Chi-squared tests Mixed exercise 6

1 $P(Y < y) = 1 - P(Y > y)$

So $P(Y < y) = 0.99 \Rightarrow P(Y > y) = 0.01$

$\chi^2_{10}(1\%) = 23.209$, so $P\left(\chi^2_{10} > 23.209\right) = 0.01 \Rightarrow y = 23.209$

2 $\chi^2_8(5\%) = 15.507$, so $P(\chi^2_8 > 15.507) = 0.05 \Rightarrow x = 15.507$

3 Degrees of freedom $= (5-1)\times(3-1) = 8$

From the tables: $\chi^2_8(5\%) = 15.507$

Critical region is $\chi^2 > 15.507$

4 Amalgamation gives a 3×4 contingency table.

Degrees of freedom $= (4-1)\times(3-1) = 6$

Critical value is $\chi^2_6(5\%) = 12.592$

5 H_0: There is no association between catching a cold and taking the new drug.
H_1: There is an association between catching a cold and taking the new drug.

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

	Cold	No cold	Total
Drug	34	66	100
Dummy pill	45	55	100
Total	79	121	200

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

Expected frequency 'Cold' and 'Taken drug' $= \dfrac{100\times 79}{200} = 39.5$

The expected frequency and test statistic (X^2) calculations are:

O_i	E_i	$\dfrac{(O_i - E_i)^2}{E_i}$
34	39.5	0.766
66	60.5	0.5
45	39.5	0.766
55	60.5	0.5

$$X^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 2.53$$

The number of degrees of freedom $\nu = (2-1)(2-1) = 1$; from the tables: $\chi^2_1(5\%) = 3.841$

As 2.53 is less than 3.841, there is insufficient evidence to reject H_0 at the 5% level. It appears taking the new drug doesn't affect the chance of a person catching a cold.

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

Total frequency $= 38 + 32 + 10 = 80$

$$\text{Mean} = \lambda = \frac{1 \times 32 + 2 \times 10}{80} = \frac{52}{80} = 0.65$$

Calculate the expected frequencies as follows:

$$E_0 = 80 \times P(X = 0) = 80 \times \frac{e^{-0.65}\, 0.65^0}{0!} = 41.764$$

$$E_1 = 80 \times P(X = 1) = 80 \times \frac{e^{-0.65}\, 0.65^1}{1!} = 27.146$$

$$E_2 = 80 \times P(X = 2) = 80 \times \frac{e^{-0.65}\, 0.65^2}{2!} = 8.823$$

$$E_{i>2} = 80 - (41.764 + 27.146 + 8.823) = 2.267$$

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

Number of breakdowns	**0**	**1**	**$\geqslant 2$**	**Total**
Observed (O_i)	38	32	10	80
Expected (E_i)	41.764	27.146	11.090	80
$\dfrac{(O_i - E_i)^2}{E_i}$	0.339	0.868	0.107	1.314

The number of degrees of freedom $\nu = 1$ (three data cells with two constraints as λ is estimated by calculation)

From the tables: $\chi_1^2(5\%) = 3.841$

As 1.314 is less than 3.841, there is insufficient evidence to reject H_0 at the 5% level. The data may be modelled by a Poisson distribution.

7 H_0: There is no association between gender and passing a driving test at the first attempt.
H_1: There is an association between gender and passing a driving test at the first attempt.

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

	Pass	**Fail**	**Total**
Male	23	27	50
Female	32	18	50
Total	55	45	100

The expected frequency and test statistic (X^2) calculations are:

O_i	E_i	$\frac{(O_i - E_i)^2}{E_i}$
23	27.5	0.736
27	22.5	0.9
32	27.5	0.736
18	22.5	0.9

$$X^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 3.272$$

The number of degrees of freedom $\nu = (2-1)(2-1) = 1$; from the tables: $\chi_1^2(10\%) = 2.705$

As 3.27 is greater than 2.705, reject H_0 at the 10% level. Conclude there is evidence of an association between gender and passing a driving test at the first attempt.

