last change: 13.nov.2009, 15:45

Multivariate Analysis
WS 2009/10

K. Utikal

1. give an example of a 2x2x2 table for three binary variables X, Y, Z with conditional independence of the X, Y on Z but not joint independence of X, Y from Z and vice versa

solution: the solution to this problem is very simple in terms of log linear models (deferred).

2. given a $J \times K$ confidence table with fixed total sample size (multinomial) write a simplified version of the test of independence of row and column variables

$$-2 \log LR = -2 \sum_{i,j} n_{ij} \left(\log \widehat{p}_{i\cdot} \widehat{p}_{\cdot j} - \log \widehat{p}_{ij} \right) \sim \chi^2 (I-1)(J-1)$$

solution:

$$\begin{aligned}
-2 \sum_{i,j} n_{ij} \left(\log \widehat{p}_{i\cdot} \widehat{p}_{\cdot j} - \log \widehat{p}_{ij} \right) &= -2 \sum_{i,j} n_{ij} \log \left(1 + \frac{\widehat{p}_{i\cdot} \widehat{p}_{\cdot j}}{\widehat{p}_{ij}} - 1 \right) \\
&\approx -2 \sum_{i,j} n_{ij} \left\{ \left(\frac{\widehat{p}_{i\cdot} \widehat{p}_{\cdot j}}{\widehat{p}_{ij}} - 1 \right) - \frac{1}{2} \left(\frac{\widehat{p}_{i\cdot} \widehat{p}_{\cdot j}}{\widehat{p}_{ij}} - 1 \right)^2 \right\} \\
&= -2 \sum_{i,j} n_{ij} \left(\frac{\widehat{p}_{i\cdot} \widehat{p}_{\cdot j} - \widehat{p}_{ij}}{\widehat{p}_{ij}} \right) + \sum_{i,j} n_{ij} \left(\frac{\widehat{p}_{i\cdot} \widehat{p}_{\cdot j} - \widehat{p}_{ij}}{\widehat{p}_{ij}} \right)^2 \\
&= -2 \sum_{i,j} n_{ij} \left(\frac{\widehat{p}_{i\cdot} \widehat{p}_{\cdot j} - \widehat{p}_{ij}}{n_{ij}/n} \right) + \sum_{i,j} n_{ij} \left(\frac{\widehat{p}_{i\cdot} \widehat{p}_{\cdot j} - \widehat{p}_{ij}}{\widehat{p}_{ij}} \right)^2 \\
&= \sum_{i,j} n_{ij} \left(\frac{\widehat{p}_{i\cdot} \widehat{p}_{\cdot j} - \widehat{p}_{ij}}{\widehat{p}_{ij}} \right)^2 = \sum_{i,j} n_{ij} \left(\frac{n\widehat{p}_{i\cdot} \widehat{p}_{\cdot j} - n\widehat{p}_{ij}}{n\widehat{p}_{ij}} \right)^2 \\
&= \sum_{i,j} n_{ij} \left(\frac{\widehat{\mu}_{ij} - n_{ij}}{n_{ij}} \right)^2 = \sum_{i,j} \frac{\left(\widehat{\mu}_{ij} - n_{ij} \right)^2}{n_{ij}}
\end{aligned}$$

3. for a two population problem (equal variances) write a simplified version of the LR-test that both means are equal.

(full) model:

$$\begin{aligned}
y_{ij} &\sim \mathcal{N}(\mu_{ij}, \sigma^2) \\
\mu_{ij} &= \mu + \alpha_j, \ i = 1, ..., n_j, \ j = 1, 2 \\
\alpha_1 + \alpha_2 &= 0
\end{aligned}$$

ml-estimators

$$\widehat{\mu}_{ij} = \overline{y}_{.j}$$

hypothesis:

$$H_0: \ \alpha_1 = \alpha_2 = 0$$

reduced model:

