Review exercise 2

1 a Mean number of defectives in a sample $= \dfrac{0\times17+1\times31+2\times19+3\times17+4\times9+5\times7+6\times3}{17+31+19+14+9+7+3}$

$$= \frac{200}{100} = 2$$

As each sample comprises 20 items, the estimated proportion of defective items on the production line, p, is given by $p = \dfrac{2}{20} = 0.1$

b Assume the number of defective items in a sample is modelled by B(20, 0.1)

$$r = 100\times \mathrm{P}(X=2) = 100\times\binom{20}{2}(0.1)^2(0.9)^{18} = 28.517 = 28.5 \text{ (1 d.p.)}$$

The value r can also be found by using the tables for the binomial cumulative distribution function

$$r = 100\times\mathrm{P}(X=2) = 100\times\left(\mathrm{P}(X\leqslant2)-\mathrm{P}(X\leqslant1)\right)$$
$$= 100\times(0.6769-0.3917) = 100\times0.2852 = 28.5 \text{ (1 d.p.)}$$

The value s can be found using similar calculations or by using the fact that the total of the expected frequencies must be the same as the total of the observed frequencies, i.e. 100 in this case, so:

$$s = 100-(12.2+27.0+28.5+19.0+3.2+0.9+0.2) = 100-91 = 9.0 \text{ (1 d.p.)}$$

c H_0: A binomial distribution is a suitable model.
H_1: A binomial distribution is not a suitable model.

The observed and expected results are shown in the table. The results for 4, 5, 6 and 7 or more have been combined to ensure that all expected frequency values are greater than 5.

x	**0**	**1**	**2**	**3**	**⩾4**	**Total**
Observed (O_i)	17	31	19	14	19	100
Expected (E_i)	12.2	27.0	28.5	19.0	13.3	100
$\dfrac{(O_i-E_i)^2}{E_i}$	1.889	0.593	3.167	1.316	2.443	9.406

So the test statistic $X^2 = \sum\dfrac{(O_i-E_i)^2}{E_i} = 9.406$

The number of degrees of freedom $\nu = 3$ (five data cells with two constraints as p has been estimated by calculation and the total frequency must be 100)
From the tables, the critical value $\chi_3^2(5\%) = 7.815$

As the text statistic 9.406 is greater than the critical value 7.815, H_0 should be rejected at the 5% significance level. A binomial distribution is not a suitable model.

d Since a binomial distribution does not fit, the laws for binomial distribution cannot be true. Defective items on the production line do not occur independently *or* not with a constant probability.

2 a A suitable distribution to model the number of heads obtained from spinning five unbiased coins is a binomial distribution, B(5,0.5)

b H_0: $B(5, 0.5)$ is a suitable model.
H_1: $B(5, 0.5)$ is not a suitable model.

Total frequency $= 6+18+29+34+10+3 = 100$
This is one constraint on the test.

Now find the expected frequencies. This can be done using the tables or by using a calculator.

For example, for $x = 1$ and using the tables, the expected frequency is:
$100\times\left(\mathrm{P}(X \leqslant 1)-\mathrm{P}(X=0)\right)=100\times(0.1875-0.0312)=15.63$

Using a calculator, the respective calculations are:

Expected frequency for 0 heads $= 100\times\binom{5}{0}0.5^5 = 3.125$

Expected frequency for 1 head $= 100\times\binom{5}{1}0.5^5 = 15.625$

Expected frequency for 2 heads $= 100\times\binom{5}{2}0.5^5 = 31.25$

Expected frequency for 3 heads $= 100\times\binom{5}{3}0.5^5 = 31.25$

Expected frequency for 4 heads $= 100\times\binom{5}{4}0.5^5 = 15.625$

Expected frequency for 5 heads $= 100\times\binom{5}{5}0.5^5 = 3.125$

The observed and expected results are shown in the table. The results for 0 and 1 and for 4 and 5 have been combined to ensure that all expected frequency values are greater than 5.

x	0 or 1	2	3	4 or 5	Total
Observed (O_i)	24	29	34	13	100
Expected (E_i)	18.75	31.25	31.25	18.75	100
$\dfrac{(O_i-E_i)^2}{E_i}$	1.47	0.162	0.242	1.763	3.637

So the test statistic $X^2=\sum\dfrac{(O_i-E_i)^2}{E_i}=3.637$

The number of degrees of freedom $\nu = 3$ (four data cells with one constraint)
From the tables, the critical value $\chi_3^2(10\%) = 6.251$

As 3.637 < 6.251, there is insufficient evidence to reject H_0 at the 10% level. B(5, 0.5) is a suitable model. There is no evidence that the coins are biased.

3 a Mean number of cuttings that did not grow $= \dfrac{1\times21+2\times30+3\times20+4\times12+5\times3+6\times2+7\times1}{11+21+30+20+12+3+2+1}$

$$= \frac{223}{100} = 2.23$$

As each sample comprises 10 cuttings, the probability, p, of a randomly selected cutting not growing is given by $p = \dfrac{2.23}{10} = 0.223$

b The gardener's proposed model for the number of cuttings taken from a plant that do not grow is B(10,0.2). The expected frequencies for the number of cuttings that do not grow from 100 randomly selected plants can be found using tables or by direct calculation.

Using tables:

$r = 100\times \mathrm{P}(X=0) = 100\times0.1074 = 10.74$

$s = 100\times\left(\mathrm{P}(X\leqslant2)-\mathrm{P}(X\leqslant1)\right) = 100(0.6778-0.3758) = 30.20$

By calculation:

$r = 100\times(0.8)^{10} = 10.7374 = 10.74$ (2 d.p.)

$s = 100\times\begin{pmatrix}10\\2\end{pmatrix}(0.8)^8(0.2)^2 = 30.20$ (2 d.p.)

The final value, t, can be found by using the fact that the total of the expected frequencies must equal the total of the observed frequencies. Here it is 100.

$t = 100-(r+26.84+s+20.13+8.81)$

$= 100-(10.74+26.84+30.20+20.13+8.81)$

$= 100-96.72 = 3.28$

c H_0: B(10, 0.2) is a suitable model for the data.

H_1: B(10, 0.2) is not a suitable model for the data.

d Since $t<5$, the last two groups are combined to ensure that all expected frequencies are greater than 5. Thus, the number of degrees of freedom $\nu = 5-1=4$ (as there are 5 data cells and one constraint that the expected frequencies must sum to 100). Note that the parameter p is given and not calculated so this is not a constraint.

e Critical value $\chi^2_4(5\%) = 9.488$

The test statistic is less than the critical value ($4.17 < 9.488$), so there is insufficient evidence to reject H_0 at the 5% level. The binomial distribution with $p=0.2$ is a suitable model for the number of cuttings that do not grow.

