STATS 305B: Methods for Applied Statistics I **Winter 2016/17**

Lecture 4: January 17

Lecturer: Jonathan Taylor *Scribes: Kenneth Tay*

4.1 (Agresti 2.2) Odds Ratio for 2×2 Table

In a 2×2 contingency table, the odds ratio θ is defined by

$$\theta := \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}.$$

In the independent rows (sampling) model, this is the same as

$$\theta := \frac{\pi_1(1-\pi_1)}{\pi_2(1-\pi_2)},$$

where $\pi_i = P(Y = \text{success} \mid X = i)$.

Useful Fact: The odds ratio does not change when we switch the rows and columns! The example below demonstrates this.

4.1.1 Example: Lung Cancer and Smoking

This is a case-control set-up: Each lung cancer patient was asked whether he/she was a smoker. Then, for each lung cancer patient, a random patient without lung cancer (matched on other covariates) was asked whether he/she was a smoker. This is an example of an independent columns model.

The counts obtained were as follows:

	Lung cancer	No lung cancer
Smoker	688	650
Non-smoker	21	59
Total	709	709

From the table, we can compute

$$\begin{aligned} \hat{P}(\text{smoker} \mid \text{lung cancer}) &= \frac{688}{709}, \\ \hat{P}(\text{smoker} \mid \text{no lung cancer}) &= \frac{650}{709}, \\ \therefore \hat{\theta} &\approx 3.0. \end{aligned}$$

Assuming lung cancer is rare in the population, we can use the rare disease hypothesis to obtain an estimate of the relative risk:

$$\hat{r} = \frac{P(\text{lung cancer} \mid \text{smoker})}{P(\text{lung cancer} \mid \text{non-smoker})} \approx \hat{\theta} \approx 3.0.$$

Why not use Bayes rule to compute $\hat{r}$? In order to do so, we need an estimate for the probabilty of lung cancer, $\hat{P}(\text{lung cancer})$. However, due to study design, our estimate would be 0.5, which is clearly wrong.

4.2 (Agresti 3.1) Inference for Two-Way Tables

4.2.1 Estimation using asymptotic normal approximations

Say we have an 2×2 contingency table.

4.2.1.1 Odds ratio

For large samples, $\hat{\theta}$ has an asymptotic normal distribution around θ. However, convergence takes a long time as the sampling distribution of $\hat{\theta}$ can be highly skewed. (E.g. if $\theta = 1$, the estimate $\hat{\theta}$ can never be too much smaller than θ as it must be non-negative, but it can be much much bigger.) Hence, it is better to look at $\log \hat{\theta}$ instead, then exponentiate to get results for $\hat{\theta}$.

For large samples, we have

$$\log \hat{\theta} \approx \mathcal{N}\left(\log \theta, \frac{1}{Y_{11}} + \frac{1}{Y_{12}} + \frac{1}{Y_{21}} + \frac{1}{Y_{22}}\right).$$

Denoting the variance above by $\hat{\sigma}^2$, we get Wald confidence intervals for $\log \theta$ and θ:

$$\begin{aligned} \log \theta &: \left(\log \hat{\theta} - z_{\alpha/2}\hat{\sigma}, \log \hat{\theta} + z_{\alpha/2}\hat{\sigma}\right), \\ \theta &: \left(\hat{\theta} \cdot \exp[-z_{\alpha/2}\hat{\sigma}], \hat{\theta} \cdot \exp[z_{\alpha/2}\hat{\sigma}]\right). \end{aligned}$$

(Note: The result above applies for all 4 designs (i.e. Poisson, multinomial, independent rows, independent columns).)

4.2.1.2 Difference of proportions

Here, we have the normal approximation

$$\hat{\pi}_1 - \hat{\pi}_2 \approx \mathcal{N}\left(\pi_1 - \pi_2, \frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}\right).$$

(This applies for all 4 designs.)

4.2.1.3 Relative risk

Here, we have the normal approximation

$$\frac{\hat{\pi}_1}{\hat{\pi}_2} \approx \mathcal{N}\left(\frac{\pi_1}{\pi_2}, \frac{\hat{\pi}_1}{(1-\hat{\pi}_1)n_1} + \frac{\hat{\pi}_2}{(1-\hat{\pi}_2)n_2}\right).$$

4.2.2 (Agresti 3.2.6) Using profile likelihood for odds ratio

Assume we are in the multinomial model for a 2 contingency table. This model is typically parametrized by

$$\pi = \begin{pmatrix} \pi_{11} & \pi_{12} \\ \pi_{21} & \pi_{22} \end{pmatrix}.$$

(Note that there are only 3 free parameters since they have to sum up to 1.) Another way to parametrize this model is through the parameters $(\theta, \pi_{1+}, \pi_{+1})$. Each combination of these parameters gives rise to a π, which we denote by $\pi = h(\theta, \pi_{1+}, \pi_{+1})$.

For each t, let

$$\begin{pmatrix} \hat{\pi}_{1+}(Y,t) \\ \hat{\pi}_{1+}(Y,t) \end{pmatrix} := MLE \begin{pmatrix} \pi_{1+} \\ \pi_{+1} \end{pmatrix}$$

given $\theta = t$. Then, the profile likelihood for θ is

$$L_{Profile}(\theta \mid Y) = L\Big(h(\theta, \hat{\pi}_{1+}(Y,\theta), \hat{\pi}_{1+}(Y,\theta)) \mid Y\Big),$$

and the $(1-\alpha)$-level profile confidence interval for θ is given by

$$\Big\{\theta : -2\left[\log L_{Profile}(\theta \mid Y) - \log L(\hat{\pi} \mid Y)\right] \leq \chi^2_{1,1-\alpha}\Big\},$$

where $\hat{\pi}$ is the unrestricted MLE for π.

4.3 (Agresti 3.2) Tests of Independence

Say we have an $I \times J$ contingency table, and that we are in a Poisson or multinomial model. We wish to test whether the response and explanatory variables are independent. (This does not make sense for the independent rows or columns model.)

Here, the null hypothesis is $H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$ for all i, j.

Let $E_{ij} = Y_{..}\hat{\pi}_{i+}\hat{\pi}_{+j}$, i.e. the expected number of observations in cell (i,j) under H_0.

There are 2 common test statistics we use:

1. **Pearson's χ^2.**

$$X^2 = \sum_{i,j} \frac{(Y_{ij} - E_{ij})^2}{E_{ij}} \overset{H_0}{\approx} \chi^2_{(I-1)(J-1)}.$$

2. **Likelihood ratio test statistic.**

$$G^2 = \sum_{i,j} Y_{ij} \log\left(\frac{Y_{ij}}{E_{ij}}\right) \overset{H_0}{\approx} \chi^2_{(I-1)(J-1)}.$$

The 2 statistics above are asymptotically equivalent.