$$y_{ij} \sim \mathcal{N}(\mu_{ij}, \sigma^2)$$
$$\mu_{ij} = \mu, \ i = 1, ..., n_j, \ j = 1, 2$$

ml-estimator

$$\widetilde{\mu}_{ij} = \overline{y}_{..}$$

likelihood ratio test

$-2\log\text{LR} = n\log\frac{\|\mathbf{y}-\widetilde{\mu}\|^2}{\|\mathbf{y}-\widehat{\mu}\|^2} = n\log\frac{SS_{red}}{SS_{full}} \approx n\left(\frac{\|\mathbf{y}-\widetilde{\mu}\|^2}{\|\mathbf{y}-\widehat{\mu}\|^2} - 1\right)$

$= n\frac{\|\mathbf{y}-\widetilde{\mu}\|^2 - \|\mathbf{y}-\widehat{\mu}\|^2}{\|\mathbf{y}-\widehat{\mu}\|^2} = n\frac{\sum_{ij}(y_{ij}-\overline{y}_{..})^2 - \sum_{ij}\left(y_{ij}-\overline{y}_{.j}\right)^2}{\sum_{ij}\left(y_{ij}-\overline{y}_{.j}\right)^2} = n\frac{SS_{err\_red} - SS_{err\_full}}{SS_{err\_full}}$

for further simplification use ANOVA decomposition:

$(y_{ij} - \overline{y}_{..}) = \left(y_{ij} - \overline{y}_{.j}\right) + \left(\overline{y}_{.j} - \overline{y}_{..}\right)$

$\implies \sum_{ij}\left(y_{ij} - \overline{y}_{..}\right)^2 = \sum_{ij}\left(\overline{y}_{.j} - \overline{y}_{..}\right)^2 + \sum_{ij}\left(y_{ij} - \overline{y}_{.j}\right)^2$

$\implies \sum_{ij}\left(y_{ij} - \overline{y}_{..}\right)^2 - \sum_{ij}\left(y_{ij} - \overline{y}_{.j}\right)^2 = \sum_{ij}\left(\overline{y}_{.j} - \overline{y}_{..}\right)^2 = SS_{err\_red} - SS_{err\_full}$

$\implies$ Likelihood ratio test: $\quad -2\log\text{LR} = n\frac{\sum_{ij}\left(\overline{y}_{.j}-\overline{y}_{..}\right)^2}{\sum_{ij}\left(y_{ij}-\overline{y}_{.j}\right)^2} = (n_1 + n_2)\,\frac{n_1(\overline{y}_{.1}-\overline{y}_{..})^2 + n_2(\overline{y}_{.2}-\overline{y}_{..})^2}{\sum_i (y_{i1}-\overline{y}_{.1})^2 + \sum_i (y_{i2}-\overline{y}_{.2})^2}$

note that $n_1\left(\overline{y}_{.1} - \overline{y}_{..}\right)^2 + n_2\left(\overline{y}_{.2} - \overline{y}_{..}\right)^2 = n_1\left(\overline{y}_{.1} - \frac{n_1\overline{y}_{.1}+n_2\overline{y}_{.2}}{n_1+n_2}\right)^2 + n_2\left(\overline{y}_{.2} - \frac{n_1\overline{y}_{.1}+n_2\overline{y}_{.2}}{n_1+n_2}\right)^2$

$= n_1\left(n_2\frac{\overline{y}_{.1}-\overline{y}_{.2}}{n_1+n_2}\right)^2 + n_2\left(n_1\frac{\overline{y}_{.1}-\overline{y}_{.2}}{n_1+n_2}\right)^2 = \left(\overline{y}_{.1} - \overline{y}_{.2}\right)^2\,\frac{n_1 n_2^2 + n_2 n_1^2}{(n_1+n_2)^2} = \left(\overline{y}_{.1} - \overline{y}_{.2}\right)^2\,\frac{n_1 n_2}{n_1+n_2}$

hence -2LR $= (n_1 + n_2)\,\frac{n_1 n_2}{n_1+n_2}\,\frac{(\overline{y}_{.1}-\overline{y}_{.2})^2}{\sum_i(y_{i1}-\overline{y}_{.1})^2 + \sum_i(y_{i2}-\overline{y}_{.2})^2} = \frac{(\overline{y}_{.1}-\overline{y}_{.2})^2}{\frac{1}{n_1 n_2}\sum_i(y_{i1}-\overline{y}_{.1})^2 + \sum_i(y_{i2}-\overline{y}_{.2})^2}$