4 H_0: A Poisson distribution is a suitable model.
H_1: A Poisson distribution is not a suitable model.

As λ is not given, it must be estimated from the observed frequencies.

$$\text{Mean} = \lambda = \frac{(1\times 65)+(2\times 22)+(3\times 12)+(4\times 2)}{99+65+22+12+2} = \frac{153}{200} = 0.765$$

Calculate the expected frequencies as follows:

$$E_0 = 200\times \mathrm{P}(X=0) = 200\times \frac{\mathrm{e}^{-0.765}\,0.765^0}{0!} = 93.06678\ldots$$

$$E_1 = 200\times \mathrm{P}(X=1) = 200\times \frac{\mathrm{e}^{-0.765}\,0.765^1}{1!} = 71.19609\ldots$$

$$E_2 = 200\times \mathrm{P}(X=2) = 200\times \frac{\mathrm{e}^{-0.765}\,0.765^2}{2!} = 27.23250\ldots$$

$$E_3 = 200\times \mathrm{P}(X=3) = 200\times \frac{\mathrm{e}^{-0.765}\,0.765^3}{3!} = 6.944288\ldots$$

$$E_{i\geq 4} = 200-(E_0+E_1+E_2+E_3) = 200-198.43965\ldots = 1.56034\ldots$$

Combine the classes for 3 and $\geqslant 4$ so that all the expected frequencies are greater than 5, and then calculate the test statistic

x	0	1	2	$\geqslant 3$	Total
Observed (O_i)	99	65	22	14	200
Expected (E_i)	93.0667	71.1960	27.2325	8.50463	200
$\dfrac{(O_i-E_i)^2}{E_i}$	0.3783	0.5392	1.0054	3.5509	5.474

The number of degrees of freedom $\nu = 4-2 = 2$ (four data cells with two constraints as λ is estimated by calculation)
From the tables: $\chi^2_2(5\%) = 5.991$

As $5.474 < 5.991$, there is insufficient evidence to reject H_0 at the 5% level. A Poisson distribution is a suitable model. The number of computer failures a day can be modelled by a Poisson distribution.

5 a H_0: Mathematics grades and English grades are independent.
H_1: Mathematics grades and English grades are not independent.

Another way to express the hypotheses is:
H_0: There is no association between Mathematics grades and English grades.
H_1: There is an association between Mathematics grades and English grades.

These are the observed frequencies (O_i) with totals for each row and column:

		Maths grades			
		A or B	**C or D**	**E or U**	**Total**
English grades	**A or B**	25	25	10	60
	C to U	5	30	15	50
	Total	30	55	25	110

Calculate the expected frequencies (E_i) for each cell. Show the working for at least expected frequency. For example:

Expected frequency 'Mathematics A or B and English A or B' $= \frac{60 \times 30}{110} = 16.364$

The expected frequencies (E_i) are:

		Maths grades		
		A or B	**C or D**	**E or U**
English grades	**A or B**	16.364	30	13.636
	C to U	13.636	25	11.364

The test statistic (X^2) calculations are:

O_i	E_i	$\frac{(O_i - E_i)^2}{E_i}$
25	16.364	4.5576
25	30	0.8333
10	13.636	0.9695
5	13.636	5.4694
30	25	1.0000
15	11.364	1.1637

$$X^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 13.993$$

The number of degrees of freedom $\nu = (2-1)(3-1) = 2$; from the tables: $\chi^2_2(10\%) = 4.605$

As 13.993 > 4.605, reject H_0 at the 10% level. There is evidence that the Mathematics and English results are not independent.

b By increasing the number of data cells, some may have expected frequencies less than 5, and this will require merging some rows and/or columns to perform the chi-squared test.

6 H_0: Gender and acceptance/rejection of a flu jab are independent.
H_1: Gender and acceptance/rejection of a flu jab are not independent.

Another way to express the hypotheses is:
H_0: There is no association between gender and acceptance/rejection of a flu jab.
H_1: There is an association between gender and acceptance/rejection of a flu jab.

These are the observed frequencies (O_i) with totals for each row and column:

		Accepted	**Rejected**	**Total**
Gender	**Male**	170	110	280
	Female	280	140	420
	Total	450	250	700

Calculate the expected frequencies (E_i) for each cell. For example:

Expected frequency 'Male and Rejected' $= \frac{280 \times 250}{700} = 100$

The expected frequencies (E_i) are:

		Accepted	**Rejected**
Gender	**Male**	180	100
	Female	270	150

The test statistic (X^2) calculations are:

O_i	E_i	$\frac{(O_i - E_i)^2}{E_i}$
170	180	0.555
110	100	1.000
280	270	0.370
140	150	0.667

$$X^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 2.593$$

The number of degrees of freedom $\nu = (2-1)(2-1) = 1$; from the tables: $\chi_1^2(5\%) = 3.841$

As 2.593 < 3.841, there is insufficient evidence to reject H_0 at the 5% level. There is no association between a person's gender and their acceptance or rejection of a flu jab.

7 H_0: Gender and type of course taken are independent.
H_1: Gender and type of course taken are not independent.

Another way to express the hypotheses is:
H_0: There is no association between the gender of student and the type of course taken.
H_1: There is an association between the gender of student and the type of course taken.

These are the observed frequencies (O_i) with totals for each row and column:

		Course			
		Arts	**Science**	**Humanities**	**Total**
Gender	**Boy**	30	50	35	115
	Girl	40	20	42	102
	Total	70	70	77	217

Calculate the expected frequencies (E_i) for each cell. For example:

Expected frequency 'Boy and Arts' $= \dfrac{115 \times 70}{217} = 37.09\ldots$

The expected frequencies (E_i) are:

		Arts	**Science**	**Humanities**
Gender	**Male**	37.1	37.1	40.8
	Female	32.9	32.9	36.2

The test statistic (X^2) calculations are:

O_i	E_i	$\dfrac{(O_i - E_i)^2}{E_i}$
30	37.1	1.359
50	37.1	4.485
35	40.8	0.825
40	32.9	1.532
20	32.9	5.058
42	36.2	0.929

$$X^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 14.188$$

The number of degrees of freedom $\nu = (2-1)(3-1) = 2$; from the tables: $\chi_2^2(1\%) = 9.210$

As 14.188 > 9.210, reject H_0 at the 1% level. There is evidence of an association between the gender of student and the type of course taken.

8 H_0: There is no association between the treatment of the trees and their survival.
H_1: There is an association between the treatment of the trees and their survival.

These are the observed frequencies (O_i) with totals for each row and column:

	No action	**Remove diseased branches**	**Spray with chemicals**	**Total**
Tree died within 1 year	10	5	6	21
Tree survived for 1–4 years	5	9	7	21
Tree survived beyond 4 years	5	6	7	18
Total	20	20	20	60

Calculate the expected frequencies (E_i) for each cell. For example:

Expected frequency 'Tree died within 1 year and No action' $= \frac{21 \times 20}{60} = 7$

The expected frequencies (E_i) are:

	No action	**Remove diseased branches**	**Spray with chemicals**
Tree died within 1 year	7	7	7
Tree survived for 1–4 years	7	7	7
Tree survived beyond 4 years	6	6	6

The test statistic (X^2) calculations are:

O_i	E_i	$\frac{(O_i - E_i)^2}{E_i}$
10	7	1.2857
5	7	0.5714
6	7	0.1429
5	7	0.5714
9	7	0.5714
7	7	0.0000
5	6	0.1667
6	6	0.0000
7	6	0.1667

$$X^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 3.476$$

The number of degrees of freedom $\nu = (3-1)(3-1) = 4$; from the tables: $\chi_4^2(5\%) = 9.488$

As $3.476 < 9.488$, there is insufficient evidence to reject H_0 at the 5% level. There is no evidence of an association between the treatment of the trees and their survival.

9 H_0: There is no association between age and colour preference.
H_1: There is an association between age and colour preference.

