Review Exercise 2

1 **a** Let the random variable X be the height of a three-year-old child and let $\bar{X}$ be the sample mean found from a sample of 100 three-year-old children.

As the sample is large, using the central limit theorem $\bar{X} \approx\sim \mathrm{N}\left(90, \frac{5^2}{100}\right)$, i.e. $\bar{X} \approx\sim \mathrm{N}(90, 0.25)$

b Using the normal distribution function on a calculator

$\mathrm{P}(\bar{X} \geqslant 91) = 1 - \mathrm{P}(\bar{X} < 91) = 1 - 0.9772 = 0.0228$ (4 d.p.) =0.0228

2 Using the central limit theorem $\bar{X} \sim \mathrm{N}\left(10, \frac{3^2}{5}\right)$, i.e. $\bar{X} \sim \mathrm{N}(10, 1.8)$

Using the normal distribution function on a calculator

$\mathrm{P}(\bar{X} < 10) = 0.5$ and $\mathrm{P}(\bar{X} \leqslant 7) = 0.0127$

So $\mathrm{P}(7 < \bar{X} < 10) = 0.5 - 0.0127 = 0.4873$ (4 d.p.)

3 **a** Probabilities sum to 1, so:

$0.4 + 2k + 0.3 + k = 1 \Rightarrow 3k = 0.3 \Rightarrow k = 0.1$

b To use the central limit theorem, find the mean and variance of X.

$\mathrm{E}(X) = 0.4 + (2 \times 0.2) + (3 \times 0.3) + (4 \times 0.1) = 2.1$

$\mathrm{E}\left(X^2\right) = 0.4 + (4 \times 0.2) + (9 \times 0.3) + (16 \times 0.1) = 5.5$

$\mathrm{Var}(X) = \mathrm{E}\left(X^2\right) - (\mathrm{E}(X))^2 = 5.5 - 4.41 = 1.09$

So by the central limit theorem, $\bar{X} \approx\sim \mathrm{N}\left(2.1, \frac{1.09}{200}\right)$, i.e. $\bar{X} \approx\sim \mathrm{N}(2.1, 0.00545)$

Using the normal distribution function on a calculator

$\mathrm{P}(\bar{X} > 2.09) = 1 - \mathrm{P}(\bar{X} \leqslant 2.09) = 1 - 0.4461 = 0.5539$ (4 d.p.)

c This estimate is accurate since the sample size, $n = 200$, is large.

4 **a** Let the random variable X be the number of calls the centre receives every minute, then $X \sim \mathrm{Po}(15)$

Using a calculator,

$\mathrm{P}(X < 10) = \mathrm{P}(X \leqslant 9) = 0.0699$ (4 d.p.)

b Let the random variable Y be the number of calls the centre receives in a 30-minute period, then $Y \sim \mathrm{Po}(450)$

Using a calculator

$\mathrm{P}(Y \leqslant 420) = 0.0810$ (4 d.p.)

© Pearson Education Ltd 2019. Copying permitted for purchasing institution only. This material is not copyright free.

4 c Let the random variable X be the number of calls the centre receives every minute, then $X \sim \text{Po}(15)$. As this is a Poisson distribution the mean and the variance of X is λ, i.e. 15

The number of calls made in 30-minute window is $30\bar{X}$, where $\bar{X}$ is the sample mean of 30 consecutive 1-minute samples.

Model this using the central limit theorem, $\bar{X} \approx\sim \text{N}\left(15, \frac{15}{30}\right)$, i.e. $\bar{X} \approx\sim \text{N}(15, 0.5)$

Require $\text{P}(30\bar{X} \leqslant 420) = \text{P}(\bar{X} \leqslant 14)$

Using a calculator $\text{P}(\bar{X} \leqslant 14) = 0.0786$ (4 d.p.)

The answer is close to the value found in part **b**, so the central limit theorem provides a good approximation in this case.

5 Let the random variable X be the number of attempts a student makes before selecting a green ball then $X \sim \text{Geo}(0.25)$

$$\text{E}(X) = \frac{1}{p} = \frac{1}{0.25} = 4 \quad \text{Var}(X) = \frac{1-p}{p^2} = \frac{0.75}{0.0625} = 12$$

Let $\bar{X}$ be the sample mean of the number attempts required by all 20 students

Then by the central limit theorem $\bar{X} \approx\sim \text{N}\left(4, \frac{12}{20}\right)$, i.e. $\bar{X} \approx\sim \text{N}(4, 0.6)$

Using the normal distribution function on a calculator

$\text{P}(\bar{X} > 4.5) = 1 - \text{P}(\bar{X} \leqslant 4.5) = 1 - 0.7407 = 0.2593$ (4 d.p.) $= 0.0228$

