STAT 770 Sep. 9 Lecture Part B
LRT in Contingency Table Setting

Reading for this lecture: Agresti Ch. 2 through Sec. 2.2, plus Ch. 16 through Sec. 16.3.4.

General X^2 form of LRT for Contingency Tables.

Special cases of row-column independence in 2×2 tables,

differences between proportions

LRT in Contingency Table Setting

Recall: $Y_a = (Z_a, X_a)$ Multinomial with probabilities $p_{z,c}$

$\theta = \{p_{z,c} : \ (z,c) \in \mathcal{K}\}, \quad \beta = (\theta_1, \ldots, \theta_d), \ d = |\mathcal{K}| - 1$

$$L(\beta; \underline{\mathbf{Y}}) \ = \ (\text{multinom. coeff.}) \cdot \prod_{(z,c)\in\mathcal{K}} p_{z,c}^{N_{z,c}}$$

Lower dimensional model $p_{z,c} = \pi_{z,c}(\gamma_0, \lambda)$ is Null Hypothesis
(Many examples will follow !)

So LRT $\Lambda = G^2 = -2 \log \left[L(\{\pi_{z,c}(\gamma_0, \hat{\lambda}_r\}) \Big/ L(\{\hat{p}_{x,c}\}) \right]$

$$= 2 \sum_{(z,c)\in\mathcal{K}} N_{z,c} \log \Big(\frac{N_{z,c}/n}{\pi_{z,c}(\gamma_0, \hat{\lambda}_r)} \Big)$$

Consequences for General Models

(I) G^2 is a **goodness-of-fit** test statistic for the model

$p_{z,c} = \pi_{z,c}(\gamma_0, \lambda)$ (λ general, $d-q$ dimensional, unknown)

(II) $G^2 = 2\sum_{k\in\mathcal{K}} N_k \log\left(\frac{N_k}{n\tilde{\pi}_k}\right)$ with $\tilde{\pi}_k$ $\sqrt{n}$-consistent for p_k which means the same as $\sqrt{n}\,(\tilde{\pi}_k - p_k) = O_P(1)$ or $\sqrt{n}(\tilde{\pi}_k - N_k/n) = O_P(1)$ for large n which implies that as $n \to \infty$

$$G^2 = \sum_{k\in\mathcal{K}} \frac{(N_k - n\tilde{\pi}_k)^2}{n\tilde{\pi}_k} + o_P(1) = \sum_{k\in\mathcal{K}} \frac{(O_k - E_k)^2}{E_k} + o_P(1)$$

and Wilks' Theorem gives $X^2 = \sum_{k\in\mathcal{K}} \frac{(O_k - E_k)^2}{E_k} \xrightarrow{\mathcal{D}} \chi^2_q$

Proof of Assertion (II) on Last Slide

This is a Taylor Series proof, using $N_k/(np_k) - 1 = o_P(1)$ and

$$N_k \log(N_k/(n\tilde{\pi}_k)) = -N_k \log(1 - \tfrac{N_k - n\tilde{\pi}_k}{N_k})$$

$$= N_k\Big[\tfrac{N_k - n\tilde{\pi}_k}{N_k} + \tfrac{(N_k - n\tilde{\pi}_k)^2}{2N_k^2} + O_P\Big(\tfrac{(N_k - n\tilde{\pi}_k)^3}{N_k^3}\Big)\Big]$$

$$= N_k - n\tilde{\pi}_k + \tfrac{(N_k - n\tilde{\pi}_k)^2}{2n\tilde{\pi}_k} + O_P\Big(\tfrac{(N_k - n\tilde{\pi}_k)^3}{n^2}\Big)$$

since $\log(1-z) = -z - \frac{z^2}{2} - O_P(z^3)$ for small z.

Sum over $k \in \mathcal{K}$ to find [using $\sum_k N_k = n = \sum_k (n\tilde{\pi}_k)$] that

$$G^2 = 2 \sum_{k \in \mathcal{K}} N_k \log\left(\frac{N_k}{n\tilde{\pi}_k}\right) = \sum_{k \in \mathcal{K}} \frac{(N_k - n\tilde{\pi}_k)^2}{n\tilde{\pi}_k} + O_P(n^{-1/2})$$

Row-column independence in 2×2 Tables

Here $Z_a \in \{0,1\}$ are random, $\mathcal{K} = \{0,1\}^2$, $K = 4$ and

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

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

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

The unrestricted MLE is $\widehat{p}_{zc} = N_{zc}/n$, $z, c = 0, 1$, while the restricted MLE maximizes the likelihood

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

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

$X^2 \overset{\mathcal{D}}{\approx} \chi^2_1$ **from (II) above has the familiar form**
$\sum_{(z,c)} (O_{z,c} - E_{z,c})^2/E_{z,c}$, **with** $O_{z,c} = N_{z,c}$, $E_{z,c} = n\pi_{z,c}$

R Code to Check χ^2_1 Distribution

```
>tmp=array(rmultinom(1e5, 40, prob=c(.16,.24,.24,.36)), c(2,2,1e5))
   aux = apply(tmp,3, function(tab2)  c(chisq.test(tab2)$stat,
               chisq.test(tab2, corr=F)$stat) )
   round( rbind(Xsq.corr = quantile(aux[,1], prob=(1:9)/10),
               Xsq = quantile(aux[,2], prob=(1:9)/10),
               chisq = qchisq((1:9)/10, 1) ), 3)
           10%   20%   30%   40%   50%   60%   70%   80%   90%
Xsq.corr 0.000 0.000 0.004 0.038 0.111 0.264 0.508 0.938 1.742
Xsq      0.017 0.067 0.152 0.302 0.444 0.750 1.125 1.710 2.824
chisq    0.016 0.064 0.148 0.275 0.455 0.708 1.074 1.642 2.706
```

Similar accuracy when $n = 80$

NB Yates over-corrects badly, used only when conditioning on marginals!!

Testing Equality of Row Proportions in 2×2 Table

In this setting, Z_a values are fixed by design, so the row-totals $N_{z+} = n_z$ are nonrandom and known, and $N_{z1} \sim \text{Binom}(n_z, \pi_z)$, with $\pi_z = p_{z1}/p_{z+}$.

Here we can take $\beta = (\gamma, \lambda)$ in different ways,

with $H_0 : \gamma = 1$ and $\lambda = \pi_0$ under H_0.

Example 1. Relative Risk, RR: $\beta = (\pi_1/\pi_0,\, \pi_0)$

Example 2. Odds Ratio, OR: $\beta = \Big([\pi_1/(1-\pi_1)]\Big/[\pi_0/(1-\pi_0)],\, \pi_0\Big)$

In RR, the restricted MLE (under $\gamma = 1$) maximizes

$$\Big[\prod_{z=0}^{1} \binom{n_z}{N_{z1}}\Big] \pi_0^{N_{11}+N_{01}} (1-\pi_0)^{N_{10}+N_{00}} = c \cdot \pi_0^{N_{+1}} (1-\pi_0)^{N_{+0}}$$

In both RR and OR, $\hat{\lambda} = N_{+1}/n$ **and** $E_{z,c} = n_z \pi_0^c (1-\pi_0)^{1-c}$