Calculate the expected frequencies (E_i) for each cell. For example:

Expected frequency 'Aged 4 and prefers blue' $= \frac{18 \times 22}{50} = 7.92$

The expected frequencies (E_i) are:

		Red	**Blue**
Age in years	**4**	10.08	7.92
	8	9.52	7.48
	12	8.4	6.6

The test statistic (X^2) calculations are:

O_i	E_i	$\frac{(O_i - E_i)^2}{E_i}$
12	10.08	0.3657
6	7.92	0.4655
10	9.52	0.0242
7	7.48	0.0308
6	8.4	0.6857
9	6.6	0.8727

$$X^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 2.4446$$

The number of degrees of freedom $\nu = (3-1)(2-1) = 2$; from the tables: $\chi_2^2(5\%) = 5.991$

As $2.4446 < 5.991$, there is insufficient evidence to reject H_0 at the 5% level. There is no association between age and colour preference.

10 H_0: Deliveries of mail are uniformly distributed.
H_1: Deliveries of mail are not uniformly distributed.

The celebrity thinks the deliveries of mail are uniformly delivered over the 6 days so all the expected frequencies = total observed frequencies (120) divided by 6, i.e. 20.

x	**Mon**	**Tues**	**Wed**	**Thurs**	**Fri**	**Sat**	**Total**
Observed (O_i)	20	15	18	23	19	25	120
Expected (E_i)	20	20	20	20	20	20	120
$\frac{(O_i - E_i)^2}{E_i}$	0	1.25	0.2	0.45	0.05	1.25	3.2

The degrees of freedom $\nu = 5$ (six data cells with one constraint); from the tables: $\chi_5^2(1\%) = 15.086$

As $3.2 < 15.086$, there is insufficient evidence to reject H_0 at the 1% level. There is no evidence to suggest that mail is not uniformly distributed.

11 a The sum of the observed frequencies $= 30+18+12+1+1 = 62$

As $X \sim \text{Geo}(0.4)$, calculate the expected frequencies using the equation:
$E_i = \text{P}(X = i) \times 62 = 0.4(0.6)^{i-1} \times 62 = 24.8(0.6)^{i-1}$

So $E_1 = 24.8(0.6)^0 = 24.8$
$E_2 = 24.8(0.6)^1 = 14.88$
$E_3 = 24.8(0.6)^2 = 8.928$
$E_4 = 24.8(0.6)^3 = 5.3586$
$E_{i \geq 5} = 62 - 24.8 = 14.88 = 8.928 - 5.3586 = 8.0352$

b H0: $X \sim \text{Geo}(0.4)$ is a suitable model.
H1: $X \sim \text{Geo}(0.4)$ is not a suitable model.

c The observed and expected frequencies and the calculations for the goodness of fit are:

n	**1**	**2**	**3**	**4**	**⩾5**	**Total**
Observed (O_i)	30	18	12	1	1	62
Expected (E_i)	24.8	14.88	8.928	5.3568	8.0352	62
$\frac{(O_i - E_i)^2}{E_i}$	1.0903	0.6542	1.0570	3.5435	6.1597	12.5047

There are 5 cells and 1 constraint (that the sum of the expected frequencies must equal 62).
So the number of degrees of freedom, $\nu = 5 - 1 = 4$
From the tables: $\chi_4^2(1\%) = 13.277$

As 12.5047 < 13.277, there is insufficient evidence to reject H0 at the 1% level. There is no evidence to suggest that the model is not suitable.

d *[impression 2 onwards]*
At the 2.5% significance level, the critical value is $\chi_4^2(2.5\%) = 11.143$
As 12.5047 > 11.143, reject H0 at the 2.5% level. The evidence suggests that Geo(0.4) is not a suitable model.

d *[In the first impression of the book, 2% was used instead of 2.5%]*
At the 2% significance level, the critical value is $\chi_4^2(2\%) = 11.668$
As 12.5047 > 11.668, reject H0 at the 2% level. The evidence suggests that Geo(0.4) is not a suitable model.

12 a Use the fact that $G_X(1) = 1$

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

$\Rightarrow k \times 6^2 = 1 \Rightarrow k = \frac{1}{36}$

b P(X = 1) is the coefficient of the t term in the probability function.

$G_X(t) = \frac{1}{36}(3+2t+t^2)(3+2t+t^2) = \frac{1}{36}(9+12t+10t^2+4t^3+t^4)$

$\Rightarrow P(X=1) = \frac{12}{36} = \frac{1}{3}$

13 a $X \sim \text{Geo}(0.24)$

b $P(X=7) = 0.24(1-0.24)^6 = 0.24 \times 0.76^6 = 0.0462$ (4 d.p.)

c The probability generating function for a geometric distribution is:

$G_X(t) = \frac{pt}{1-(1-p)t}$

So in this case, $G_X(t) = \frac{0.24t}{1-(1-0.24)t} = \frac{0.24t}{1-0.76t}$

d The random variable Z is the number of times Billy hits gold in 15 arrows at the target, so $Y \sim B(15, 0.24)$

The probability generating function for a binomial distribution is:

$G_Y(t) = (1-p+pt)^n$

So in this case, $G_Y(t) = (1-0.24+0.24t)^{15} = (0.76+0.24t)^{15}$

e The random variable Y is the number of shots it takes Billy to hit four golds, so $Z \sim \text{Neg B}(4, 0.24)$

The probability generating function for a negative binomial distribution is:

$G_Z(t) = \left(\frac{pt}{(1-(1-p)t}\right)^r$

So in this case, $G_Z(t) = \left(\frac{0.24t}{(1-(1-0.24)t}\right)^4 = \left(\frac{0.24t}{1-0.76t}\right)^4$

14 The random variable X can be modelled by a Poisson distribution, $X \sim \text{Po}(1.7)$

From the properties of the Poisson distribution, $\text{P}(X = x) = \dfrac{e^{-1.7}1.7^x}{x!}$, so

$$\begin{aligned}
\text{G}_X(t) &= \sum \text{P}(X = x)t^x \\
&= \sum \frac{\text{e}^{-1.7}1.7^x}{x!} \times t^x = \text{e}^{-1.7}\sum\frac{1.7^x t^x}{x!} = \text{e}^{-1.7}\sum\frac{(1.7t)^x}{x!} \\
&= \text{e}^{-1.7}\left(\frac{(1.7t)^0}{0!} + \frac{(1.7t)^1}{1!} + \frac{(1.7t)^2}{2!} + \frac{(1.7t)^3}{3!} + \ldots\right) \\
&= \text{e}^{-1.7}\left(1 + 1.7t + \frac{(1.7t)^2}{2!} + \frac{(1.7t)^3}{3!} + \ldots\right)
\end{aligned}$$

Compare the expression in brackets with the infinite series for $\mathbf{e}^{\mathbf{x}}$: $\text{e}^x = 1 + x + \dfrac{x^2}{2!} + \dfrac{x^3}{3!} + \ldots + \dfrac{x^r}{r!} + \ldots$

$$\begin{aligned}
\text{So } \text{G}_X(t) &= \text{e}^{-1.7} \times \text{e}^{1.7t} \\
&= \text{e}^{1.7t-1.7} \\
&= \text{e}^{1.7(t-1)} \qquad \text{(as required)}
\end{aligned}$$

