Goodness-of-Fit Tests

Starter

1. (Review of last lesson) During an influenza epidemic 15 boys and 8 girls became ill out of a class of 22 boys and 28 girls. Assuming this class may be treated as a random sample of the age group, test at the $5\,\%$ level hypothesis whether there is a connection between sex and susceptibility to influenza.

Working: The table for the ***observed*** frequencies is:

Observed frequencies	Boys	Girls	Totals
Became ill	15	8	23
Did not become ill	7	20	27
Totals	22	28	50

The table for the ***expected*** frequencies (if independent) is:

Expected frequencies	Boys	Girls	Totals
Became ill	10.12	12.88	23
Did not become ill	11.88	15.12	27
Totals	22	28	50

Apply Yates' correction to the calculation of χ^2_{calc}

$$\chi^2_{calc} = \sum \frac{(|O_i - E_i| - 0.5)^2}{E_i}$$

$$= \frac{\left(\left|15 - 10.12\right| - 0.5\right)^2}{10.12} + \frac{\left(\left|8 - 12.88\right| - 0.5\right)^2}{12.88} + \frac{\left(\left|7 - 11.88\right| - 0.5\right)^2}{11.88} + \frac{\left(\left|20 - 15.12\right| - 0.5\right)^2}{15.12}$$

$$= 6.269$$

H_0 : there is an association between sex and susceptibility to influenza

H_1 : there is no association between sex and susceptibility to influenza

Degrees of freedom, $\nu = 1 \times 1 = 1$

The critical value at the $5\,\%$ level is $\chi^2_1(5\%) = 3.841$

Since $\chi^2_{calc} = 6.269 < 3.841 = \chi^2_1(5\%)$, we reject H_0.

There is evidence of an association between sex and susceptibility to influenza.

2. One hundred digits between 0 and 9 are generated by a computer with frequencies below:

Digit	0	1	2	3	4	5	6	7	8	9
Frequency	11	8	8	7	8	9	12	9	13	15

(a) Write down the expected frequencies.

(b) Calculate the $\chi^2_{calc} = \sum \frac{(O_i - E_i)^2}{E_i}$ value for these values

(c) State how many degrees of freedom there are in the table and state the $5\,\%$ critical value from the χ^2 tables.

(d) Could the numbers have been generated randomly? Test at the $5\,\%$ level, stating your null and alternative hypotheses clearly.

Working: (a) 10

(b) $\chi^2_{calc} = \sum \frac{(O_i - E_i)^2}{E_i}$

$= \frac{1^2}{10} + \frac{2^2}{10} + \frac{2^2}{10} + \frac{3^2}{10} + \frac{2^2}{10} + \frac{1^2}{10} + \frac{2^2}{10} + \frac{1^2}{10} + \frac{3^2}{10} + \frac{5^2}{10}$

$\chi^2_{calc} = 6.2$

(c) The frequencies from $0 - 8$ can be filled in but then given there is a fixed number of total frequencies, the frequency for 9 is fixed. Therefore, the number of degrees of freedom is 9

$\nu = 10 - 1 = 9; \chi^2_9(5\%) = 16.96$

(d) H_0 : the number were generated randomly

H_1 : the number were not generated randomly

$\chi^2_{calc} = 6.2 < 16.92 = \chi^2_9(5\%);$

We do not reject H_0.

The values could be from a random number table.

E.g. 1 The data in the table are thought to be modelled by the binomial distribution B $(10, 0.2)$. Conduct a test at the $5\ \%$ significance level to check whether this is a good model.

x	0	1	2	3	4	5	6	7	8
Frequency of x	12	28	28	17	7	4	2	2	0

Working:

x	0	1	2	3	4	5	6	7	8
Frequency of x	10.7	26.8	30.2	20.1	8.8	2.6	0.55	0.08	0.01

But $E_i \geq 5$ for $5 - 8$ so we need to combine $4 - 8$

x	0	1	2	3	4—8
Observed	12	28	28	17	15
Expected	10.7	26.8	30.2	20.1	12.1

H_0 : the results can be modelled by a B $(10, 0.2)$ distribution
H_1 : the results cannot be modelled by a B $(10, 0.2)$ distribution
Degrees of freedom, $\nu = 5 - 1 = 4$
The critical value at the $5\ \%$ level is $\chi_4^2(5\%) = 9.488$

$$\chi^2_{calc} = \sum \frac{(O_i - E_i)^2}{E_i}$$

$$= \frac{(12 - 10.7)^2}{10.7} + \frac{(28 - 26.8)^2}{26.8} + \frac{(28 - 30.2)^2}{30.2} + \frac{(17 - 20.1)^2}{20.1} + \frac{(15 - 12.1)^2}{12.1}$$

$\chi^2_{calc} = 1.55$
Since $\chi^2_{calc} = 1.55 < 9.488 = \chi_4^2(5\%)$, we do not reject H_0.
B $(10, 0.2)$ is a possible model for the data

E.g. 2 The table below shows the number of employees in thousands at five factories and the number of accidents in 3 years.

