Contingency Tables

Classification of subjects based on two (or more) factors.

A/B	1	$\cdots$	J
1	Y_{11}	$\cdots$	Y_{1J}
$\vdots$	$\vdots$	$\ddots$	$\vdots$
I	Y_{I1}	$\cdots$	Y_{IJ}

Table: An example of two-way contingency table

- Data are organized in a cross-classified table
- Variables are categorical

Primary goal: Test whether the two variables are associated (for more factors, test different types of association)

Sampling Distributions

The probability model for a contingency table depends on the study design.

- Poisson model
  - Number of new cases in gender-by-age groups ($2 \times J$) during a year in an epidemiological study.
- Multinomial model
  - Cross-sectional study of patients with melanoma. Patients are classified based on tumor types and locations.
- Product multinomial model
  - Case control study of lung cancer and smoking. Subjects in each group are further divided into the smoking group and non-smoking group.

- The three probability models are highly related.
- In fact, they are equivalent in modeling cell means.
- For two-way contingency table, they all lead to the same χ^2 test.

Poisson Model

- Observe a set of Poisson process, one for each cell of the contingency table. No prior knowledge regarding the total number of observations.

$$Y_{ij} \sim Poisson(\mu_{ij})$$

- If no constraint on μ_{ij}, this is a saturated model with IJ observations and IJ parameters.

$$\hat{\mu}_{ij} = y_{ij}$$

- Under the independent assumption, we have $\mu_{ij} = \mu_{i\cdot}\mu_{\cdot j}$. Therefore, we have $(i = 1, \cdots I; j = 1, \cdots, J)$

$$\log \mu_{ij} = \mu + \alpha_i + \beta_j$$

where $\sum \alpha_i = \sum \beta_j = 0$

- This leads to a Poisson regression model with $I + J - 1$ parameters.
- Testing the association is equivalent to checking the goodness-of-fit of the additive model.

$$D(\text{or } G) \sim \chi^2((I-1)(J-1)), \text{ under } H_0$$

- If D (or G) is larger than some threshold, we reject the null, and claim there is significant association between the two factors.

Multinomial Model

If the total counts $Y_{++} = \sum_{i=1}^{I} \sum_{j=1}^{J} Y_{ij} = n$ is fixed, the joint distribution of $(Y_{11}, \cdots, Y_{IJ})$ is multinomial.

$$f(\boldsymbol{y}|y_{++} = n) = n! \prod_{i=1}^{I} \prod_{j=1}^{J} \frac{\pi_{ij}^{y_{ij}}}{y_{ij}!}.$$

For a saturated model, the number of parameters is $IJ - 1$.

- Under the independent assumption, we have $\pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}$. Therefore, we can obtain the MLE as

$$\hat{\pi}_{i\cdot} = \frac{Y_{i+}}{n}, \hat{\pi}_{\cdot j} = \frac{Y_{+j}}{n}$$

- Use the likelihood ratio test, we have

$$LR \sim \chi^2((I-1)(J-1)), \text{ under } H_0$$

- If LR is larger than some threshold, we reject the null, and claim there is significant association between the two factors.

Product Multinomial Model

If there are more fixed marginal totals than just the overall total n, then an appropriate model is a product multinomial model.

- In Product Multinomial model (with fixed row totals, for example), the saturated model has
$$\hat{\pi}_{j|i} = \frac{y_{ij}}{n_{i+}}$$
- Under the independent assumption, the model is homogeneous across rows
$$\pi_{j|i} = \pi_{\cdot j}$$
- This corresponds to a multinomial model with only intercepts.
- Use deviance analysis to test the independence hypothesis

$$LR \sim \chi^2((I-1)(J-1)), \text{ under } H_0$$

Example: Malignant Melanoma Study

The following data are from a cross-sectional study of patients with a form of skin cancer called malignant melanoma. For a sample of 400 patients, the site of the tumor and its histological type were recorded.

	Site		
Tumor	HeadNeck	Trunk	Extremities
Hutchinson	22	2	10
Superficial	16	54	115
Nodular	19	33	73
Indeterminate	11	17	28

The question of interest is whether there is any association between tumor type and site.