15 a $\text{G}_X(t) = \dfrac{4}{(3-t)^2} = 4(3-t)^{-2}$

$\text{G}'_X(t) = 4(-2)(-1)(3-t)^{-3} = 8(3-t)^{-3}$

$\text{G}'_X(1) = 8(3-1)^{-3} = \dfrac{8}{2^3} = 1$

Hence, mean $= \text{E}(X) = \text{G}'_X(1) = 1$

$\text{G}''_X(t) = 8(-3)(-1)(3-t)^{-4} = 24(3-t)^{-4}$

$\text{G}''_X(1) = 24(3-1)^{-4} = \dfrac{24}{2^4} = 1.5$

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

$\text{Var}(X) = 1.5 + 1 - 1^2 = 1.5$

Standard deviation $= \sqrt{\text{Var}} = \sqrt{1.5} = 1.2247$ (4 d.p.)

b Using the general result $\text{G}_X{}^{(n)}(0) = n!\text{P}(x = n)$ gives:

$\text{P}(X = 0) = \text{G}_X(0) = \dfrac{4}{3^2} = \dfrac{4}{9}$

$\text{P}(X = 1) = \text{G}'_X(0) = \dfrac{8}{3^3} = \dfrac{8}{27}$

Alternatively find, respectively, the constant and the coefficient of t from the probability generating function:

$$\begin{aligned}
\text{G}_X(t) &= 4(3-t)^{-2} = 4\left(3^{-2}\right)\left(1-\tfrac{t}{3}\right)^{-2} = \tfrac{4}{9}\left(1+\left(-\tfrac{t}{3}\right)\right)^{-2} \\
&= \tfrac{4}{9}\left(1+(-2)\left(-\tfrac{t}{3}\right)+\ldots\right) \qquad \text{(using the binomial expansion)}
\end{aligned}$$

15 b (continued)

$\mathrm{P}(X=0)=\frac{4}{9}$

$\mathrm{P}(X=1)=\frac{4}{9}(-2)\left(-\frac{1}{3}\right)=\frac{8}{27}$

16 a Using the fact that $\mathrm{G}_X(1)=1$

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

So $4^2k=1\Rightarrow k=\frac{1}{16}$

b $\mathrm{G}_X(t)=\frac{1}{16}t^2\left(1+3t^2\right)^2=\frac{1}{16}t^2\left(1+6t^2+9t^4\right)=\frac{1}{16}\left(t^2+6t^4+9t^6\right)$

$\mathrm{P}(X=4)$ is the coefficient of t^4

So $\mathrm{P}(X=4)=\frac{6}{16}=\frac{3}{8}$

16 c $\mathrm{G}_X(t)=\frac{1}{16}\left(t^2+6t^4+9t^6\right)$

$\mathrm{G}'_X(t)=\frac{1}{16}\left(2t+24t^3+54t^5\right)$

$\mathrm{G}'_X(1)=\frac{1}{16}(2+24+54)=\frac{80}{16}=5$

$\mathrm{E}(X)=\mathrm{G}'_X(1)=5$

$\mathrm{G}''_X(t)=\frac{1}{16}\left(2+72t^2+270t^4\right)$

$\mathrm{G}''_X(1)=\frac{1}{16}(2+72+270)=\frac{344}{16}=21.5$

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

$=21.5+5-25=1.5$

d $\mathrm{G}_Y(t)=\left(\frac{1}{4}\right)^2(1+3t)^2=\frac{1}{16}\left(1+6t+9t^2\right)$

$\mathrm{G}'_Y(t)=\frac{1}{16}(6+18t)$

$\mathrm{E}(Y)=\mathrm{G}'_Y(1)=\frac{24}{16}=\frac{3}{2}=1.5$

As X and Y are independent $\mathrm{G}_{X+Y}(t)=\mathrm{G}_X(t)\times\mathrm{G}_Y(t)$

$\mathrm{G}_{X+Y}(t)=\frac{1}{16}\left(t^2+6t^4+9t^6\right)\times\frac{1}{16}\left(1+6t+9t^2\right)$

$=\frac{1}{256}\left(t^2+6t^3+15t^4+36t^5+63t^6+54t^7+81t^8\right)$

$\mathrm{G}'_{X+Y}(t)=\frac{1}{256}\left(2t+18t^2+60t^3+180t^4+378t^5+378t^6+648t^8\right)$

$\mathrm{E}(X+Y)=\mathrm{G}'_{X+Y}(1)=\frac{1}{256}(2+18+60+180+378+378+648)=\frac{1664}{256}=\frac{13}{2}=6.5$

17 a Using the fact that $G_X(1)=1$

$k(1+4+2)^2=1$

So $7^2k=1 \Rightarrow k=\frac{1}{49}$

b Expanding the probability generating function gives

$$\mathrm{G}_X(t)=\frac{1}{49}\left(t+4t^2+2t^3\right)\times\left(t+4t^2+2t^3\right)$$
$$=\frac{1}{49}\left(t^2+8t^3+20t^4+16t^5+4t^6\right)$$

So $\mathrm{P}(X=3)=\frac{8}{49}$

17 c $\mathrm{G}'_X(t)=\dfrac{1}{49}\left(2t+24t^2+80t^3+80t^4+24t^5\right)$

$$\mathrm{E}(X)=\mathrm{G}'_X(1)=\frac{1}{49}(2+24+80+80+24)=\frac{210}{49}=\frac{30}{7}$$

$$\mathrm{G}''_X(t)=\frac{1}{49}\left(2+48t+240t^2+320t^3+120t^4\right)$$

$$\mathrm{G}''_X(1)=\frac{1}{49}(2+48+240+320+120)=\frac{730}{49}$$

$$\mathrm{Var}(X)=\mathrm{G}''_X(1)+\mathrm{G}'_X(1)-(\mathrm{G}'_X(1))^2$$
$$=\frac{730}{49}+\frac{30}{7}-\left(\frac{30}{7}\right)^2=\frac{730+210-900}{49}=\frac{40}{49}$$

The values $\mathrm{G}'_X(1)$ and $\mathrm{G}''_X(1)$ can also be found by differentiating the probability generating function using the product rule:

$$\mathrm{G}_X(t)=\frac{1}{49}\left(t+4t^2+2t^3\right)^2$$

$$\mathrm{G}'_X(t)=\frac{2}{49}\left(t+4t^2+2t^3\right)\left(1+8t+6t^2\right)$$

$$\Rightarrow \mathrm{G}'_X(1)=\frac{2}{49}(1+4+2)(1+8+6)=\frac{2\times7\times15}{49}=\frac{30}{7}$$

$$\mathrm{G}''_X(t)=\frac{2}{49}\left(t+4t^2+2t^3\right)(8+12t)+\frac{2}{49}\left(1+8t+6t^2\right)\left(1+8t+6t^2\right)$$

$$\Rightarrow \mathrm{G}''_X(t)=\frac{2}{49}\left((7\times20)+(15\times15)\right)=\frac{730}{49}$$

d If $Y=aX+b,\ \mathrm{G}_Y(t)=t^b\mathrm{G}_X(t^a)$

$$\mathrm{G}_Y(t)=t^{-2}\mathrm{G}_X(t^3)$$
$$=t^{-2}\times\frac{1}{49}(t^3+4\left(t^3\right)^2+2\left(t^3\right)^3)^2$$
$$=\frac{1}{49}t^{-2}\left(t^3+4t^6+2t^9\right)^2$$

18 a i Type I error – H_0 is rejected, when it is in fact true.

ii Type II error – H_0 is accepted, when it is in fact false.

b $H_0: \lambda = 5$, $H_1: \lambda > 5$

Assume H_0, so that $X \sim \text{Po}(5)$
Using the Poisson cumulative distribution tables:
$P(X \geqslant 7 \mid \lambda = 5) = 1 - P(X \leqslant 6 \mid \lambda = 5) = 1 - 0.7622 = 0.2378$

As 0.2378 > 0.05, there is insufficient evidence to reject H0 at the 5% significance level. So there is no evidence of an increase in the number of chicks reared per year.

