How to estimate the multinomial covariance matrix

Haziq Jamil

2023-08-31

Let $\mathbf{Y} = (Y_1, \dots, Y_p)^\top$ be a multivariate Bernoulli random vector. Suppose we have $i = 1, \dots, n$ observations of $\mathbf{Y}_i$, where each observation was sampled according to some complex sampling design with wight w_i. What is the estimate of the covariance matrix $\mathbf{\Sigma} = \operatorname{Var} \mathbf{Y}$?

My solution: Let $\mathbf{p}$ be the weighted sample proportion estimator

$$\mathbf{p} = \frac{\sum_{i=1}^n w_i \mathbf{y}_i}{\sum_{i=1}^n w_i}.$$

The estimator for the covariance matrix is

$$\hat{\mathbf{\Sigma}} = \frac{1}{\sum_{i=1}^n w_i} \sum_{i=1}^n \frac{n}{n-1} w_i (\mathbf{y}_i - \mathbf{p})(\mathbf{y}_i - \mathbf{p})^\top.$$

Question

- Does the estimator change depending on the sampling design chosen? For e.g.
  - Population consists of 2000 schools of types A (400 units), B (1000 units) and C (600 units). School types correlate with abilities of students, hence provides a way of stratifying the population. Each school organises students into classrooms (clusters).

  A stratified cluster design goes like this: For each stratum, sample 50 schools via SRS. Then within each school, select 1 class by SRS, and all students in that class are added to the sample. The probability of selection of a student in PSU b from school type $a \in \{A, B, C\}$ is

$$\Pr(\text{Selection}) = \frac{50}{\text{\# schools of type } a} \times \frac{1}{\text{\# classes in school } b}.$$