

Week 8 Solutions

Tutorial Exercise.

- The data below come from a 1992 survey in Dayton Ohio on high school students' use of alcohol, tobacco and marijuana. Students were asked if they had ever used each of the drugs.

Alcohol use	Cigarette use	Marijuana use	
		Yes	No
Yes	Yes	911	538
	No	44	456
No	Yes	3	43
	No	2	279

(a) Let

$$p_{ijk} = P(\text{alcohol} = i, \text{smoking} = j, \text{marijuana} = k)$$

so that the table may be written as:

Alcohol use	Cigarette use	Marijuana use	
		Yes	No
Yes	Yes	x_{111}	x_{112}
	No	x_{121}	x_{122}
No	Yes	x_{211}	x_{212}
	No	x_{221}	x_{222}

Under complete independence

$$\begin{aligned} P(\text{alcohol} = i, \text{smoking} = j, \text{marijuana} = k) \\ = P(\text{alcohol} = i)P(\text{smoking} = j)P(\text{marijuana} = k) \end{aligned}$$

Marginalising over the categories 'smoking' and 'marijuana' we have

$$\begin{aligned} P(\text{alcohol} = i) &= \sum_j \sum_k P(\text{alcohol} = i, \text{smoking} = j, \text{marijuana} = k) \\ &= \sum_j \sum_k p_{ijk} \\ &= p_{i\bullet\bullet} \end{aligned}$$

Similarly,

$$P(\text{smoking} = j) = p_{\bullet j \bullet} \quad \text{and} \quad P(\text{smoking} = k) = p_{\bullet \bullet k}$$

Hence,

$$np_{ijk} = np_{i\bullet\bullet}p_{\bullet j\bullet}p_{\bullet\bullet k}.$$

We make the estimates

$$\hat{p}_{i\bullet\bullet} = \frac{x_{i\bullet\bullet}}{n} \quad \hat{p}_{\bullet j\bullet} = \frac{x_{\bullet j\bullet}}{n} \quad \text{and} \quad \hat{p}_{\bullet\bullet k} = \frac{x_{\bullet\bullet k}}{n}.$$

So the estimate for the cell (i, j, k) under independence is

$$np_{ijk} = n \frac{x_{i\bullet\bullet}}{n} \times \frac{x_{\bullet j\bullet}}{n} \times \frac{x_{\bullet\bullet k}}{n} = \frac{x_{i\bullet\bullet}x_{\bullet j\bullet}x_{\bullet\bullet k}}{n^2}$$

Then for cell $(1, 1, 1)$ we have

$$x_{1\bullet\bullet} = 911 + 538 + 44 + 456 = 1949,$$

$$x_{\bullet 1\bullet} = 911 + 538 + 3 + 43 = 1495,$$

$$x_{\bullet\bullet 1} = 911 + 44 + 3 + 2 = 960$$

and $n = 2276$. Hence, the estimate for the cell $(1, 1, 1)$ under independence is

$$\frac{x_{1\bullet\bullet}x_{\bullet 1\bullet}x_{\bullet\bullet 1}}{n^2} = \frac{1949 \times 1495 \times 960}{2276^2} = 539.98 \quad (\text{to 2 dp}).$$

- (b) If alcohol use and marijuana use are independent given smoking status, then (from the Categorical Data Analysis notes page 41 after reordering)

$$p_{ijk} = \frac{p_{ij\bullet}p_{\bullet jk}}{p_{\bullet j\bullet}}.$$

Then with $n = 2276$

$$\hat{p}_{11\bullet} = \frac{911 + 538}{n} = \frac{1449}{n}$$

$$\hat{p}_{12\bullet} = \frac{44 + 456}{n} = \frac{500}{n}$$

$$\hat{p}_{21\bullet} = \frac{3 + 43}{n} = \frac{46}{n}$$

$$\hat{p}_{22\bullet} = \frac{2 + 279}{n} = \frac{281}{n}$$

$$\hat{p}_{\bullet 11} = \frac{911 + 3}{n} = \frac{914}{n}$$

$$\hat{p}_{\bullet 12} = \frac{538 + 43}{n} = \frac{581}{n}$$

$$\hat{p}_{\bullet 21} = \frac{44 + 2}{n} = \frac{46}{n}$$

$$\hat{p}_{\bullet 22} = \frac{456 + 279}{n} = \frac{735}{n}$$

$$\hat{p}_{\bullet 1 \bullet} = \frac{1495}{n}$$

$$\hat{p}_{\bullet 2 \bullet} = \frac{781}{n}$$

Expected cell frequencies under conditional independence

$$np_{1,1,1} = (1, 1, 1) = \frac{\hat{p}_{11\bullet}\hat{p}_{\bullet 11}}{\hat{p}_{\bullet 1\bullet}} = \frac{1449 \times 914}{1495} = 885.877$$

$$(1, 1, 2) = \frac{\hat{p}_{11\bullet}\hat{p}_{\bullet 12}}{\hat{p}_{\bullet 1\bullet}} = \frac{1449 \times 581}{1495} = 563.123$$