6 a Let the random variable X be the number of questions attempted before the student gets a 4th answer correct then, $X \sim \text{Negative B}(4, 0.2)$

$$\text{P}(X = 12) = \binom{11}{3} \times (0.2)^4 (0.8)^8 = 0.0443 \text{ (4 d.p.)}$$

b $\text{E}(X) = \frac{r}{p} = \frac{4}{0.2} = 20$

c By the central limit theorem, for a sample of 15 students $\bar{X} \approx\sim \text{N}\left(\mu, \frac{\sigma^2}{15}\right)$

$$\text{Var}(X) = \sigma^2 = \frac{r(1-p)}{p^2} = \frac{4 \times 0.8}{0.04} = 80$$

So $\bar{X} \approx\sim \text{N}\left(20, \frac{80}{15}\right)$

Using the normal distribution function on a calculator

$\text{P}(\bar{X} < 19) = 0.3325$ (4 d.p.)

© Pearson Education Ltd 2019. Copying permitted for purchasing institution only. This material is not copyright free.

7 a The table shows the ranks and d and d^2 for each pair of ranks. Remember to rank the data. (Note, it does not matter whether the data is ranked from highest to lowest or vice versa as long as it is same for both judges. It is ranked from highest to lowest here.)

Display	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*
Judge P	25	19	21	23	28	17	16	20
Judge Q	20	9	21	13	17	14	11	15
Rank, P	2	6	4	3	1	7	8	5
Rank, Q	2	8	1	6	3	5	7	4
d	0	–2	3	–3	–2	2	1	1
***d*²**	0	4	9	9	4	4	1	1

$$\sum d^2 = 32$$

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6\times 32}{8(8^2-1)} = 1 - \frac{8}{21} = \frac{13}{21} = 0.619 \text{ (3 s.f.)}$$

b $\text{H}_0: \rho = 0$ There is no correlation between the judges rankings

$\text{H}_1: \rho > 0$ There is a positive correlation between the judges rankings

Sample size = 8

Significance level = 0.05

The critical value for r_s for a 0.05 significance level with a sample size of 8 is $r_s = 0.6429$

As $0.619 < 0.6429$, accept H₀. There is insufficient evidence at the 5% significance level of a positive correlation between rankings of the judges – the competitor's claim is justified.

8 a The table shows the ranks and d and d^2 for each pair of ranks.

Stand	**Distance**	**Price**	r_{distance}	r_{price}	d	d^2
A	50	1.75	1	9	–8	64
B	175	1.20	2	7	–5	25
C	270	2.00	3	10	–7	49
D	375	1.05	4	6	–2	4
E	425	0.95	5	4	1	1
F	580	1.25	6	8	–2	4
G	710	0.80	7	2	5	25
H	790	0.75	8	1	7	49
I	890	1.00	9	5	4	16
J	980	0.85	10	3	7	49
					Total	286

$$\sum d^2 = 286$$

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6\times 286}{10(10^2-1)} = -0.733 \text{ (3 s.f.)}$$

© Pearson Education Ltd 2019. Copying permitted for purchasing institution only. This material is not copyright free.

8 b $\mathrm{H}_0: \rho = 0,\ \mathrm{H}_1: \rho < 0$

Sample size = 10

Significance level = 0.05

The critical value for r_s for a 0.05 significance level with a sample size of 10 is $r_s = -0.5636$.

As $-0.733 < -0.5636$, r_s lies within the critical region, so reject H_0. There is sufficient evidence at the 5% significance level that the price of an ice cream and the distance from the pier are negatively correlated. The further from the pier you go, the less you are likely to pay of an ice cream.

9 a There is no reason to assume that the variables are normally distributed. Therefore use Spearman's rank correlation coefficient.

b Let x be the deaths from pneumoconiosis and y the deaths from lung cancer.
The table shows the ranks and d and d^2 for each pair of ranks.

Age group	20–29	30–39	40–49	50–59	60–69	70+
x	12.5	5.9	18.5	19.4	31.2	31
y	3.7	9	10.2	19	13	18
Rank, x	5	6	4	3	1	2
Rank, y	6	5	4	1	3	2
d	−1	1	0	2	−2	0
d^2	1	1	0	4	4	0

$$\sum d^2 = 10$$

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6 \times 10}{6(6^2-1)} = 1 - \frac{10}{35} = \frac{5}{7} = 0.714 \text{ (3 s.f.)}$$

c $\mathrm{H}_0: \rho = 0,\ \mathrm{H}_1: \rho > 0$

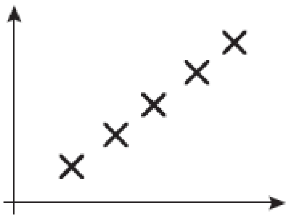
Sample size = 6

Significance level = 0.05

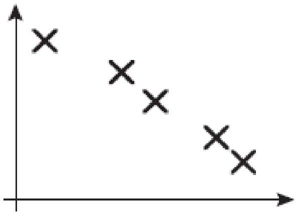
The critical value for r_s for a 0.05 significance level with a sample size of 6 is $r_s = 0.8286$.

