Chi-Squared (Goodness of Fit) Q1 [16 marks] (8/6/21)

Exam Boards

OCR : Statistics (Year 2)

MEI: Statistics a

AQA: -

Edx: S1 (Year 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

[16 marks]

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

[2 marks]

$H_0 \Rightarrow X \sim B(5, p)$ for some p [1 mark]

The sample mean of the 'observed' data is

$\frac{0+24+68+69+28+15}{100} = 2.04$ [2 marks]

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

[2 marks]

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

[4 marks]

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

[2 marks]

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

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

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

[1 mark]

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.

[1 mark]