$= \left(\frac{\overline{y}_{.1}-\overline{y}_{.2}}{\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\frac{\sum_i(y_{i1}-\overline{y}_{.1})^2 + (y_{i2}-\overline{y}_{.2})^2}{n_1+n_2}}}\right)^2 \to \chi^2(1) \text{ as } n \to \infty$

Alternative: $F - test$

$$\mathtt{F} = \frac{(\mathtt{SSerr\_red} - \mathtt{SSerr\_full})/(\dim\beta - \dim\beta_1)}{\mathtt{SSerr\_full}/(\mathtt{n} - \dim\beta)} \sim F(\dim\beta - \dim\beta_1, \mathtt{n} - \dim\beta)$$

where

$$\texttt{F} = \frac{\sum_{ij}\left(\overline{y}_{\cdot j} - \overline{y}_{\cdot\cdot}\right)^2}{\frac{1}{n-2}\sum_{ij}\left(y_{ij} - \overline{y}_{\cdot j}\right)^2} \sim F(1, n-2)$$

Note: $t \sim t(\nu) \implies t^2 \sim F(1,\nu)$. It can be shown that

$$F = \left(\frac{\overline{y}_{\cdot 1} - \overline{y}_{\cdot 2}}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\frac{\sum_i (y_{i1}-\overline{y}_{\cdot 1})^2 + (y_{i2}-\overline{y}_{\cdot 2})^2}{n_1+n_2-2}}}\right)^2$$

4. given a one-way anova with on four groups

$$\begin{aligned} y_{ij} &= \mu + \alpha_j + \epsilon_{ij} \\ \epsilon_{ij} &\text{ iid} \sim \mathcal{N}(0,\sigma^2) \\ i &= 1,...,n_j,\; j = 1,...,5 \end{aligned}$$

write the LR-test for the null hypothesis

a) that all α_i are equal.

b) that the first two groups are equal and the two last groups are also equal but may be differing from the first two.

5. Rewrite the following estimations for a one-way with sum-restriction

Coefficients:	Estimate Std.	Error	t-value	Pr(>|t|)
(Intercept)	10.499056	0.004066	2582.2	<2e-16 ***
groupf2	1.006183	0.005750	175.0	<2e-16 ***
groupf3	-1.498169	0.005750	-260.5	<2e-16 ***

solution:

$$\left.\begin{aligned} \mu + \alpha_1 &= 10.50 \\ \alpha_2 - \alpha_1 &= 1.01 \\ \alpha_3 - \alpha_1 &= -1.50 \end{aligned}\right\} \implies \left\{\begin{aligned} \mu + \alpha_2 &= 11.51 \\ \mu + \alpha_3 &= 9.00 \end{aligned}\right\} \implies \mu + \overline{\alpha} = 10.33$$

$$\implies \left\{\begin{aligned} \mu + \alpha_1 - (\mu + \overline{\alpha}) &= 0.17 \\ \mu + \alpha_2 - (\mu + \overline{\alpha}) &= 1.18 \end{aligned}\right.$$

Coefficients:	Estimate Std.	Error	t value	Pr(>|t|)
(Intercept)	10.335061	0.002347	4402.7	< 2e-16 ***
groupf1	0.163995	0.003320	49.4	3.10e-15 ***
groupf2	1.170178	0.003320	352.5	< 2e-16 ***

6.