As $0.714 < 0.8286$, accept H_0. There is insufficient evidence at the 5% significance level to suggest a positive association between the rates of deaths from pneumoconiosis and lung cancer.

10 a i As $r = 1$, there is a perfect positive correlation. The points form a straight line, with a positive gradient.



10 a ii As $r_s = -1$ but $r > -1$ there is an imperfect negative correlation. The points approximate a straight line, with a negative gradient.



b i Rearranging the data first, the table d and d^2 for each pair of ranks.

Dog	***A***	***B***	***C***	***D***	***E***	***F***	***G***
Judge 1	1	4	2	3	5	6	7
Judge 2	1	2	4	3	5	7	6
d	0	2	−2	0	0	−1	1
***d*²**	0	4	4	0	0	1	1

$$\sum d^2 = 10$$

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6 \times 10}{7(7^2-1)} = 1 - \frac{10}{35} = \frac{5}{7} = 0.821 \text{ (3 s.f.)}$$

ii $\text{H}_0: \rho = 0,\ \text{H}_1: \rho > 0$

Sample size = 7

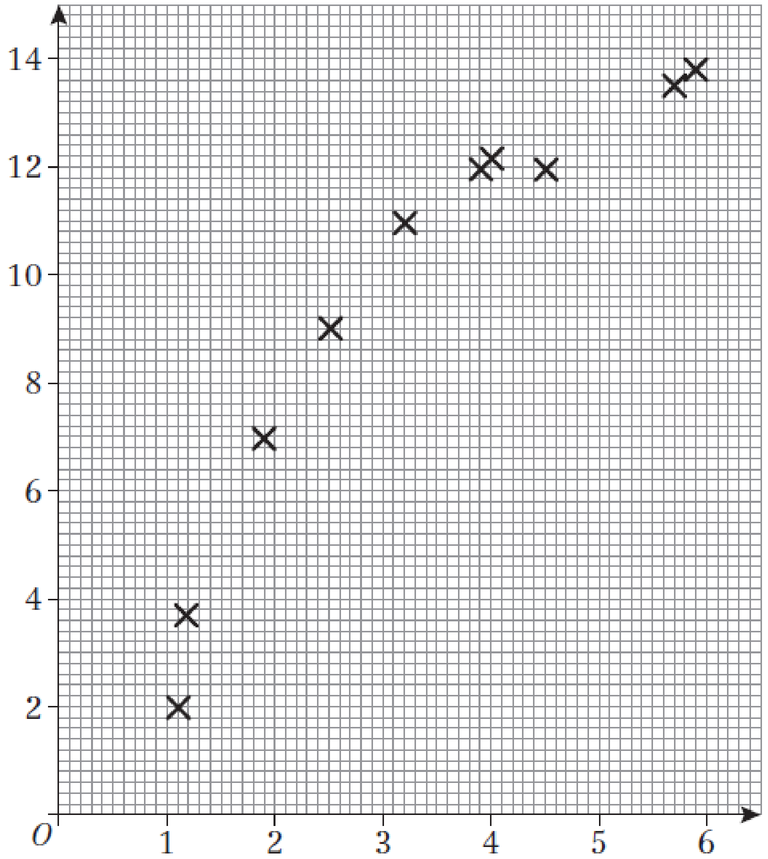
Significance level = 0.05

The critical value for r_s for a 0.05 significance level with a sample size of 7 is $r_s = 0.7143$.

As $0.821 > 0.7143$, r_s lies within the critical region, so reject H_0. There is sufficient evidence at the 5% significance level that the judges are in agreement, i.e. there is evidence of a (positive) correlation between the ranks awarded by the judges.

© Pearson Education Ltd 2019. Copying permitted for purchasing institution only. This material is not copyright free.