18 c Find the critical region for this test.
Assume H_0, so that $X \sim \text{Po}(5)$
Significance level 5%, so require c such that $P(X \geqslant c) > 0.05$
From the Poisson cumulative distribution tables
$P(X \geqslant 9) = 1 - P(X \leqslant 8) = 1 - 0.9319 = 0.0681$
$P(X \geqslant 10) = 1 - P(X \leqslant 9) = 1 - 0.9682 = 0.0318$
As $P(X \geqslant 9) > 0.05$ and $P(X \geqslant 10) < 0.05$, the critical value is 10
Hence the critical region is $X \geqslant 10$
So $P(\text{Type I error}) = P(X \geqslant 10) = 0.0318$

d $P(\text{Type II error}) = P(X \leqslant 9 \mid \lambda = 8) = 0.7166$

19 a Let the random variable $\bar{X}$ be the mean weight of a block of butter from the random sample of 15 blocks.

Then $\bar{X} \sim N\left(250, \frac{4^2}{15}\right)$

Standardise the $\bar{X}$ variable

$$Z = \frac{\bar{X} - 250}{\frac{4}{\sqrt{15}}} = \frac{\sqrt{15}}{4}(\bar{X} - 250)$$

Significance level 2%, so require 1% in each tail

From the tables, the critical region for Z is $Z > 2.3263$ or $Z < -2.3263$

So the critical values for $\bar{X}$ are given by

$\sqrt{15}(\bar{X} - 250) = \pm(4 \times 2.3263) = \pm 9.3052$

$\Rightarrow \bar{X} = 247.60$ and $\bar{X} = 252.40$ (to 2 d.p.)

So the critical region for this test is $\bar{X} < 247.60$ or $\bar{X} > 252.40$

b Using the normal cumulative distribution function on a calculator:

$$\begin{aligned} P(\text{Type II error}) &= P(247.6 < \bar{X} < 252.4 \mid \mu = 254) \\ &= P(\bar{X} < 252.4 \mid \mu = 254) - P(\bar{X} < 247.6 \mid \mu = 42) \\ &= 0.0607 - 0.000 = 0.0607 \text{ (4 d.p.)} \end{aligned}$$

20 a $H_0 : \lambda = 8 \quad H_1 : \lambda \neq 8$

Assume H_0, so that $X \sim \text{Po}(8)$

Significance level 10%

If $X = c_1$ is the upper boundary of the lower critical region, require $P(X \leqslant c_1)$ to be as close as possible to 5%

From the tables

$P(X \leqslant 3) = 0.0424$ and $P(X \leqslant 4) = 0.0996$

0.0424 is closer to 0.05, so $c_1 = 3$ and the lower critical region is $X \leqslant 3$

If $X = c_2$ is the lower boundary of the upper critical region, require $P(X \geqslant c_2)$ to be as close as possible to 5%

From the tables

$P(X \geqslant 13) = 1 - P(X \leqslant 12) = 1 - 0.9362 = 0.0638$

and $P(X \geqslant 14) = 1 - P(X \leqslant 13) = 1 - 0.9658 = 0.0342$

0.0638 is closer to 0.05, so $c_2 = 13$ and the upper critical region is $X \geqslant 13$

Critical region is $X \leqslant 3$ or $X \geqslant 13$

b i
$$\begin{aligned} P(\text{Type II error}) &= P(4 \leqslant X \leqslant 12 \mid \lambda = 10) \\ &= P(X \leqslant 12 \mid \lambda = 10) - P(X \leqslant 3 \mid \lambda = 10) \\ &= 0.7916 - 0.0103 = 0.7813 \end{aligned}$$

ii Power $= 1 - P(\text{Type II error}) = 1 - 0.7813 = 0.2187$

21 a Let the random variable X be the number of trains arriving late from the sample of 12 trains, so the null hypothesis is $X \sim B(12, 0.1)$

$$\begin{aligned} \text{Size} &= P(\text{Type I error}) = P(X \geqslant 3 \mid p = 0.1) \\ &= 1 - P(X \leqslant 2 \mid p = 0.1) = 1 - 0.8891 = 0.1109 \end{aligned}$$

b
$$\begin{aligned} \text{Power function} &= P(\text{reject } H_0 \text{ when it is false}) \\ &= P(X \geqslant 3 \mid X \sim B(12, p)) = 1 - P(X \leqslant 3 \mid X \sim B(12, p)) \\ &= 1 - P(X = 0 \mid X \sim B(12, p)) - P(X = 1 \mid X \sim B(12, p)) - P(X = 2 \mid X \sim B(12, p)) \\ &= 1 - (1-p)^{12} - 12p(1-p)^{11} - 66p^2(1-p)^{10} \\ &= 1 - (1-p)^{10}((1-p)^2 + 12p(1-p) + 66p^2) \\ &= 1 - (1-p)^{10}(1 - 2p + p^2 + 12p - 12p^2 + 66p^2) \\ &= 1 - (1-p)^{10}(1 + 10p + 55p^2) \end{aligned}$$

21 c i Power of test ($p = 0.2$) $= 1-(0.8)^{10}(1+2+2.2) = 1-0.01737\times 5.2$
$= 1-0.5583 = 0.4417$ (4 d.p.)

ii Power of test ($p = 0.6$) $= 1-(0.4)^{10}(1+6+19.8) = 1-0.0001049\times 26.8$
$= 1-0.0028 = 0.9972$ (4 d.p.)

d The second test with the larger value ($p = 0.6$) is more discriminating (powerful).

22 a i The power of the test is defined as the probability of rejecting the null hypothesis when it is in fact false.

Note that a Type II error is accepting the null hypothesis when it is false.
So the power of a test = 1 – P(Type II error).

The power of a test is also the probability of a result being in the critical region when the null hypothesis is false.

ii The size of a test is the probability of rejecting the null hypothesis when it is in fact true (a Type I error).

b $X \sim \mathrm{B}(8, 0.25)$
Size $= \mathrm{P}(\text{Type I error}) = \mathrm{P}(X > 6) = 1-\mathrm{P}(X \le 6 \mid X \sim \mathrm{B}(8, 0.25))$
$= 1-0.9996 = 0.0004$

c Power function $= \mathrm{P}(X > 6 \mid X \sim \mathrm{B}(8, p))$

$$= \mathrm{P}(X = 7 \mid X \sim \mathrm{B}(8, p)) + \mathrm{P}(X = 8 \mid X \sim \mathrm{B}(8, p))$$
$$= \frac{8}{7!1!}p^7(1-p)+p^8$$
$$= 8p^7 - 8p^8 + p^8$$
$$= 8p^7 - 7p^8$$

d Power $= 8\times 0.3^7 - 7\times 0.3^8 = 0.00129$ (3 s.f.)

e Power $= 1-\mathrm{P}(\text{Type II error})$
So $\mathrm{P}(\text{Type II error}) = 1-\text{power} = 1-0.00129 = 0.9987$ (4 d.p.)

f Increase the probability of a Type I error, e.g. increase the significance level of the test. (This test has an actual significance level of 0.42%.) Increase the value of p.

