STAT 770 Sep. 23 Lecture
More on Dependence Structure in 2-way Tables

Reading for this lecture:

Chapter 2 in Agresti.

Today's topics (not separated into parts A, B):

(1) row-column independence for larger two-way tables;

(2) Sensitivity & specificity, 'prevalence'

(3) Case-control 2-way $2 \times K$ tables

(4) Conditional Association, Stratified $(K \times 2 \times 2)$ tables

Multinomial 2×2 LRT Example (Sec.2.1.6)

$Z_a \in \{0, 1\}$ random (Seat-belt use), $X_a \in \{0, 1\}$ (Fatal accident)

$$\beta = (\gamma, \lambda_1, \lambda_2) = \Big(p_{11}/(p_{+1}p_{1+}),\, p_{+1},\, p_{1+}\Big), \qquad K = 4$$

with $\gamma = p_{11}/(p_{+1}p_{1+}) = 1$ under row-column independence.

Model is $\pi_{11}(\gamma, \lambda) = \gamma\lambda_1\lambda_2$, $\pi_{+1} = \lambda_1$, $\pi_{1+} = \lambda_2$, $\pi_{++} = 1$.

unrestricted MLE $\widehat{p}_{zc} = N_{zc}/n$, restricted MLE maximizes

$$(\lambda_1\lambda_2)^{N_{11}}\,(\lambda_1 - \lambda_1\lambda_2)^{N_{01}}\,(\lambda_2 - \lambda_1\lambda_2)^{N_{10}}\,((1 - \lambda_1)(1 - \lambda_2))^{N_{00}}$$

which occurs (**check it!**) at $(\widehat{\lambda}_1)_r = N_{+1}/n$, $(\widehat{\lambda}_2)_r = N_{1+}/n$

$$X^2 = \sum_{z,c} \frac{(N_{zc} - E_{zc})^2}{E_{zc}},\ E_{11} = \frac{N_{+1}N_{1+}}{n},\ E_{1+} = \frac{N_{1+}}{n},\ E_{+1} = \frac{N_{+1}}{n}$$

Same method applies to larger 2-way tables !

Sensitivity and Specificity in 2×2 tables

Consider table with Z_a a diagnostic prediction `Y/N` and X_a the indicator of the actual Disease condition `D/N`.

`Sensitivity:` $P(Z_a = Y \,|\, X_a = D) = \pi_{YD}/\pi_{+D}$ **True Positive**

`Specificity:` $P(Z_a = N \,|\, X_a = N) = \pi_{NN}/\pi_{+N}$ **True Negative**

`Prevalence:` $P(X_a = D)$ **delicate case when this is small**

If $P(\mathtt{TP}) = 0.96$, $P(\mathtt{TN}) = 0.97$, $P(D) = .005$, test pos: **then**
$P(X_a = D \,|\, Z_a = Y) = .005 * .96/(.005 * .96 + .995 * .03) = 0.139$

Very low prevalence leads to low Positive Predictive Value

Case-Control Studies, $2 \times K$

Collect records on Risk-factor categories $k = 1, \ldots, K$ separately for Disease Cases and for *comparable* Controls

Here row-totals $n_z = N_{z+}$ are fixed, often $n_C / n_D = 1$ or 2

Example (*Br.Med.J. 1950*): D=Lung Cancer, k=1 ⇔ Smoking

	Smoker	Non
Cases	688	21
Controls	650	59

Hugely influential, OR = 2.97; other similar studies with stricter 'smoker' def'n had **higher** OR

Critics (including R.A.Fisher!) asked whether omitted Risk-factors defining population subgroups would explain the OR

Conditional Association, Stratification/Confounding

Confounding: in Cancer/Smoking case-control studies with higher OR's, Cornfield (1956) asked: could there be K pop subgroups with different **conditional** ORs that account for overall OR ?

Notation: π_{kzx} cell-probs, N_{kzx} counts,

$n_z = N_{+z+}$ or N_{++x} fixed

Conditional OR: separate Odds Ratio for population subgroup k

$$\text{OR} = \theta = \frac{\pi_{+11}\pi_{+00}}{\pi_{+01}\pi_{+10}}, \quad \theta_k = \frac{\pi_{k11}\pi_{k00}}{\pi_{k01}\pi_{k10}}$$

When overall OR is ≥ 10 , some subgroup ORs would have to be absurdly large !

Conditional Association, Stratification $K \times 2 \times 2$

`Sec.2.3.2 Race & Death Penalty` covered in `R` Script in file `R-ContingTable.RLog` using

separately coded `OR` function and `apply`

Separate Odds Ratios 0.431 and 0 stratified by Victim's Race

Combined Odds Ratio 1.45 **instance of Simpson's Paradox**

Small Additional Use of Univariate Delta Method

In last lecture, we found it convenient to talk about approximate normal distribution of log Odds Ratio estimate $\widehat{\beta}_1$ and estimated standard error $\widehat{\sigma}_{logOR}$ for Wald-type CI $\widehat{\beta}_1 \pm 1.96\,\widehat{\sigma}_{logOR}$.

Can form confidence interval for Odds Ratio $\psi = e^{\beta_1}$ in 2 ways:

(i) Transform the previous interval:

$$\left(\text{exp}\left\{\widehat{\beta}_1 - 1.96\,\widehat{\sigma}_{logOR}\right\}, \text{exp}\left\{\widehat{\beta}_1 + 1.96\,\widehat{\sigma}_{logOR}\right\} \right)$$

(ii) Wald interval for transformed parameter: $e^{\widehat{\beta}_1} \pm 1.96\,\widehat{\sigma}_{OR}$

where Delta Method gives $\sqrt{n}\,(e^{\widehat{\beta}_1} - e^{\beta_1}) \approx e^{\beta_1}\sqrt{n}\,(\widehat{\beta}_1 - \beta_1)$

which implies $\widehat{\sigma}_{OR} = e^{\widehat{\beta}_1} \cdot \widehat{\sigma}_{logOR}$