Chi Squared - Exercises (with Sol'ns)(8 pages; 14/2/26)

Contingency tables

(1) The table below shows data that have been collected for 1000 people (in an area where there are a lot of ladders) to test the theory that walking under a ladder brings bad luck.

On a particular day, each person is asked whether they walked under a ladder, and whether they experienced bad luck that day.

Walked under ladder? Experienced Bad luck?	Yes	No	Totals
Yes	51	84	135
No	269	596	865
Totals	320	680	1000

(i) Test the hypothesis that walking under ladders brings bad luck, assuming a 5% significance level.

(ii) Discuss any practical difficulties with the data.

Solution – Version 1 (MEI & Edx)

(i) H_0: There is no association between walking under ladders and experiencing bad luck

H_1: There is an association between walking under ladders and experiencing bad luck

The table of expected frequencies is:

Walked under ladder? Experienced Bad luck?	Yes	No	Totals
Yes	43	92	135
No	277	588	865
Totals	320	680	1000

where $320 \times \frac{135}{1000} = 43$

[dealing with the smallest observed frequency first, so that any rounding errors appear in the larger cells]

Then $X^2 = \frac{(51-43)^2}{43} + \frac{(84-92)^2}{92} + \frac{(269-277)^2}{277} + \frac{(596-588)^2}{588}$

$= \frac{64}{43} + \frac{64}{92} + \frac{64}{277} + \frac{64}{588} = 2.524$ (3dp)   [2 marks]

d.f. $= (2-1) \times (2-1) = 1$, so that the critical value of χ^2 at the 5% level is 3.841

As $2.524 < 3.841$, H_0 is accepted: there is not sufficient evidence at the 5% level to conclude that there is any association between walking under ladders and experiencing bad luck

[Had H_1 been accepted, we would then have to examine the data to establish whether bad luck was more or less likely if a ladder had been walked under.]

(ii)

- subjective nature of 'bad luck'

- individuals are allowed to choose whether to walk under ladders: it might be the case that people who are cautious tend to avoid walking under ladders, and experience less 'bad luck' generally (as they are more cautious)

- individuals might walk under more than one ladder in the course of the day (and so be more prone to bad luck)

- individuals walking under their ladder earlier in the day may be prone to more bad luck

[Unusual points, such as the last two, are very unlikely to appear on mark schemes, and so might not attract any marks.]

Solution (Version 2: OCR & AQA)

(i) H_0: There is no association between walking under ladders and experiencing bad luck

H_1: There is an association between walking under ladders and experiencing bad luck

The table of expected frequencies is:

Walked under ladder? Experienced Bad luck?	Yes	No	Totals
Yes	43	92	135
No	277	588	865
Totals	320	680	1000

where $320 \times \frac{135}{1000} = 43$

[dealing with the smallest observed frequency first, so that any rounding errors appear in the larger cells]

$X^2 = \frac{(51-43-0.5)^2}{43} + \frac{(84-92-0.5)^2}{92} + \frac{(269-277-0.5)^2}{277} + \frac{(596-588-0.5)^2}{588}$

$= \frac{63.5}{43} + \frac{63.5}{92} + \frac{63.5}{277} + \frac{63.5}{588} = 2.504$ (3dp)

[with Yates's correction]

d.f. $= (2-1) \times (2-1) = 1$, so that the critical value of χ^2 at the 5% level is 3.841 [1 mark]

As $2.524 < 3.841$, H_0 is accepted: there is not sufficient evidence at the 5% level to conclude that there is any association between walking under ladders and experiencing bad luck

[Had H_1 been accepted, we would then have to examine the data to establish whether bad luck was more or less likely if a ladder had been walked under.]

(ii)

- subjective nature of 'bad luck'

- individuals are allowed to choose whether to walk under ladders: it might be the case that people who are cautious tend to avoid walking under ladders, and experience less 'bad luck' generally (as they are more cautious)

- individuals might walk under more than one ladder in the course of the day (and so be more prone to bad luck)

- individuals walking under their ladder earlier in the day may be prone to more bad luck

[Unusual points, such as the last two, are very unlikely to appear on mark schemes, and so might not attract any marks.]

Goodness of Fit

(1) The number of days in a 5 day week on which a particular train is cancelled is thought to have a Binomial distribution.

Some data were supposed to have been collected over 100 weeks, in order to investigate this model. However, there is reason to believe that no such data exist, and the 'observed' frequencies in the table below were in fact made up to fit the expected distribution. Investigate this assertion at the 5% significant level.

Number of days on which the train is cancelled	0	1	2	3	4	5
'Observed' Frequencies	9	24	34	23	7	3

Solution

Let X be the number of days on which the train is cancelled.

H_0: X genuinely follows a Binomial distribution

H_1: The data have been rigged to suggest a Binomial distribution

$H_0 \Rightarrow X \sim B(5, p)$ for some p

The sample mean of the 'observed' data is

$\frac{0+24+68+69+28+15}{100} = 2.04$

Then, as $E(X) = 5p$, the estimated value for p is $\frac{2.04}{5} = 0.408$

Based on $B(5,0.408)$, $P(X = r) = \binom{5}{r}(0.408)^r(0.592)^{5-r}$,

and, on multiplying by 100, the expected frequencies are as shown in the table below.

Number of days on which the train is cancelled	0	1	2	3	4	5
'Observed' frequency	9	24	34	23	7	3
Expected frequency	7.27	25.06	34.54	23.80	8.20	1.13

In order to carry out the Goodness of Fit test, the expected frequencies of cells have to be at least 5, and so the last two cells are combined:

Number of days on which the train is cancelled	0	1	2	3	4 or 5	
'Observed' frequency, O_i	9	24	34	23	10	
Expected frequency, E_i	7.27	25.06	34.54	23.80	9.33	
$\frac{(O_i-E_i)^2}{E_i}$	0.41	0.04	0.01	0.03	0.05	0.54

$\nu = 5[\text{no. of cells}] - 1[\text{as total freq. is fixed}]$

$-1[\text{estimation of p}] = 3$

and the left-hand critical value at 5% significance is then 0.352

As $0.54 > 0.352$, we accept H_0, and conclude that there is not sufficient evidence of data rigging, at the 5% significance level.