MATH 60604A
Statistical modelling
§ 4e - Contingency tables

HEC Montréal
Department of Decision Sciences

Two-way contingency tables

The most common format for aggregated count data with categorical predictors is **contingency tables**, in which each cell is a count for a given combination of levels of the categorical variables.
Consider X_1 and X_2 two categorical variables with J and K levels and the associated contingency table.

$$
\begin{array}{c|cccc}
 & \mathrm{X}_2 = 1 & \mathrm{X}_2 = 2 & \cdots & \mathrm{X}_2 = K \\
\hline
\mathrm{X}_1 = 1 & Y_{11} & Y_{12} & \cdots & Y_{1K} \\
\mathrm{X}_1 = 2 & Y_{21} & Y_{22} & \cdots & Y_{2K} \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
\mathrm{X}_1 = J & Y_{J1} & Y_{J2} & \cdots & Y_{JK}
\end{array}
$$

Testing independence in contingency tables

- We consider two competing Poisson models: under $\mathscr{H}_0$, a model with the two categorical variables, but **no interaction**. For $(j = 1, \ldots, J; k = 1, \ldots, K)$, the mean number in cell (j, k) is

$$\mu_{jk} = \exp\left(\beta_0 + \alpha_j \mathbf{1}_{\mathrm{X}_1 = j} + \gamma_k \mathbf{1}_{\mathrm{X}_2 = k}\right),$$

  with $\alpha_1 = 0$ and $\gamma_1 = 0$ for identifiability.
- The alternative model is the saturated model (including an additional interaction between X_1 and X_2). The null hypothesis of independence is simply a test that the additional parameters associated to the interaction are equal to zero.

Main effects and saturated models for two-way contingency tables

- Under $\mathscr{H}_0$, the model includes only X_1 and X_2 (main effects).
  - One can show that the fitted values of the null model are the product of the sample proportion in each line/column.
  - We denote the fitted values for cell (j, k) is denoted $\widehat{\mu}_{jk}$.
- The saturated model, under the alternative $\mathscr{H}_1$, includes additional parameters for the interaction.
  - the saturated model has $n = JK$ parameters and the fitted values are simply Y_{jk}.

Statistics for the test of independence in contingency tables

- The likelihood ratio test statistic is the deviance,

$$D = 2\sum_{j=1}^{J}\sum_{k=1}^{K} Y_{jk}\ln\left(\frac{Y_{jk}}{\widehat{\mu}_{jk}}\right),$$

  which follows $\chi^2_{(J-1)(K-1)}$ under the null hypothesis of independence.
- Alternatively, we can use the score test statistic (with the same null distribution),

$$X^2 = \sum_{j=1}^{J}\sum_{k=1}^{K}\frac{(Y_{jk}-\widehat{\mu}_{jk})^2}{\widehat{\mu}_{jk}}.$$

Political affiliation in USA

We consider the two by three contingency table of political affiliation by party in the US as a function of gender in 2000.

	Party Identification			
Gender	Democrat	Independent	Republican	Total
Females	762	327	468	1557
	(703.7)	(319.6)	(533.7)	
Males	484	239	477	1200
	(542.3)	(246.4)	(411.3)	
Total	1246	566	945	2757

The number in parenthesis represent the fitted values from the additive Poisson model without interaction (main effects).

Data reproduced from Table 2.5, Agresti (2007), *An Introduction to Categorical Data Analysis*, Wiley.

Result of the independence test for political affiliation

- By fitting the model without interaction, we get the X^2 and the likelihood ratio statistic in the output (30.07 and 30.02, respectively).
- Both should behave as χ_2^2 variables if gender was independent of political affiliation.
- The p-values are smaller than 10^{-4} and we conclude against independence, meaning gender has an effect on political affiliation.

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	2	30.0167	15.0083
Scaled Deviance	2	30.0167	15.0083
Pearson Chi-Square	2	30.0701	15.0351
Scaled Pearson X2	2	30.0701	15.0351