Factory	A	B	C	D	E
Employees (000s)	4	3	5	1	2
Accidents	22	14	25	8	12

Using a $2.5\ \%$ level significance, test the hypothesis that the number of accidents per 1000 employees is constant at each factory.

Hint: work out the number of accidents per 1000 employees.

Working: Total employees $= 15000$ and total accidents $= 81$

So accidents per 1000 employees $= 5.4$

Expected frequencies for the 5 factories are $21.6,\ 16.2,\ 27,\ 5.4,\ 10.8$

H_0 : the number of accidents is constant at each factory

H_1 : the number of accidents is not constant at each factory

Degrees of freedom $= 5 - 1 = 4$

$$\chi^2_{calc} = \sum \frac{(O_i - E_i)^2}{E_i}$$

$$= \frac{(22-21.6)^2}{21.6} + \frac{(14-16.2)^2}{16.2} + \frac{(25-27)^2}{27} + \frac{(8-5.4)^2}{5.4} + \frac{(12-10.8)^2}{10.8}$$

$\chi^2_{calc} = 1.84$

Since $\chi^2_{calc} = 1.84 < 11.14 = \chi^2_4(2.5\%)$, we do not reject H_0.

There is evidence to suggest the number of accidents is constant at each factory.

E.g. 3 The number of telephone calls arriving at an exchange in $6-$minute periods were recorded over a period of 8 hours, with the following results:

Number of calls	0	1	2	3	4	5	6	7	8
Frequency	8	19	26	13	7	5	1	1	0

Can these results be modelled by a Poisson distribution? Test at the 10% level.

Working: H_0 : the results can be modelled by a Poisson distribution
H_1 : the results cannot be modelled by a Poisson distribution
Since we don't have λ we must estimate it from the data.

$$\lambda = \frac{176}{80} = 2.2$$

Number of calls	0	1	2	3	4	5	6	7	8
Frequency	8	19	26	13	7	5	1	1	0
Expected	8.86	19.5	21.5	15.7	8.65	3.81	1.4	0.44	0.12

But $E_i \geq 5$ for $5-8$ so we need to combine $5-8$

Number of calls	0	1	2	3	4	5−8
Frequency	8	19	26	13	7	7
Expected	8.86	19.5	21.5	15.7	8.65	5.8

Degrees of freedom, $\nu = 6 - 1 - 1 = 4$ ***λ is estimated***
The critical value at the 10% level is $\chi^2_4(10\%) = 7.779$

$$\chi^2_{calc} = \sum \frac{(O_i - E_i)^2}{E_i}$$

$$= \frac{(8-8.86)^2}{8.86} + \frac{(19-19.5)^2}{19.5} + \frac{(26-21.5)^2}{21.5} + \frac{(13-15.7)^2}{15.7} + \frac{(7-8.65)^2}{8.65} + \frac{(7-5.8)^2}{5.8}$$

$\chi^2_{calc} = 2.11$
Since $\chi^2_{calc} = 2.11 < 7.779 = \chi^2_4(10\%)$, we do not reject H_0.
Poisson is a possible model for the data.

E.g. 4 A marksman fires 6 shots at a target and records the number of bull's eye hits. After a series of 100 such trials he analyses his scores and the frequencies are below.

Number of hits	0	1	2	3	4	5	6
Frequency	0	26	36	20	10	6	2

(a) Estimate the probability of hitting a bull's eye.
(b) Use a test at the $5\,\%$ significance level to see if these results are consistent with the assumption of a binomial distribution.

Working: (a) Total hits $= 1\times 26 + 2\times 36 + 3\times 20 + 4\times 10 + 5\times 6 + 6\times 2 = 240$
Total shots $= 6\times 100 = 600$
An estimate for the probability of hitting the target is $\dfrac{240}{600} = 0.4$.

(b) Expected $= 100\times {}^6C_x \times 0.4^x \times 0.6^{n-x}$

Number of hits	0	1	2	3	4	5	6
Observed	0	26	36	20	10	6	2
Expected	4.6656	18.6624	31.104	27.648	13.824	3.6864	0.4096

Combine columns where $E_i < 5$.

Number of hits	0-1	2	3	4-6
Observed	26	36	20	18
Expected	23.328	31.104	27.648	17.92

$$\chi^2_{calc} = \sum \frac{(O_i - E_i)^2}{E_i}$$
$$= \frac{(26-23.328)^2}{23.328} + \frac{(36-31.104)^2}{31.104} + \frac{(20-27.648)^2}{27.648} + \frac{(18-17.92)^2}{17.92}$$
$$\chi^2_{calc} = 3.19$$

Degrees of freedom, $\nu = 4 - 1 - 1$ ***p is estimated***
Since $\chi^2_{calc} = 3.19 < 5.991 = \chi^2_2(5\%)$, we do not reject H_0.
There is evidence to suggest the bull's eye hits are binomially distributed.

N.B. Yates' correction is only used for contingency tables

Video A: Goodness of fit tests (from 9:24)
Video B: Goodness of fit tests

Solutions to Starter and E.g.s

Exercise
AS: p111 6C Qu 1i, 2-4, 7, 11, (12, 13 red) — avoid Normal distribution questions.
A2: p111 6C Qu 5, 6, 8-10