23 a Size $= \mathrm{P}(\text{Type I error}) = \mathrm{P}(\text{rejecting } \mathrm{H}_0 \text{ when it is true})$

$$= \mathrm{P}(X \leqslant 2 \mid X \sim \mathrm{Po}(4)) + \mathrm{P}(3 \leqslant X \leqslant 4 \mid X \sim \mathrm{Po}(4)) \times \mathrm{P}(X \leqslant 2 \mid X \sim \mathrm{Po}(4))$$
$$= \mathrm{P}(X \leqslant 2 \mid X \sim \mathrm{Po}(4)) + \big(\mathrm{P}(X \leqslant 4 \mid X \sim \mathrm{Po}(4)) - \mathrm{P}(X \leqslant 2 \mid X \sim \mathrm{Po}(4))\big) \times \mathrm{P}(X \leqslant 2 \mid X \sim \mathrm{Po}(4))$$
$$= 0.2381 + (0.6288 - 0.2381)\times 0.2381$$
$$= 0.2381 + 0.0930 = 0.3311$$

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

$$= P(X \leqslant 2 \mid X \sim Po(\lambda)) + P(3 \leqslant X \leq 4 \mid X \sim Po(\lambda)) \times P(X \leqslant 2 \mid X \sim Po(\lambda))$$
$$= P(X=0) + P(X=1) + P(X=2) + \big(P(X=3) + P(X=4)\big) \times \big(P(X=0) + P(X=1) + P(X=2)\big)$$
$$= e^{-\lambda}\left(\frac{\lambda^0}{0!} + \frac{\lambda^1}{1!} + \frac{\lambda^2}{2!}\right) + e^{-\lambda}\left(\frac{\lambda^3}{3!} + \frac{\lambda^4}{4!}\right) \times e^{-\lambda}\left(\frac{\lambda^0}{0!} + \frac{\lambda^1}{1!} + \frac{\lambda^2}{2!}\right)$$
$$= e^{-\lambda}\left(1 + \lambda + \frac{\lambda^2}{2}\right) + e^{-\lambda}\left(\frac{\lambda^3}{6} + \frac{\lambda^4}{24}\right) \times e^{-\lambda}\left(1 + \lambda + \frac{\lambda^2}{2}\right)$$
$$= e^{-\lambda}\left(1 + \lambda + \frac{\lambda^2}{2}\right) + e^{-2\lambda}\left(\frac{\lambda^3}{6} + \frac{\lambda^4}{24}\right)\left(1 + \lambda + \frac{\lambda^2}{2}\right)$$
$$= e^{-2\lambda} \times e^{\lambda}\left(1 + \lambda + \frac{\lambda^2}{2}\right) + e^{-2\lambda}\left(\frac{\lambda^3}{6} + \frac{\lambda^4}{24}\right)\left(1 + \lambda + \frac{\lambda^2}{2}\right)$$
$$= e^{-2\lambda}\left(1 + \lambda + \frac{\lambda^2}{2}\right)\left(e^{\lambda} + \frac{\lambda^3}{6} + \frac{\lambda^4}{24}\right)$$

c Find power using the power function

$$\text{Power } = e^{-6}\left(1 + 3 + \frac{9}{2}\right)\left(e^3 + \frac{3^3}{6} + \frac{3^4}{24}\right) = e^{-6} \times 8.5 \times 27.9605 = 0.5891$$

Alternatively, find the power directly

Power = P(H_0 rejected | $\lambda = 3$)

$$= P(X \leqslant 2 \mid X \sim Po(3)) + P(3 \leqslant X \leqslant 4 \mid X \sim Po(3)) \times P(X \leqslant 2 \mid X \sim Po(3))$$
$$= P(X \leqslant 2 \mid X \sim Po(3)) + \big(P(X \leqslant 4 \mid X \sim Po(3)) - P(X \leqslant 2 \mid X \sim Po(3))\big) \times P(X \leqslant 2 \mid X \sim Po(3))$$
$$= 0.4232 + (0.8153 - 0.4232) \times 0.4232$$
$$= 0.4232 + 0.1659 = 0.5891$$

Power = 1 − P(Type II error)

P(Type II error) = 1 − 0.5891 = 0.4109

24 a H_0: $p = 0.35$ H_1: $p \neq 0.35$

b Let the random variable X = number of people cured, then H_0 is $X \sim B(20, 0.35)$

$$P(\text{Type I error}) = P(X \leqslant 3 \mid X \sim B(20, 0.35)) + P(X \geqslant 11 \mid X \sim B(20, 0.35))$$
$$= P(X \leqslant 3 \mid X \sim B(20, 0.35)) + 1 - P(X \leqslant 10 \mid X \sim B(20, 0.35))$$
$$= 0.0444 + (1 - 0.9468) = 0.0444 + 0.0532 = 0.0976$$

c $r = P(\text{Type II error} \mid X \sim B(20, 0.3)) = P(4 \leqslant X \leqslant 10 \mid X \sim B(20, 0.3))$

$$= P(X \leqslant 10 \mid X \sim B(20, 0.3)) - P(X \leqslant 3 \mid X \sim B(20, 0.3))$$
$$= 0.9829 - 0.1071 = 0.8758$$

$s = P(\text{Type II error} \mid X \sim B(20, 0.5)) = P(4 \leqslant X \leqslant 10 \mid X \sim B(20, 0.5))$

$$= P(X \leqslant 10 \mid X \sim B(20, 0.5)) - P(X \leqslant 3 \mid X \sim B(20, 0.5))$$
$$= 0.5881 - 0.0013 = 0.5868$$

24 d Power $= 1 - \text{P(Type II error)}$

$p = 0.2$, power $= 1 - 0.5880 = 0.4120$

$p = 0.4$, power $= 1 - 0.8565 = 0.1435$

e This is not a very powerful test, as the power is less than 0.5 for all values of p in the table. This means the probability of coming to the wrong conclusion is greater than the probability of coming to the correct conclusion. The power of the test does get greater as p gets further from 0.35.

25 a Size of test = P(Type I error) = P(reject H_0 when H_0 is true)

$$= P(X \geqslant 2 \mid X \sim B(5, 0.05)) = 1 - P(X \leqslant 1 \mid X \sim B(5, 0.05))$$
$$= 1 - 0.9774 = 0.0226$$

b Power function = P(reject H_0 when H_0 is false)

$$= P(X \geqslant 2 \mid X \sim B(5, p)) = 1 - P(X \leqslant 1 \mid X \sim B(5, p))$$
$$= 1 - \left(P(X = 0 \mid X \sim B(5, p)) + P(X = 1 \mid X \sim B(5, p))\right)$$
$$= 1 - \left((1-p)^5 + 5p(1-p)^4\right)$$
$$= 1 - (1-p)^4((1-p) + 5p)$$
$$= 1 - (1-p)^4(1+4p)$$

