
$$= 300 - (27.2 + 65.3 + 78.4 + 62.7 + 37.6 + 18.1 + 7.2) = 3.5$$

f As $E_{i \geq 7}$ combine with E_6 to give the data cell '6 or more goals'. There are now 7 data cells after combining these two values and two constraints as the mean has been calculated in part **a**, so there are 7 – 2 = 5 degrees of freedom.

g Test statistic $(X^2) = 15.7$; critical value is $\chi^2_5(5\%) = 11.070$

As 15.7 is greater than 11.070, H_0 should be rejected at the 5% level. The data cannot be modelled by Po(2.4)

11 a $\text{Mean} = \dfrac{0 \times 9 + 1 \times 24 + 2 \times 43 + 3 \times 34 + 4 \times 21 + 5 \times 15 + 6 \times 2}{9 + 24 + 43 + 34 + 21 + 15 + 2} = \dfrac{383}{148} = 2.59$ (2 d.p.)

b It could be assumed that the plants occur at a constant average rate and occur independently and at random in the meadow.

c $s = E_2 = 148 \times \mathrm{P}(X = 2) = 148 \times \dfrac{\mathrm{e}^{-2.59}\,2.59^2}{2!} = 37.24$ (2 d.p.)

$t = E_7 = 148 - (11.10 + 28.76 + 37.24 + 32.15 + 20.82 + 10.78 + 4.65) = 2.50$ (2 d.p.)

d H_0: The data can be modelled by a Po(2.59) distribution.
H_1: The data cannot be modelled by Po(2.59) distribution.

To get values for E greater than 5, combine the last two cells:

Number of plants	**0**	**1**	**2**	**3**	**4**	**5**	**⩾ 6**	**Total**
Observed (O_i)	9	24	43	34	21	15	2	148
Expected (E_i)	11.10	28.76	37.24	32.15	20.82	10.78	7.15	148
$\dfrac{(O_i - E_i)^2}{E_i}$	0.397	0.788	0.891	0.106	0.002	1.652	3.709	7.545

The number of degrees of freedom $\nu = 5$ (seven data cells with two constraints as λ has been estimated by calculation)

From the tables: $\chi^2_5(5\%) = 11.070$

As 7.545 is less than 11.070, there is insufficient evidence to reject H_0 at the 5% level. The data may be modelled by Po(2.59)