Testing for goodness-of-fit

- Hypothesis testing procedures are designed for problems in which the probability distribution is known and hypothesis involve the parameters of the distribution.
- Another kind of hypothesis is often encountered.
- We do not know the underlying distribution of the population and we wish to test the hypothesis that particular distribution will be satisfactory as a population model.
- We describe a formal goodness-of-fit test procedure based on the chi-square distribution.
- The test procedure requires a random sample of size n from the population whose probability distribution is unknown.
- These n observations are arranged in a frequency histogram having k-class intervals.
- Let O_i be the observed frequency of the i^{th} class.
- From the hypothesized probability distribution we compute the expected frequency denoted by E_i
- The test statistic is $\chi_0^2 = \frac{(O_i - E_i)^2}{E_i}$

which has an approximately χ^2_{k-p-1} distribution with $k-p-1$ degrees of freedom where p is the number of parameters of the hypothesized distribution estimated by sample statistics.

- This approximation improves as n increases.
- We will reject ~~H_0~~ the hypothesis that the distribution of the population is the hypothesized if the calculated value of the test statistic $\chi_0^2 > \chi^2_{\alpha, k-p-1}$

Example

The number of defects in a printed circuit board is hypothesized to follow a Poisson distribution. A random sample of $n = 60$ printed boards has be collected and the following number of defects observed

No of defects	Observed frequency (O_i)
0	32
1	15
2	9
3	4

The mean of the assumed Poisson distribution in this example is unknown and must be estimated from the sample data. The estimate of the mean is

$$(32 + (15 \times 1) + (9 \times 2) + (4 \times 3))/60 = 0.75$$

From the Poisson distribution with parameter 0.75 we may compute p_i the theoretical hypothesized probability associated with the ith class interval. p_i are found as follows.

$$p_1 = P(X=0) = \frac{e^{-0.75} 0.75^0}{0!} = 0.472$$

$$p_2 = P(X=1) = (e^{-0.75} 0.75^1)/1! = 0.354$$

$$p_3 = P(X=2) = \frac{e^{-0.75} 0.75^2}{2!} = 0.133$$

$$p_4 = P(X=3) = \frac{e^{-0.75} 0.75^3}{3!} = 0.041$$

Note:

$\Rightarrow p_4 = P(X \geq 3)$ $\quad 1 - p(X \leq 3)$

$= 1 - (p_1 + p_2 + p_3)$

The E_i are computed by multiplying the sample size $n=60$ times p_i i.e $E_i = np_i$

No of defects	p_i	E_i
0	0.472×60	28.32
1	0.354×60	21.24
2	0.133×60	7.98
3 (or more)	0.041×60	2.46

Since the expected frequency in the last cell is less than 3, we combine the last 2 cells

No of defects	O_i	E_i
0	32	28.32
1	15	21.24
2 (or more)	13	10.44

The degree of freedom are $K-p-1 = 3-1-1 = 1$

H_0: The form of the dist of defects is Poisson
H_1: The form of the dist of defects is not a Poisson

$\alpha = 0.05$

$$\chi_0^2 = \sum^{K} \frac{(O_i - E_i)^2}{E_i}$$

Reject H_0 if $\chi^2_{0\ 0.05,1} = 3.84$

$$\chi_0^2 = \frac{(32-28.32)^2}{28.32} + \frac{(15-21.24)^2}{21.24} + \frac{(13-10.44)^2}{10.44}$$

$$= 2.94$$

Since $\chi_0^2 = 2.94 < \chi^2_{0.05} = 3.84$ we are unable to reject H_0 that the distribution of defects printed on the circuit board is Poisson.

Contigency table tests

- It is used to test statistical independence of two categorical variables classified in a table.
- A contigency table is a two-way classification of data i.e a company has to choose among 3 pension plans. The management wish to know whether the preference of plans is independent of job classification using $\alpha = 0.05$
- The opinion of a random sample of 500 employees are shown below

Job classification	Pension plan 1	2	3	Total
Salary workers	160	140	40	340
Hourly workers	40	60	60	160
Total	200	200	100	500

H_0: preference is independent of salaried workers versus hourly job classification

H_1: preference is not independent of salaried versus hourly job classification.

$$\chi^2_{calculated} = \sum_{i=1}^{m} \sum_{j=1}^{p} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

m = rows

p = cols

Where O_{ij} is the observed frequency in row i column j

E_{ij} is the expected frequency in row i column j and is given by

$$E_{ij} = \frac{i^{th} \text{ row total} \times j^{th} \text{ column total}}{\text{sample size}}$$

$$= \frac{R_i - C_j}{n}$$

The critical value is given by $\chi^2_{crit} = \chi^2_{(m-1)\times(p-1)}{}^{\alpha}$. Reject H_0 iff $\chi^2_{calc} > \chi^2_{crit}$

$$E_{ij} \Rightarrow E_{11} = \frac{340 \times 200}{500} = 136$$

$$E_{12} = \frac{340 \times 200}{500} = 136$$

$$E_{13} = \frac{340 \times 100}{500} = 68$$

(b)

$$E_{21} = \frac{160 \times 200}{500} = 64$$

$$E_{22} = \frac{160 \times 200}{500} = 64$$

$$E_{23} = \frac{160 \times 100}{500} = 32$$

Job classification	Pension 1	2	3	Total
Salaried Workers	160 (136)	140 (136)	40 (68)	340
Hourly Workers	40 (64)	60 (64)	60 (32)	160
				500

$$\chi^2_{cale} = \frac{(160-136)^2}{136} + \frac{(140-136)^2}{136} + \frac{(40-68)^2}{68} + \frac{(40-64)^2}{64} + \frac{(60-64)^2}{64} + \frac{(60-32)^2}{32}$$

$$= 49.63$$

$$\chi^2_{chit} = \chi^2_{(m-1)(p-1)} = \chi^2_2(0.05) = 5.991$$

Conclusion: Since $\chi^2_0 = 49.63 > \chi^2_2(0.05) = 5.991$
We reject H_0 and conclude preference for pension plans is not dependent on job classification.

Exercise

1) Grades in statistics and OR courses taken simultaneously were as follows for a group of students.

(7)

Statistics grades	OR grades A	B	C	Others	Total
A	25	6	7	13	51
B	17	16	15	6	54
C	18	4	8	10	50
Others	10	8	11	20	49
Total	70	34	51	49	204

Are the grades in statistics and OR related? $\alpha = 0.01$ in teaching your conclusion.