$$(1, 2, 1) = \frac{\hat{p}_{12\bullet}\hat{p}_{\bullet 21}}{\hat{p}_{\bullet 2\bullet}} = \frac{500 \times 46}{781} = 29.449$$

$$(1, 2, 2) = \frac{\hat{p}_{12\bullet}\hat{p}_{\bullet 22}}{\hat{p}_{\bullet 2\bullet}} = \frac{500 \times 735}{781} = 470.551$$

$$(2, 1, 1) = \frac{\hat{p}_{21\bullet}\hat{p}_{\bullet 11}}{\hat{p}_{\bullet 1\bullet}} = \frac{46 \times 914}{1495} = 28.123$$

$$(2, 1, 2) = \frac{\hat{p}_{21\bullet}\hat{p}_{\bullet 12}}{\hat{p}_{\bullet 1\bullet}} = \frac{46 \times 581}{1495} = 17.877$$

$$(2, 2, 1) = \frac{\hat{p}_{22\bullet}\hat{p}_{\bullet 21}}{\hat{p}_{\bullet 2\bullet}} = \frac{281 \times 46}{781} = 16.551$$

$$(2, 2, 2) = \frac{\hat{p}_{22\bullet}\hat{p}_{\bullet 22}}{\hat{p}_{\bullet 2\bullet}} = \frac{281 \times 735}{781} = 264.449$$

(c) Test the fit of this model using Pearson's X^2 .

$$\chi^2 = \sum_i \sum_j \sum_k \left(\frac{o_{ijk}^2}{e_{ijk}} \right) - 2276 = 80.815$$

$$p\text{-value} = P(\chi_2^2 \geq 80.815) < 0.0001$$

So we have strong evidence against the conditional independence model.

(d) Express the conditional independence model in log-linear format.

$$\log p_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\beta\gamma)_{jk}$$

(No 3 factor and no alcohol times marijuana interaction term.)

The d.f = 2, 8 observations with 6 parameters and the deviance (form output) = 92.018.

- (e) Use the log-linear model output and an appropriate model to predict the chance of a randomly selected student having used alcohol but never smoked cigarettes or marijuana.

From the computer output, the model with all 2 factors interactions gives reasonable fit with Dev = 0.374 (1 d.f). No simpler log-linear model fits this data set. We want to estimate p_{122} , probability of yes to alcohol, no to cigarette and no to Marijuana. Empirical estimate is $\frac{456}{2276} = 0.2$. Using the model

$$\log(np_{122}) = \mu + \alpha_1 + \beta_2 + \gamma_2 + (\alpha\beta)_{12} + (\beta\gamma)_{22}$$

we have

$$\begin{aligned} \log(n\hat{p}_{122}) &= \hat{\mu} + \hat{\beta}_2 + \hat{\gamma}_2 + (\hat{\beta\gamma})_{22} \\ &= 6.814 - 3.015 - 0.525 + 2.847 \\ &= 6.121 \\ \hat{p}_{122} &= \frac{e^{6.12}}{2276} = 0.2 \end{aligned}$$

2. If R is completely independent of S and T then

$$P(R = i, S = j, T = k) = P(R = i)P(S = j, T = k)$$

or equivalently, $np_{ijk} = np_{i\bullet\bullet}p_{\bullet jk}$. We have, by assumption, that

$$\log(np_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k + (\beta\gamma)_{jk}.$$

Then

$$\begin{aligned} np_{i\bullet\bullet} &= \sum_j \sum_k \exp \{ \mu + \alpha_i + \beta_j + \gamma_k + (\beta\gamma)_{jk} \} \\ &= \exp \{ \mu + \alpha_i \} \sum_j \sum_k \exp \{ \beta_j + \gamma_k + (\beta\gamma)_{jk} \} \\ np_{\bullet jk} &= \sum_i \exp \{ \mu + \alpha_i + \beta_j + \gamma_k + (\beta\gamma)_{jk} \} \\ &= \exp \{ \mu + \beta_j + \gamma_k + (\beta\gamma)_{jk} \} \sum_i \exp \{ \alpha_i \} \end{aligned}$$

and

$$\begin{aligned} np_{i\bullet\bullet}p_{\bullet jk} &= n^{-1} \exp \{ \mu + \alpha_i \} \sum_j \sum_k \exp \{ \beta_j + \gamma_k + (\beta\gamma)_{jk} \} \\ &\quad \times \exp \{ \mu + \beta_j + \gamma_k + (\beta\gamma)_{jk} \} \sum_i \exp \{ \alpha_i \} \\ &= n^{-1} \exp \{ \mu + \alpha_i + \beta_j + \gamma_k + (\beta\gamma)_{jk} \} \sum_i \sum_j \sum_k \exp \{ \mu + \alpha_i + \beta_j + \gamma_k + (\beta\gamma)_{jk} \} \\ &= n^{-1} (np_{ijk}) \sum_i \sum_j \sum_k np_{ijk} = np_{ijk} \end{aligned}$$

as required (since $np_{ijk} = \exp \{ \mu + \alpha_i + \beta_j + \gamma_k + (\beta\gamma)_{jk} \}$ and $\sum_i \sum_j \sum_k p_{ijk} = p_{\bullet\bullet\bullet} = 1$).