8 a We would expect each box to have an equal chance of being opened, and so would expect each box to have been opened 20 times.

b 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 observed and expected results are:

Box number	**1**	**2**	**3**	**4**	**5**
Observed (O_i)	20	16	25	18	21
Expected (E_i)	20	20	20	20	20
$\frac{(O_i - E_i)^2}{E_i}$	0	0.8	1.25	0.2	0.05

$$X^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 2.3$$

Degrees of freedom $\nu = 4$ (five data cells with a single constraint); from the tables: $\chi_4^2(5\%) = 9.488$

As 2.3 is less than 9.488, there is insufficient evidence to reject H_0 at the 5% level. The data may be modelled by a discrete uniform distribution.

9 a Total number of dead flies $= 0 \times 1 + 1 \times 1 + 2 \times 5 + 3 \times 11 + 4 \times 24 + 5 \times 8 = 180$

Total number of flies sprayed $= 50 \times 5 = 250$

So $\text{P(fly dies when sprayed)} = \dfrac{180}{250} = 0.72$

b H_0: A B(5, 0.72) distribution is a suitable model for the data.
H_1: The data cannot be modelled by a B(5, 0.72) distribution.

Find the expected frequencies by multiplying the total frequency 50 samples by the probability $P(X = i)$ using the probability equation for a binomial random variable.

$$E_0 = 50 \times P(X = 0) = 50 \times \binom{5}{0} \times 0.72^0 \times 0.28^5 = 0.086$$

$$E_1 = 50 \times P(X = 1) = 50 \times \binom{5}{1} \times 0.72^1 \times 0.28^4 = 1.1064$$

$$E_2 = 50 \times P(X = 2) = 50 \times \binom{5}{2} \times 0.72^2 \times 0.28^3 = 5.6900$$

Combine to get all E values to be 5 or more

Similarly $E_3 = 14.6313$, $E_4 = 18.8117$, $E_5 = 9.6746$

After combining the relevant cells, this gives:

Number of dead flies	**⩽ 2**	**3**	**4**	**5**	**Total**
Observed (O_i)	7	11	24	8	50
Expected (E_i)	6.8825	14.6313	18.8117	9.6476	50
$\dfrac{(O_i - E_i)^2}{E_i}$	0.0020	0.9012	1.4309	0.2905	2.62

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 2.62 is less than 5.991, there is insufficient evidence to reject H_0 at the 5% level.
The distribution B(5, 0.72) may be a suitable model for the data.

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

Total frequency $= 112 + 56 + 40 = 208$

$$\text{Mean} = \lambda = \frac{1 \times 56 + 2 \times 40}{208} = \frac{136}{208} = 0.654 \text{ (3 d.p.)}$$

Calculate the expected frequencies as follows:

$$E_0 = 208 \times \mathrm{P}(X = 0) = 208 \times \frac{e^{-0.654}\, 0.654^0}{0!} = 108.152$$

$$E_1 = 208 \times \mathrm{P}(X = 1) = 208 \times \frac{e^{-0.654}\, 0.654^1}{1!} = 70.731$$

$$E_2 = 208 \times \mathrm{P}(X = 2) = 208 \times \frac{e^{-0.654}\, 0.654^2}{2!} = 23.129$$

$$E_{i>2} = 208 - (108.152 + 70.731 + 23.129) = 5.988$$

This gives all E values of 5 or more:

Number of accidents	**0**	**1**	**2**	**$\geqslant 3$**	**Total**
Observed (O_i)	112	56	40	0	208
Expected (E_i)	108.152	70.731	23.129	5.988	208
$\dfrac{(O_i - E_i)^2}{E_i}$	0.1369	3.0680	12.2062	5.988	21.499

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 21.5 is greater than 5.991, reject H_0 at the 5% level. This suggests that the data cannot be modelled by Po(0.654)

11 H_0: Rocks in site B occur with the same distribution as seen in the sample from site A
H_1: Rocks in site B do not occur with the same distribution as seen in the sample from site A

In the sample from site A, Igneous : Sedimentary : Other = 6 : 11 : 3
Applying this to the total 60 stones collected in site B to obtain expected values:

Rock type	**Igneous**	**Sedimentary**	**Other**	**Total**
Observed (O_i)	10	35	15	60
Expected (E_i)	18	33	9	60
$\dfrac{(O_i - E_i)^2}{E_i}$	3.556	0.121	4	7.677

Degrees of freedom $\nu = 3 - 1 = 2$, and from the tables: $\chi^2_2(5\%) = 5.991$

As 7.677 is greater than 5.991, reject H_0 at the 5% level. The distribution of rocks at Site B does not match the distribution seen in the sample from site A.

