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 $H = \dfrac{1}{T^4}$, clearly stating its interval.

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

Working: (a) $F(t) = \int_0^t 4t^3 dx = \left[t^4\right]_0^t = t^4$

$$\therefore F(t) = \begin{cases} 0 & t \leq 0 \\ t^4 & 0 < t \leq 1 \\ 1 & t > 1 \end{cases}$$

Let $G(h)$ be the cumulative distribution function of H.

$$G(h) = P(H \leq h)$$

Since $Y = \dfrac{1}{\sqrt{X}}$***:***

$$G(h) = P\left(\frac{1}{T^4} \leq h\right)$$

Rearrange to make X ***the subject of the inequality:***

$$G(h) = P\left(T^4 \geq \frac{1}{h}\right)$$

$$= P\left(T \geq \frac{1}{\sqrt[4]{h}}\right) + P\left(T \leq -\frac{1}{\sqrt[4]{h}}\right)$$

Since $t > 0$, $P\left(T \leq -\dfrac{1}{\sqrt[4]{h}}\right) = 0$

$$G(h) = P\left(T \geq \frac{1}{\sqrt[4]{h}}\right)$$

Put it in terms of $P(X \leq \ldots)$***:***

$$G(h) = 1 - P\left(T \leq \frac{1}{\sqrt[4]{h}}\right)$$

Replace $P(X \leq \ldots)$ ***by*** $F(\ldots)$***:***

$$G(h) = 1 - F\left(\frac{1}{\sqrt[4]{h}}\right)$$

Replace $F(\ldots)$ ***by its function:***

From $F(t) = \begin{cases} 0 & t \leq 0 \\ t^4 & 0 < t \leq 1 \\ 1 & t > 1 \end{cases}$, we get

$$G(h) = 1 - F\left(\frac{1}{\sqrt[4]{h}}\right) = \begin{cases} 1 - 0 & \frac{1}{\sqrt[4]{h}} \leq 0 \\ 1 - \frac{1}{h} & 0 < \frac{1}{\sqrt[4]{h}} \leq 1 \\ 1 - 1 & \frac{1}{\sqrt[4]{h}} > 1 \end{cases}$$

$$G(h) = 1 - F\left(\frac{1}{\sqrt[4]{h}}\right) = \begin{cases} 1 & \frac{1}{\sqrt[4]{h}} \leq 0 \\ 1 - \frac{1}{h} & h \geq 1 \\ 0 & h < 1 \end{cases}$$

i.e. $G(h) = \begin{cases} 0 & h < 1 \\ 1 - \frac{1}{h} & h \geq 1 \end{cases}$

The pdf of H: $g(h) = G'(h)$

$$= \frac{d}{dy}\left(1 - \frac{1}{h}\right)$$
$$= \frac{d}{dy}\left(1 - h^{-1}\right)$$
$$= h^{-2}$$
$$= \frac{1}{h^2}$$

The probability distribution function of Y is $g(h) = \begin{cases} \frac{1}{h^2} & h \geq 1 \\ 0 & \text{otherwise} \end{cases}$.

(b) $E\big(f(h)\big) = \int_{-\infty}^{\infty} f(h)g(h)dh$

$$E(1 + 2H^{-1}) = \int_{1}^{\infty} (1 + 2h^{-1}) \times h^{-2} dh$$
$$= \int_{1}^{\infty} (h^{-2} + 2h^{-3}) dh$$
$$= \left[-h^{-1} - h^{-2}\right]_{1}^{\infty}$$
$$= \left[\frac{1}{h} + \frac{1}{h^2}\right]_{\infty}^{1}$$
$$= 2$$

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

Class	$0 \le x < 1$	$1 \le x < 2$	$2 \le x < 4$	$4 \le x < 6$
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.

Working:

(a)
$$0 \le x < 1: \quad 100 \times \int_0^1 \frac{6-x}{18} dx = \frac{275}{9} = 30.\dot{5}$$
$$1 \le x < 2: \quad 100 \times \int_1^2 \frac{6-x}{18} dx = 25$$
$$2 \le x < 4: \quad 100 \times \int_2^4 \frac{6-x}{18} dx = \frac{100}{3} = 33.\dot{3}$$
$$4 \le x < 6: \quad 100 \times \int_4^6 \frac{6-x}{18} dx = \frac{100}{9} = 11.\dot{1}$$

(b)
$$\chi^2_{calc} = \sum \frac{(O_i - E_i)^2}{E_i}$$
$$= \frac{(36 - 30.\dot{5})^2}{30.\dot{5}} + \frac{(29-25)^2}{25} + \frac{(22 - 33.\dot{3})^2}{33.\dot{3}} + \frac{(13 - 11.\dot{1})^2}{11.\dot{1}}$$
$$\chi^2_{calc} \approx 5.78$$

(c) Degrees of freedom, $\nu = 4 - 1 = 3$
The critical value at the $5\,\%$ level is $\chi^2_3(5\%) = 7.815$

(d) H_0 : the data can be modelled by the proposed model
H_1 : the data cannot be modelled by the proposed model
Since $\chi^2_{calc} \approx 5.78 < 7.815 = \chi^2_3(5\%)$, we do not reject H_0.
There is evidence to suggest that the proposed model is a good fit.

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: $k = \dfrac{1}{70-10} = \dfrac{1}{60}$

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

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

Expected frequencies:

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

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

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

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

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

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.

Working:

(a) The final two columns need to be combined since $E_i < 5$
Degrees of freedom, $\nu = 5 - 1 = 4$
The critical value at the $2.5\ \%$ level is $\chi^2_4(2.5\%) = 11.14$
Since $\chi^2_{calc} \approx 13.0 > 11.14 = \chi^2_4(2.5\%)$, we reject H_0.
There is evidence to suggest that the data does not follow an exponential distribution mean of 8.

(b) From the table, an estimation for the mean is 11.425

So $\lambda = \dfrac{1}{11.425} \approx 0.08753$ so $f(x) = 0.08753e^{-0.08753x}$

$0 \le t < 4$: $\quad 80 \times \displaystyle\int_0^4 0.08753e^{-0.08753t}dt \approx 23.6$

$4 \le t < 8$: $\quad 80 \times \displaystyle\int_4^8 0.08753e^{-0.08753t}dt \approx 16.65$

$8 \le t < 12$: $\quad 80 \times \displaystyle\int_8^{12} 0.08753e^{-0.08753t}dt \approx 11.7$

$12 \le t < 20$: $\quad 80 \times \displaystyle\int_{12}^{20} 0.08753e^{-0.08753t}dt \approx 14.1$

$20 \le t < 30$: $\quad 80 \times \displaystyle\int_{20}^{30} 0.08753e^{-0.08753t}dt \approx 8.10$

$30 \le t < 60$: $\quad 80 \times \displaystyle\int_{30}^{60} 0.08753e^{-0.08753t}dt \approx 5.37$

No columns need to be combined.

$$\chi^2_{calc} = \sum \frac{(O_i - E_i)^2}{E_i} \approx 4.41$$

Degrees of freedom, $\nu = 6 - 1 - 1 = 4$ ***mean was estimated***
The critical value at the $2.5\ \%$ level is $\chi^2_4(2.5\%) = 11.14$
Since $\chi^2_{calc} \approx 4.41 < 11.14 = \chi^2_4(2.5\%)$, we do not reject H_0.
There is evidence to suggest that the data does follow an exponential distribution with a mean of 11.425.

Video: Goodness-of-fit for uniform distribution

Solutions to Starter and E.g.s

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