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

Lecture 3: January 13

Lecturer: Jonathan Taylor *Scribes: Kenneth Tay*

3.1 (Agresti 1.6) Bayesian Models for Proportions

3.1.1 Beta-Binomial Model

A common Bayesian model for a binomial proportion π is the beta-binomial model. Here, we assume that the prior distribution on π is

$$\pi \sim \text{Beta}(\alpha_1, \alpha_2), \quad \text{i.e. prior density} \;\; g(\pi) \propto \pi^{\alpha_1 - 1}(1-\pi)^{\alpha_2 - 1}, 0 \leq \pi \leq 1.$$

The likelihood is given by

$$Y \mid \pi \sim \text{Binom}(n, \pi), \quad \text{i.e. mass function} \;\; f(y \mid \pi) = \binom{n}{y}\pi^y (1-\pi)^{n-y}, y = 0, 1, \ldots, n.$$

Hence, the posterior distribution of π given the data has density

$$\begin{aligned} h(\pi \mid y) &\propto f(y \mid \pi) g(\pi) \\ &\propto \pi^{\alpha_1 + y - 1}(1-\pi)^{\alpha_2 + n - y - 1}, \\ \pi \mid y &\sim \text{Beta}(\alpha_1 + y, \alpha_2 + n - y). \end{aligned}$$

This is an example of **conjugacy**, i.e. a Bayesian model where the posterior distribution is in the same family as the prior.

3.1.2 Multinomial-Dirichlet Model

This is basically a generalization of the beta-binomial model from 2 categories to k categories.

The prior distribution for π is given by

$$\pi \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_k), \quad \text{i.e. prior density} \;\; g(\pi) \propto \prod_{i=1}^k \pi^{\alpha_i - 1}.$$

The likelihood is given by $Y \mid \pi \sim \text{Multinom}(n, \pi)$. Using the same technique as the previous subsection, the posterior distribution turns out to be $\pi \mid y \sim \text{Dirichlet}(\alpha_1 + y_1, \ldots, \alpha_k + y_k)$.

3.2 (Agresti 2.1) Contingency Tables

Say we have two categorical variables: X with I levels, and Y with J levels. A **contingency table** is simply an $I \times J$ table summary of the data, where cell (i, j) records the number of cases with explanatory variable level i and response level j.

Typically X is an explanatory variable and Y is a response variable, and we want to see if X has any effect on Y.

Some notation:

- $Y_{ij} :=$ number of observations in cell (i, j).
- $\pi_{ij} := P\{X = i, Y = j\}$ (joint probabilities).
- $\pi_{i+} := P\{X = i\}$, $\pi_{+j} := P\{Y = j\}$ (marginal probabilities).
- $\pi_{j|i} = P\{Y = j \mid X = i\}$ (conditional probabilities).

3.2.1 Sampling Models

There are a number of ways that we could go about sampling subjects.

1. **Poisson sampling model.** Fix a λ, and let $N \sim \text{Poisson}(\lambda)$. Then, get a simple random sample of size N and check which cell of the contingency table each subject falls in.

   It can be shown that $Y_{ij} \overset{ind.}{\sim} \text{Poisson}(\lambda_{ij})$.
2. **Multinomial sampling model.** Instead of letting the sample size be Poisson-distributed, let it be some fixed number n. (Row and column counts are not fixed.) Then we have $Y \sim \text{Multinom}\left(n, (\pi_{ij})\right)$.
3. **Independent multinomial sampling model.** This can be used when either row or column totals are fixed. If row totals are fixed, then we can model the counts by $Y[i,] \sim \text{Multinom}(Y_{i+}, \pi_{\cdot|i})$. (Similar set-up if column totals are fixed.

3.2.2 (Agresti 2.2) Comparing Two Proportions

This is for the special case of the 2×2 contingency table. Say we have an explanatory variable consisting of two groups (labeled 1 or 2), and a response variable ("Yes" or "No").

Define $\pi_1 := \theta_{Yes|1}$, $\pi_2 := \theta_{Yes|2}$. Then there are few different ways to measure the relationship between π_1 and π_2:

1. **Difference of proportions** $\pi_1 - \pi_2$.
2. **Relative risk** $r := \dfrac{\pi_1}{\pi_2}$.
3. **Odds ratio** $\theta := \dfrac{Odds(\pi_1)}{Odds(\pi_2)} := \dfrac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$.

In the Poisson and Multinomial sampling models, $\theta = 1$ is equivalent to X and Y being independent. In the independent multinomial sampling model, $\theta = 1$ and $r = 1$ individually imply that $\pi_1 = \pi_2$.

Note that when π_1 and π_2 are small, $\theta \approx r$ (known as *rare disease hypothesis*). This is important because sometimes one of them can be estimated well while the other cannot.