12 a Mean $= \dfrac{1\times4+2\times7+3\times8+4\times10+5\times6+6\times7+7\times4+8\times4}{4+7+8+10+6+7+4+4} = \dfrac{214}{50} = 4.28$

b H_0: The data can be modelled by a Po(4.28) distribution.
H_1: The data cannot be modelled by Po(4.28) distribution.

Calculate the expected frequencies as follows:

$$E_0 = 50\times P(X=0) = 50\times\frac{e^{-4.28}4.28^0}{0!} = 0.6921$$

$$E_1 = 50\times P(X=1) = 50\times\frac{e^{-4.28}4.28^1}{1!} = 2.9623$$

$$E_2 = 50\times P(X=2) = 50\times\frac{e^{-4.28}4.28^2}{2!} = 6.3394$$

Combine to get all E values to be 5 or more.

$$E_3 = 50\times P(X=3) = 50\times\frac{e^{-4.28}4.28^3}{3!} = 9.0442$$

Similarly $E_4 = 9.6773$, $E_5 = 8.2838$, $E_6 = 5.9091$ and $E_{i\geqslant 7} = 7.0918$

After combining cells to ensure all values of E are greater than 5, this gives:

Weekly sales	$\leqslant 2$	**3**	**4**	**5**	**6**	$\geqslant 7$	**Total**
Observed (O_i)	11	8	10	6	7	8	50
Expected (E_i)	9.9938	9.0442	9.6773	8.2838	5.9091	7.0918	50
$\dfrac{(O_i-E_i)^2}{E_i}$	0.1013	0.1206	0.0108	0.6296	0.2014	0.1163	1.18

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 1.18 is less than 9.488, there is insufficient evidence to reject H_0 at the 5% level.
The distribution Po(4.28) may be a suitable model for the data.

13 H_0: There is no association between gender and left- and right-handedness.
H_1: There is an association between gender and left- and right-handedness.

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

	Left-handed	**Right-handed**	**Total**
Male	100	600	700
Female	80	800	880
Total	180	1400	1580

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

Expected frequency 'Male' and 'Left-handed' $= \dfrac{700 \times 180}{1580} = 79.747$

The expected frequency and test statistic (X^2) calculations are:

O_i	E_i	$\dfrac{(O_i - E_i)^2}{E_i}$
100	79.747	5.1436
600	620.253	0.6613
80	100.253	4.0915
800	779.747	0.5260

$$X^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 10.42$$

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

As 10.42 is greater than 3.841, reject H_0 at the 5% level. Conclude there is evidence of an association between gender and left- and right-handedness in this population.

14 a H_0: There is no association between gender and preferred science subject.
H_1: There is no association between gender and preferred science subject.

b Total females = 130; total biology = 68; total individuals sampled = 300

$$E_{F,\,Bio} = \frac{130 \times 68}{300} = 29.47 \text{ (2 d.p.)}$$

c The expected frequency and test statistic (X^2) calculations are:

O_i	E_i	$\frac{(O_i - E_i)^2}{E_i}$
74	$\frac{170 \times 119}{300} = 67.43$	0.6401
28	$\frac{170 \times 68}{300} = 38.53$	2.8778
68	$\frac{170 \times 113}{300} = 64.03$	0.2461
45	$\frac{130 \times 119}{300} = 51.57$	0.8370
40	$\frac{130 \times 68}{300} = 29.47$	3.7625
45	$\frac{130 \times 113}{300} = 48.97$	0.3218

$$X^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 8.685$$

d The number of degrees of freedom $\nu = (3-1)(2-1) = 2$; from the tables: $\chi_2^2(1\%) = 9.210$
As 8.685 is less than 9.210, there is insufficient evidence to reject H_0 at the 1% level.

e From the tables: $\chi_2^2(5\%) = 5.991$
As 8.685 is greater than 5.991, H_0 would be rejected at the 5% significance level.