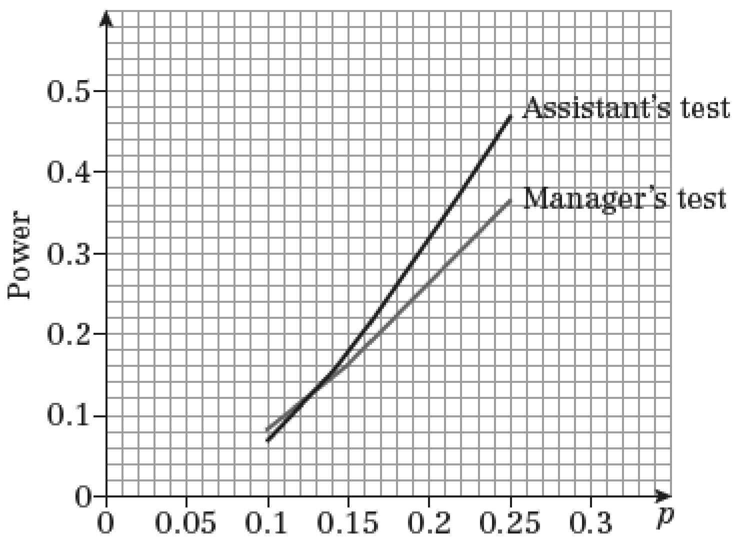
c P(Type I error) = P(reject H_0 when H_0 is true)

$$= P(X \geqslant 3 \mid X \sim B(10, 0.05)) = 1 - P(X \leqslant 2 \mid X \sim B(10, 0.05))$$
$$= 1 - 0.9885 = 0.0115$$

d a = P(reject $H_0 \mid X \sim B(10, 0.25)$)

$$= P(X \geqslant 3 \mid X \sim B(10, 0.25)) = 1 - P(X \leqslant 2 \mid X \sim B(10, 0.25))$$
$$= 1 - 0.5256 = 0.47 \text{ (2 d.p.)}$$

e Use the power function from part **b** to work out the power of the manager's test for p = 0.1, 0.15, 0.2 and 0.25. Plot the respective results for the manager's test and the assistant's test.



25 f For $p > 0.2$, the graph in part **e** shows that the assistant's test is more powerful than the manager's test, so the assistant's test is recommended for use.

g The manager's test only involves 5 bags of flour. This is a smaller sample than the assistant's test, and smaller samples are cheaper and quicker to conduct and involve less manpower. The manager may also believe that the true probability is closer to 0.05, and the graph shows that there is little difference in the power of the two tests as p gets close to 0.05.

Challenge

1 a Let $X \sim \mathrm{N}(360, 20^2)$

$r = 100 \times \mathrm{P}(350 < X < 360) = 100 \times \left(\mathrm{P}(X < 360) - \mathrm{P}(X < 350)\right)$

$= 100(0.5 - 0.3085) = 19.15$

Determine s either by using a similar method, by symmetry, or by using the fact that the expected frequencies must sum to 100:

$s = 100 - 2.28 - 13.59 - 14.98 - 19.15 - 14.98 - 13.59 - 2.28$

$= 100 - 80.85 = 19.15$

b $\mathrm{H_0}$: $\mathrm{N} \sim (360, 20^2)$ is a suitable model.

$\mathrm{H_1}$: $\mathrm{N} \sim (360, 20^2)$ is not a suitable model.

The observed and expected results and the calculation of the test statistic are shown in the table. The first two results and the last two results have been combined to ensure that all expected frequency values are greater than 5.

x	<340	340–	350–	360–	370–	>380	Total
Observed (O_i)	10	28	20	16	18	8	100
Expected (E_i)	15.87	14.98	19.15	19.15	14.98	15.87	100
$\dfrac{(O_i - E_i)^2}{E_i}$	12.171	11.316	0.038	0.518	0.609	3.903	18.555

The number of degrees of freedom $\nu = 5$ (six data cells with one constraint, that total frequency must be 100, the values for μ and σ are given)

From the tables, the critical value $\chi_5^2(1\%) = 15.086$

As 18.555 > 15.086, $\mathrm{H_0}$ should be rejected at the 1% significance level. The distribution $\mathrm{N} \sim (360, 20^2)$ is not a suitable model for the data.

2 a i $\mathrm{E}(S) = \mathrm{G}'_S(1) \qquad \mathrm{E}(N) = \mathrm{G}'_N(1) \qquad \mathrm{E}(X) = \mathrm{G}'_X(1)$

$\mathrm{G}_S(t) = \mathrm{G}_N(\mathrm{G}_X(t))$

$\mathrm{G}'_S(t) = \mathrm{G}'_N(\mathrm{G}_X(t)) \times \mathrm{G}'_X(t)$

$\mathrm{E}(S) = \mathrm{G}'_S(1) = \mathrm{G}'_N(\mathrm{G}_X(1)) \times \mathrm{G}'_X(1)$

$= \mathrm{G}'_N(1)\mathrm{G}'_X(1)$ as $\mathrm{G}_X(1) = 1$

$= \mathrm{E}(N)\mathrm{E}(X)$ as required

Challenge

2 a ii $\mathrm{G}_S(t)=\mathrm{G}_N(\mathrm{G}_X(t))$

$\mathrm{G}'_S(t)=\mathrm{G}'_N(\mathrm{G}_X(t))\times\mathrm{G}'_X(t)$

$\mathrm{G}''_S(t)=\mathrm{G}'_X(t)\times\mathrm{G}''_N(\mathrm{G}_X(t))\mathrm{G}'_X(t)+\mathrm{G}''_X(t)\mathrm{G}'_N(\mathrm{G}_X(t))$ (using the product rule)

$\mathrm{G}''_S(1)=\mathrm{G}'_X(1)\times\mathrm{G}''_N(\mathrm{G}_X(1))\mathrm{G}'_X(1)+\mathrm{G}''_X(1)\mathrm{G}'_N(\mathrm{G}_X(1))$

$=(\mathrm{G}_X(1))^2\times\mathrm{G}''_N(1)+\mathrm{G}''_X(1)\mathrm{G}'_N(1)$

$=(\mathrm{E}(X))^2\mathrm{G}''_N(1)+\mathrm{G}''_X(1)\mathrm{G}'_N(1)$

$=(\mathrm{E}(X))^2\mathrm{G}''_N(1)+\mathrm{G}''_X(1)\mathrm{E}(N)$

$\mathrm{Var}(S)=\mathrm{G}''_S(1)+\mathrm{G}'_S(1)-(\mathrm{G}'_S(1))^2$

$=(\mathrm{E}(X))^2\mathrm{G}''_N(1)+\mathrm{G}''_X(1)\mathrm{E}(N)+\mathrm{E}(S)-(\mathrm{E}(S))^2$

$=(\mathrm{E}(X))^2\mathrm{G}''_N(1)+\mathrm{G}''_X(1)\mathrm{E}(N)+\mathrm{E}(N)\mathrm{E}(X)-(\mathrm{E}(N)\mathrm{E}(X))^2$ (using result from part i)

$=\mathrm{E}(N)\left(\mathrm{G}''_X(1)+\mathrm{E}(X)\right)+(\mathrm{E}(X))^2\left(\mathrm{G}''_N(1)-(\mathrm{E}(N)^2\right)$

$=\mathrm{E}(N)\left(\mathrm{G}''_X(1)+\mathrm{E}(X)\right)+(\mathrm{E}(X))^2\left(\mathrm{G}''_N(1)-(\mathrm{E}(N)^2\right)-\mathrm{E}(N)(\mathrm{E}(X))^2+\mathrm{E}(N)(\mathrm{E}(X)^2$

(these last two terms add to zero)

$=\mathrm{E}(N)\left(\mathrm{G}''_X(1)+\mathrm{E}(X)-(\mathrm{E}(X))^2\right)+(\mathrm{E}(X))^2\left(\mathrm{G}''_N(1)-(\mathrm{E}(N))^2+\mathrm{E}(N)\right)$

