BMI206: Statistical Methods for Bioinformatics

Sampling in $I \times J$ Tables

Let $\mathbf{Y} = (Y_{11}, Y_{21}, \ldots, Y_{IJ})$ be the random vector of cell counts from a study that produces an $I \times J$ table. Let $\mathbf{n} = (n_{11}, n_{21}, \ldots, n_{IJ})$ denote the corresponding observed counts.

1. When the counts are collected over a fixed time/place, but none of the totals are fixed, the sampling scheme is **Poisson**:

$$Y_{ij} \sim Pois(\mu_{ij}), \text{ for all } i, j$$

The cell counts are separate, independent Poisson random variables, each with its own rate parameter μ_{ij}. Hence, the probability density function for all of the counts is the product of the independent Poisson densities:

$$P(\mathbf{Y} = \mathbf{n}) = \prod_i \prod_j \frac{\exp(-\mu_{ij})\mu_{ij}^{n_{ij}}}{n_{ij}!}$$

2. When the total sample size ($n = n_{..}$) is fixed, but none of the marginal totals are, the sampling scheme is **Multinomial**:

$$\mathbf{Y} \sim Mult(\pi_{11}, \pi_{21}, \ldots, \pi_{IJ})$$

The cell counts form a single Multinomial sample with probability density function:

$$P(\mathbf{Y} = \mathbf{n}) = \frac{n!}{n_{11}!n_{21}!\ldots n_{IJ}!} \prod_i \prod_j \pi_{ij}^{n_{ij}}$$

- Since $\sum_i \sum_j \pi_{ij} = 1$, $Y_{ij} \sim Bin(n_{..}, \pi_{ij})$. In other words, the probability $P(Y_{ij} = n_{ij})$ can be computed from the Binomial density function with $\pi = \pi_{ij}$ and $1 - \pi = \sum_{k \neq i} \sum_{l \neq j} \pi_{kl}$.

3. When the row counts $(n_{1.}, n_{2.}, \ldots, n_{I.})$ are fixed and each row is sampled independently, the sampling scheme is **Product Multinomial**:

$$(Y_{i1}, Y_{i2}, \ldots, Y_{iJ}) \sim Mult(\pi_{j|i}) = Mult(\pi_{1|i}, \pi_{2|i}, \ldots, \pi_{J|i}), \text{ for all } i$$

The cell counts in each row form a single, independent Multinomial sample. Hence, the probability density function is the product of the I individual Multinomial densities:

$$P(\mathbf{Y} = \mathbf{n}) = \prod_i \frac{n_{i.}!}{n_{i1}!n_{i2}!\ldots n_{iJ}!} \prod_j \pi_{j|i}^{n_{ij}}$$

- If $J = 2$ (*i.e.* the unfixed variable is binary), then each row is an independent Binomial sample, and we call the sampling scheme (Product) **Binomial**.
- The fixed marginal totals can be for the independent variable (*e.g.* a clinical trial), which is conventionally depicted in the rows, or for the dependent variable (*e.g.* a case-control study), which is depicted in the columns. In other words, all of the above holds with the roles of i and j switched.

4. When the row and column totals are both fixed in a 2 × 2 table, the sampling scheme is **Hypergeometric**. The probability density function is simply the probability for one of the cells, since the other three are determined by the fourth plus the marginal totals:

$$P(Y_{11} = n_{11}) = \frac{n_{1.}!n_{2.}!n_{.1}!n_{.2}!}{n_{11}!n_{12}!n_{21}!n_{22}!n_{..}!}$$