15 a i $P(X=1)=\dfrac{e^{-2.15}\times 2.15^1}{1!}=0.2504$ (4 d.p.)

ii $P(X>2)=1-P(X\leqslant 2)=1-0.6361=0.3639$ (4 d.p.)

b Mean calls received $=\dfrac{\sum fx}{\sum f}=\dfrac{10\times 0+12\times 1+14\times 2+12\times 3+8\times 4+3\times 5+1\times 6}{60}=\dfrac{129}{60}=2.15$

c Expected frequency $E_x=60\times P(X=x)$

$a=60\times P(X=2)=60\times 0.2692=16.15$ (2 d.p.)

$b=60-(6.99+15.03+a+11.58+6.22+2.67)=1.36$

d H_0: The data is drawn from a Poisson distribution.
H_1: The data is not drawn from a Poisson distribution.

e From part **c**, the observed and expected frequencies are:

Number of calls	0	1	2	3	4	5	$\geqslant 6$	Total
Observed (O_i)	10	12	14	12	8	3	1	60
Expected (E_i)	6.99	15.03	16.15	11.58	6.22	2.67	1.36	60

The final three cells should be combined so that the expected value in each cell is at least 5.

f The calculation of the test statistic is:

Number of calls	0	1	2	3	$\geqslant 4$	Total
Observed (O_i)	10	12	14	12	12	60
Expected (E_i)	6.99	15.03	16.15	11.58	10.25	60
$\dfrac{(O_i-E_i)^2}{E_i}$	1.2962	0.6108	0.2862	0.1523	0.2988	2.507

Degrees of freedom $\nu=3$ (five data cells with two constraints as λ is estimated by calculation)
From the tables: $\chi^2_3(5\%)=7.815$

As 2.507 is less than 7.815, there is insufficient evidence to reject H_0 at the 5% level and to conclude that the data is not drawn from Poisson distribution.

16 a Each friend has an equal probability to offer a lift (success) and David tries until he gets a success. This suggests that a geometric distribution would be appropriate.

b Let the random variable X be the number of calls David has to make to get a lift.

If $X \sim \text{Geo}(p)$ then $\text{E}(X) = \frac{1}{p}$

Using the data:

$$\text{Mean} = \bar{x} = \frac{\sum fx}{\sum f} = \frac{130 \times 1 + 54 \times 2 + 24 \times 3 + 28 \times 4 + 13 \times 5 + 5 \times 6 + 1 \times 7}{255} = \frac{524}{255}$$

Estimate $p = \frac{1}{\bar{x}} = \frac{255}{524} \approx 0.4866$

c H_0: The data is drawn from a geometric distribution.
H_1: The data is not drawn from a geometric distribution.

As $p = \frac{255}{524}$, calculate the expected frequencies using the equation:

$$E_i = 255 \times \text{P}(X = i) = 255 \times \frac{255}{524}\left(\frac{269}{524}\right)^{i-1} = 124.09\left(\frac{269}{524}\right)^{i-1}$$

The observed and expected frequencies are:

Number of calls	**1**	**2**	**3**	**4**	**5**	**6**	**$\geqslant 7$**	**Total**
Observed (O_i)	130	54	24	28	13	5	1	60
Expected (E_i)	124.09	63.70	32.70	16.79	8.62	4.42	4.67	60

The final two cells should be combined so that the expected value in each cell is at least 5.

Number of calls	**1**	**2**	**3**	**4**	**5**	**$\geqslant 6$**	**Total**
Observed (O_i)	130	54	24	28	13	6	60
Expected (E_i)	124.09	63.70	32.70	16.79	8.62	9.09	60
$\frac{(O_i - E_i)^2}{E_i}$	0.2815	1.4771	2.3147	7.4844	2.226	1.0504	14.83

Degrees of freedom $\nu = 4$ (six data cells with two constraints as p is estimated by calculation)
From the tables: $\chi_4^2(5\%) = 9.488$

As 14.83 is greater than 9.488, reject H_0 at the 5% significance level and conclude that the data is not drawn from geometric distribution.

17 a Attempts are not independent (unless his short-term memory is as poor as his long-term memory). For example, if he tries 2, he is not likely to try 2 again on the next try. Also, whether he gets through is also dependent on whether they are at home, so is not solely dependent on the accuracy of his attempt.

b H_0: The data is drawn from a Geo$\left(\frac{1}{3}\right)$ distribution.