$=\mathrm{E}(N)\left(\mathrm{G}''_X(1)+\mathrm{E}(X)-(\mathrm{E}(X))^2\right)+(\mathrm{E}(X))^2\left(\mathrm{G}''_N(1)+\mathrm{E}(N)-(\mathrm{E}(N))^2\right)$

$=\mathrm{E}(N)\mathrm{Var}(X)+(\mathrm{E}(X))^2\mathrm{Var}(N)$

Alternatively:

$\mathrm{LHS}=\mathrm{Var}(S)=\mathrm{G}''_S(1)+\mathrm{G}'_S(1)-(\mathrm{G}'_S(1))^2$

$=\mathrm{G}''_S(1)+\mathrm{E}(S)-(\mathrm{E}(S))^2$

$=\mathrm{G}''_S(1)+\mathrm{E}(N)\mathrm{E}(X)-(\mathrm{E}(N)\mathrm{E}(X))^2$

$=\mathrm{G}''_S(1)+\mathrm{E}(N)\mathrm{E}(X)-(\mathrm{E}(N))^2(\mathrm{E}(X))^2$

$=(\mathrm{E}(X))^2\mathrm{G}''_N(1)+\mathrm{G}''_X(1)\mathrm{E}(N)+\mathrm{E}((N)\mathrm{E}(X)-(\mathrm{E}(N))^2(\mathrm{E}(X))^2$ (using result for $\mathrm{G}''_S(1)$)

$=\mathrm{E}(N)\mathrm{G}''_X(1)+\mathrm{E}(N)\mathrm{E}(X)+(\mathrm{E}(X))^2\mathrm{G}''_N(1)-(\mathrm{E}(X))^2(\mathrm{E}(N))^2$ (equation I)

$\mathrm{RHS}=\mathrm{E}(N)\mathrm{Var}(X)+(\mathrm{E}(X))^2\mathrm{Var}(N)$

$=\mathrm{E}(N)\left(\mathrm{G}''_X(1)+\mathrm{E}(X)-(\mathrm{E}(X))^2\right)+(\mathrm{E}(X))^2\left(\mathrm{G}''_N(1)+\mathrm{E}(N)-(\mathrm{E}(N))^2\right)$

(multiply out brackets and cancel like terms)

$=\mathrm{E}(N)\mathrm{G}''_X(1)+\mathrm{E}(N)\mathrm{E}(X)+(\mathrm{E}(X))^2\mathrm{G}''_N(1)-(\mathrm{E}(X))^2(\mathrm{E}(N))^2$ (equation II)

As equation I = equation II, LHS = RHS, hence $\mathrm{Var}(S)=\mathrm{E}(N)\mathrm{Var}(X)+(\mathrm{E}(X))^2\mathrm{Var}(N)$

Challenge

2 b Let the random variable N denote the number of people who use the external cash machine each hour, so $N \sim \text{Po}(\lambda)$ and $\text{E}(N) = \text{Var}(N) = \lambda$

Let the random variable X denote the number of balance enquiries made by a customer at a visit, so, $X \sim \text{B}(1, p)$, with $\text{E}(X) = np = p$ and $\text{Var}(X) = np(1-p) = p(1-p)$

Let the random variable S denote the total number of balance enquiries made in an hour, so,
$\text{E}(S) = \text{E}(N)\text{E}(X) = \lambda p$ (using the result from part a i)

$$\begin{aligned}\text{Var}(S) &= \text{E}(N)\text{Var}(X) + (\text{E}(X))^2\,\text{Var}(N) \quad \text{(using the result from part a ii)}\\ &= \lambda p(1-p) + p^2\lambda\\ &= \lambda p - \lambda p^2 + \lambda p^2\\ &= \lambda p\end{aligned}$$

The trials (customers) are independent, occur randomly and at a fixed rate.
$\text{E}(S) = \text{Var}(S) = \lambda p$ so $S \sim \text{Po}(\lambda p)$

c Let the random variable M denote the number of people who use the internal cash machine each hour, so $M \sim \text{B}(n, q)$, with $\text{E}(M) = nq$ and $\text{Var}(M) = nq(1-q)$

Let the random variable Y denote the number of balance enquiries made by a customer at the internal cash machine, so, $Y \sim \text{B}(1, p)$, with $\text{E}(Y) = np = p$ and $\text{Var}(Y) = np(1-p) = p(1-p)$

Let the random variable T denote the total number of balance enquiries made in an hour at the internal cash machine, so, $\text{E}(T) = \text{E}(M)\text{E}(Y) = nqp$ (using the result from part a i)

$$\begin{aligned}\text{Var}(T) &= \text{E}(M)\text{Var}(Y) + (\text{E}(Y))^2\,\text{Var}(M) \quad \text{(using the result from part a ii)}\\ &= nqp(1-p) + p^2nq(1-q)\\ &= nqp - nqp^2 + nqp^2 - p^2nq^2\\ &= nqp(1-pq)\end{aligned}$$

Trials are independent. The number of trials is fixed. The probability of success is constant.
So $S \sim \text{B}(n, pq)$

Challenge

2 d i External cash machine: $N \sim \text{Po}(75)$, so $\text{E}(N) = \text{Var}(N) = 75$

Let the random variable C represent the amount of money withdrawn from a cash machine by a single customer and the random variable D represent the amount of money withdrawn from the external cash machine in an hour.

$$\text{E}(C) = \sum c\,\text{P}(C = c) = 0 \times 0.1 + 10 \times 0.3 + 20 \times 0.4 + 50 \times 0.2$$
$$= 0 + 3 + 8 + 10 = 21$$

Using result from part **a i**

Mean $= \text{E}(D) = \text{E}(N)\text{E}(C) = 75 \times 21 = £1575$

$$\text{Var}(C) = \text{E}(C^2) - (\text{E}(C))^2 = \sum c^2\,\text{P}(C = c) - (\text{E}(C))^2$$
$$= 0^2 \times 0.1 + 10^2 \times 0.3 + 20^2 \times 0.4 + 50^2 \times 0.2 - 21^2$$
$$= 30 + 160 + 500 - 441 = 249$$

Using result from part **a ii**

$$\text{Var}(D) = \text{E}(N)\text{Var}(C) + (\text{E}(C))^2\,\text{Var}(N)$$
$$= 75 \times 249 + 21^2 \times 75 = 51750$$

Standard deviation $= \sqrt{\text{Var}(D)} = \sqrt{51\,750} = 227.49 = £227$ (to the nearest pound)

ii Internal cash machine: $M \sim \text{B}(80, 0.25)$, so $\text{E}(M) = nq = 20$ and $\text{Var}(M) = nq(1-q) = 15$

Let the random variable F represent the amount of money withdrawn from the internal cash machine in an hour.

Using result from parts **a i** and **d i**

Mean $= \text{E}(F) = \text{E}(M)\text{E}(C) = 20 \times 21 = £420$

Using result from parts **a ii** and **d i**

$$\text{Var}(F) = \text{E}(M)\text{Var}(C) + (\text{E}(C))^2\,\text{Var}(M)$$
$$= 20 \times 249 + 21^2 \times 15 = 11595$$

Standard deviation $= \sqrt{\text{Var}(F)} = \sqrt{11\,595} = 107.68 = £108$ (to the nearest pound)