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.

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.

Notes

The χ^2 value is an immensely useful statistic as it allows us to test whether data from a frequency table fits a particular distribution, whether uniform, Binomial, Poisson etc.

Expected frequency = probability × total observed frequency

Degrees of freedom, ν, for goodness of fit

From the starter, we could fill in the observed frequencies for digits 0—8 and then the observed frequency for 9 is fixed. That is why there are 9 degrees of freedom.

In such situations, degrees of freedom, ν = number of cells after combining $- 1$

N.B. Remember to use Menu >> 7: Distribution to find the expected probabilities faster.
It is useful to have a 2nd calculator to multiply these probabilities by the total frequency.

E.g. 1 The data in the table are thought to be modelled by the binomial distribution $\mathrm{B}(10, 0.2)$. 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 $\mathrm{B}(10, 0.2)$ distribution
H_1 : the results cannot be modelled by a $\mathrm{B}(10, 0.2)$ distribution
Degrees of freedom, $\nu = 5 - 1 = 4$
The critical value at the $5\,\%$ level is $\chi^2_4(5\%) = 9.488$

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

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

$\chi^2_{calc} = 1.55$
Since $\chi^2_{calc} = 1.55 < 9.488 = \chi^2_4(5\%)$, we do not reject H_0.
$\mathrm{B}(10, 0.2)$ 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.

Degrees of freedom, ν, for goodness of fit — REVISITED
Degrees of freedom becomes more complicated when a required parameter for the distribution is not given. For example:

- Binomial distribution — the value of p is not given
- Poisson distribution — the value of λ is not given

In such cases:

Degrees of freedom, ν = cells after combining — parameters estimated — 1

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.

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

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

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

Summary
Expected frequency = probability × total observed frequency
Degrees of freedom becomes more complicated when a required parameter for the distribution is not given. For example:

- Binomial distribution — the value of p is not given
- Poisson distribution — the value of λ is not given

In such cases:

Degrees of freedom, ν = cells after combining − parameters estimated − 1