H_1: The data is not drawn from a Geo$\left(\frac{1}{3}\right)$ distribution.

As $p = \frac{1}{3}$, calculate the expected frequencies using the equation:

$$E_i = 52 \times \text{P}(X = i) = 5\frac{52}{3}\left(\frac{2}{3}\right)^{i-1}$$

The observed and expected frequencies are:

Number of calls	**1**	**2**	**3**	**4**	**5**	**6**	**$\geqslant 7$**	**Total**
Observed (O_i)	27	10	10	4	0	0	1	52
Expected (E_i)	17.33	11.56	7.70	5.14	3.42	2.28	4.57	52

The final three cells should be combined so that the expected value in each cell is at least 5.

Number of calls	**1**	**2**	**3**	**4**	**$\geqslant 5$**	**Total**
Observed (O_i)	27	10	10	4	1	52
Expected (E_i)	17.33	11.56	7.70	5.14	10.27	52
$\dfrac{(O_i - E_i)^2}{E_i}$	5.3958	0.2105	0.6870	0.2528	8.3674	14.914

Degrees of freedom $\nu = 4$ (five data cells with a single constraint)
From the tables: $\chi_3^2(5\%) = 9.488$

As 2.507 is less than 7.815, there is insufficient evidence to reject H_0 at the 5% level and to conclude that the data is not drawn from Poisson distribution.

Challenge

a Let l_i be the midpoint of each of the groups. Then the estimate of the mean is given by:

$$\bar{l} = \frac{\sum l_i f_i}{n} = \frac{2.5\times 7 + 7.5\times 63 + 12.5\times 221 + 17.5\times 177 + 22.5\times 32}{500} = \frac{7070}{500} = 14.14$$

So the estimate of the mean is 14.14 minutes and the estimate of the variance is given by:

$$\sigma^2 = \frac{\sum l_i^2 f_i}{n} - \bar{l}^2$$
$$= \frac{2.5^2\times 7 + 7.5^2\times 63 + 12.5^2\times 221 + 17.5^2\times 177 + 22.5^2\times 32}{500} - 14.14^2$$
$$= \frac{108525}{500} - = 217.05 - 199.9396 = 17.11 \text{ (2 d.p.)}$$

b H_0: Call length can be modelled by a normal distribution.
H_1: Call length cannot be modelled by a normal distribution.

First, extend the categories so that the values taken by l lie in the interval $(-\infty, \infty)$ to make the data interval compatible with a normal distribution.
Then given $X \sim N(14.14,\ 17.11)$ use the normal cumulative distribution function on a calculator to find the expected frequencies.

$$E_{l<5} = 500\,P(X<5) = 500\times 0.01356 = 6.78$$
$$E_{5\leqslant l<10} = 500\,P(5\leqslant X<10) = 500\big(P(X<10)-P(X<5)\big) = 500(0.15845-0.01356) = 72.445$$
$$E_{10\leqslant l<15} = 500\,P(10\leqslant X<15) = 500\big(P(X<15)-P(X<10)\big) = 500(0.58235-0.15845) = 211.95$$
$$E_{15\leqslant l<20} = 500\,P(15\leqslant X<20) = 500\big(P(X<20)-P(X<15)\big) = 500(0.92171-0.58235) = 169.68$$
$$E_{l>20} = 500\,P(X>20) = 500(1-P(X<20) = 500(1-0.92171) = 39.145$$

Length of call	$l<5$	$5\leqslant l<10$	$10\leqslant l<15$	$15\leqslant l<20$	$l\geqslant 20$	**Total**
Observed (O_i)	7	63	221	177	32	500
Expected (E_i)	6.78	72.445	211.95	169.68	39.145	52
$\frac{(O_i-E_i)^2}{E_i}$	0.0071	1.2314	0.3864	0.3158	1.3042	3.244

Calculating the expected values uses total, mean estimate and variance estimate calculated from the data, consuming three degrees of freedom, so $\nu = 5-3 = 2$
From the statistical table: $\chi_2^2(5\%) = 5.991$

As 3.244 is less than 5.991, there is insufficient evidence to reject H_0 at the 5% level. A normal distribution is a suitable model for this data.