11 a



b The product moment correlation coefficient measures the linear correlation between two variables, i.e. it is a measure of the strength of the linear link between the variables.

c The summary statistics for t and p are:

$$\sum t = 33.9 \qquad \sum p = 96.4$$

$$S_{tt} = \sum t^2 - \frac{\left(\sum t\right)^2}{n} = 141.51 - \frac{33.9^2}{10} = 26.589$$

$$S_{pp} = \sum p^2 - \frac{\left(\sum p\right)^2}{n} = 1081.74 - \frac{96.4^2}{10} = 152.444$$

$$S_{tp} = \sum tp - \frac{\sum t \sum p}{n} = 386.32 - \frac{33.9 \times 96.4}{10} = 59.524$$

d $r = \dfrac{59.524}{\sqrt{152.444 \times 26.589}} = 0.93494\ldots = 0.935$ (3 s.f.)

e $\mathrm{H}_0 : \rho = 0, \ \mathrm{H}_1 : \rho > 0$

Sample size = 10

Significance level = 0.01

The critical value for r for a 0.01 significance level with a sample size of 10 is $r = 0.7155$.

As $0.934 > 0.7155$, r lies within the critical region, so reject H_0. There is evidence at the 1% significance level of a positive correlation between the reactant and the product in the chemistry experiment.

© Pearson Education Ltd 2019. Copying permitted for purchasing institution only. This material is not copyright free.

11 f The test for linear correlation is significant but the scatter diagram suggest that there is a non-linear relationship between the variables. The product–moment correlation coefficient should not be used here since the association/relationship is not linear.

12 a $\mathrm{H}_0 : \rho = 0,\ \mathrm{H}_1 : \rho < 0$

Sample size = 7

Significance level = 0.01

The critical value for r_s for a 0.01 significance level with a sample size of 7 is $r_s = -0.8929$.

As $-0.93 < -0.8929$, r_s lies within the critical region, so reject H$_0$. There is sufficient evidence at the 1% significance level that the speed of flow gets slower the wider the river is.

b i This would have no effect on the coefficient since the rank of the flow at *G* stays the same.

ii Spearman's rank correlation will decrease (i.e. get closer to –1) since the new observation is further supporting the hypothesis.

c Where two or more data values are equal (so there is a tied rank), these observations should be assigned a rank equal to the mean of the tied ranks. Then the product moment correlation coefficient formula should be used to find the Spearman's rank correlation coefficient.

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

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

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

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

$$r = 100 \times \mathrm{P}(X = 2) = 100 \times \left(\mathrm{P}(X \leqslant 2) - \mathrm{P}(X \leqslant 1)\right)$$

$$= 100 \times (0.6769 - 0.3917) = 100 \times 0.2852 = 28.5 \text{ (1 d.p.)}$$

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:

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

© Pearson Education Ltd 2019. Copying permitted for purchasing institution only. This material is not copyright free.

13 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
$\frac{(O_i - E_i)^2}{E_i}$	1.889	0.593	3.167	1.316	2.443	9.406

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

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.

© Pearson Education Ltd 2019. Copying permitted for purchasing institution only. This material is not copyright free.

14 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.

© Pearson Education Ltd 2019. Copying permitted for purchasing institution only. This material is not copyright free.

15 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:

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

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

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 $\chi_4^2(5\%) = 9.488$

The test statistic is less than the critical value (4.17 < 9.488), so there is insufficient evidence to reject H₀ at the 5% level. The binomial distribution with $p = 0.2$ is a suitable model for the number of cuttings that do not grow.

© Pearson Education Ltd 2019. Copying permitted for purchasing institution only. This material is not copyright free.

16 H₀: A Poisson distribution is a suitable model.
H₀: 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
$\frac{(O_i-E_i)^2}{E_i}$	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 H₀ 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.

© Pearson Education Ltd 2019. Copying permitted for purchasing institution only. This material is not copyright free.

17 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:

		Mathematics 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.

© Pearson Education Ltd 2019. Copying permitted for purchasing institution only. This material is not copyright free.

18 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.

© Pearson Education Ltd 2019. Copying permitted for purchasing institution only. This material is not copyright free.

19 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' $= \frac{115 \times 70}{217} = 37.09...$

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	$\frac{(O_i - E_i)^2}{E_i}$
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.

© Pearson Education Ltd 2019. Copying permitted for purchasing institution only. This material is not copyright free.

20 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.

© Pearson Education Ltd 2019. Copying permitted for purchasing institution only. This material is not copyright free.

21 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' $= \dfrac{18\times 22}{50} = 7.92$

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	$\dfrac{(O_i-E_i)^2}{E_i}$
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.

22 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
$\dfrac{(O_i-E_i)^2}{E_i}$	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.

© Pearson Education Ltd 2019. Copying permitted for purchasing institution only. This material is not copyright free.

Challenge

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

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
$\frac{(O_i - E_i)^2}{E_i}$	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, 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.

© Pearson Education Ltd 2019. Copying permitted for purchasing institution only. This material is not copyright free.