Goodness-of-fit tests with continuous distributions

Starter

1. **(Review of last lesson)** The continuous random variable T has probability density function given by $f(t) = \begin{cases} 4t^3 & 0 < t \leq 1 \\ 0 & \text{otherwise} \end{cases}$.

   (a) Find the probability density function of H, where $H = \frac{1}{T^4}$, clearly stating its interval.

   (b) Find $E(1 + 2H^{-1})$.

2. A model is proposed for a continuous random variable. The proposed probability density function is $f(x) = \begin{cases} \frac{6-x}{18} & 0 \leq x \leq 6 \\ 0 & \text{otherwise} \end{cases}$. The observed frequencies of 100 items of data are in the table below.

Class	$0 \leq x < 1$	$1 \leq x < 2$	$2 \leq x < 4$	$4 \leq x < 6$
Frequency	36	29	22	13

   (a) Calculate the expected frequencies for each class.

   (b) Calculate the value of χ^2_{calc} using the formula $\chi^2_{calc} = \sum \frac{(O_i - E_i)^2}{E_i}$.

   (c) State the number of degrees of freedom and hence state the value of $\chi^2_{\nu}(5\%)$

   (d) Carry out a goodness-of-fit test at the 5% significance level to see whether the proposed model fits the data. State the null and alternative hypotheses clearly.

Notes

For continuous random variables, integration is used to calculate the expected frequencies.

Here are the key points from goodness-of-fit tests:

Expected frequency = probability × total observed frequency

With continuous functions, you may need to integrate to find the probabilities.

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

The expected frequencies must be greater than or equal to 5 i.e. $E_i \geq 5$. If this is not the case, adjacent cells must be combined.

Degrees of freedom, ν = number of cells after combining $-$ 1

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

E.g. 1 Test at the $10\,\%$ significance level whether the following data follows a continuous uniform distribution.

Class	$10 \le x < 20$	$20 \le x < 35$	$35 \le x < 50$	$50 \le x < 60$	$60 \le x < 70$
Frequency	21	18	17	12	22

Working:

$$k = \frac{1}{70-10} = \frac{1}{60}$$

$$f(x) = \begin{cases} \frac{1}{60} & 10 \le x \le 70 \\ 0 & \text{otherwise} \end{cases}$$

Sum of observed frequencies is $18 + 21 + 31 + 20 = 90$

Expected frequencies:

$10 \le x < 20$: $\quad 90 \times \frac{20-10}{60} = 15$

$20 \le x < 35$: $\quad 90 \times \frac{35-20}{60} = 22.5$

$35 \le x < 50$: $\quad 90 \times \frac{50-35}{60} = 22.5$

$50 \le x < 60$: $\quad 90 \times \frac{60-50}{60} = 15$

$60 \le x < 70$: $\quad 90 \times \frac{70-60}{60} = 15$

H_0 : the data can be modelled as a continuous uniform distribution.
H_1 : the data cannot be modelled as a continuous uniform distribution.

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

$$= \frac{(21-15)^2}{15} + \frac{(18-22.5)^2}{22.5} + \frac{(17-22.5)^2}{22.5} + \frac{(12-15)^2}{15} + \frac{(22-15)^2}{15}$$

$$\approx 8.51$$

Degrees of freedom, $\nu = 5 - 1 = 4$
The critical value at the $10\,\%$ level is $\chi^2_4(10\%) = 7.779$
Since $\chi^2_{calc} \approx 8.51 > 7.779 = \chi^2_4(10\%)$, we reject H_0.
There is evidence to suggest that the data does not follow a continuous uniform distribution.

E.g. 2 A train station collected data on the lateness of trains over a period of time.

Minutes late	$0 \le t < 4$	$4 \le t < 8$	$8 \le t < 12$	$12 \le t < 20$	$20 \le t < 30$	$30 \le t < 60$
Frequency	24	18	16	8	8	6

(a) Test at the $2.5\,\%$ significance level whether this follows an exponential distribution with mean of 8, given that $\chi^2_{calc} \approx 13.0$ and the expected value for $30 \le t < 60$ is about 1.84.

(b) Test at the $2.5\,\%$ significance level whether an alternative exponential distribution would be more suitable.

Video: Goodness-of-fit for uniform distribution

Solutions to Starter and E.g.s

Exercise
p145 7I Qu 1-4 (5 red)

Summary
Expected frequency = probability × total observed frequency

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

With continuous functions, you may need to integrate to find the probabilities.
The expected frequencies must be greater than or equal to 5 i.e. $E_i \geq 5$. If this is not the case, adjacent cells must be combined.
Degrees of freedom, ν = number of cells after combining $- 1$
Degrees of freedom, ν = cells after combining — number of parameters estimated — 1