多項分布の一般化について

金沢大・経　星野伸明

多項分布は多くの良い性質を持つが、平均ベクトルと分散共分散行列の構造が固定的で、平均より分散が大きい（過分散）構造を記述できない。従って本講演では、それが可能となるように多項分布を一般化した分布族について、二次モーメントなどを議論する。

母数 $(n, \pi_1, \pi_2, \ldots, \pi_J)$ の多項分布は、母数 λ_j のポアソン分布に独立に従う度数 $F_j, j \in [J]$, の同時分布を度数の総和 $N := \sum_{j=1}^J F_j = n$ で条件付けして得られる。ここで「セル確率」は $\pi_j = \lambda_j / \sum_{j=1}^J \lambda_j$ となる。このような多項分布の構成法において、ポアソン分布を複合ポアソン (Compound Poisson, CP) 分布に置き換えれば、多項分布の良い性質をかなり保持する分布族が生成される。

母数 $(\lambda, \boldsymbol{q} := (q_i)_{i=1}^\infty)$ の CP 分布は確率母関数

$$G(z; \lambda) = \exp(\lambda(g(z) - 1)), \quad \lambda > 0, \tag{1}$$

ただし $g(z) = \sum_{i=1}^\infty q_i z^i$、で定義される。ここで正準条件 $\sum_{i=1}^\infty q_i = 1$, $q_i \geq 0$ を要求するので、確率関数の系列 $\boldsymbol{q}$ は正の整数上の分布を規定する。式 (1) で定義される分布を $\mathtt{CP}(\lambda, \boldsymbol{q})$ で表す。

ここで $\boldsymbol{F}_J := (F_1, F_2, \ldots, F_J) \sim \perp\!\!\!\perp_{j=1}^J \mathtt{CP}(\lambda_j, \boldsymbol{q})$ の時、$\boldsymbol{F}_J$ の $N = n$ 所与の条件付き分布を母数 $(n, \pi_1, \pi_2, \ldots, \pi_J, \sum \lambda_j, \boldsymbol{q})$ の "Conditional CP (CCP)" 分布と呼ぶことにする。セル確率は多項分布の場合と同様に定義される。CCP 分布族は多項＝ディィリクレ混合分布（負の超幾何分布）や疑似多項分布を含む。

CCP 分布の周辺モメントは以下のようになる。

命題 1 *Let $\boldsymbol{F}_J \sim \perp\!\!\!\perp_{j=1}^J \mathtt{CP}(\lambda_j, \boldsymbol{q}), \lambda_j > 0$. Let $Y \sim \boldsymbol{q}$ independently from $\boldsymbol{F}_J$. Write $\mu := \sum_j \lambda_j$. Then for $n \in \mathbb{N}$,*

$$\begin{aligned} E(F_j | N = n) &= n\pi_j, \ j \in [J], \\ V(F_j | N = n) &= n\pi_j(1 - \pi_j) f(n, \mu, \boldsymbol{q}), \ j \in [J], \end{aligned} \tag{2}$$

where

$$\begin{aligned} f(n, \mu, \boldsymbol{q}) &= 1 + \frac{\mu E(Y(Y-1) | N + Y = n) P(N + Y = n)}{n P(N = n)}, \\ Cov(F_i, F_j | N = n) &= -n\pi_i\pi_j f(n, \mu, \boldsymbol{q}), \ i \in [J], j \in [J], i \neq j. \end{aligned}$$

ここで $f(n, \mu, \boldsymbol{q}) \geq 1$ であり、CCP 分布は過分散を記述出来る。またこの不等式の等号は $q_1 = 1$ の場合のみ成立し、それは多項分布のケースである。

命題 1 の系として相関係数行列が即座に導かれるが、それは母数 $(n, \mu, \boldsymbol{q})$ に依存せず、多項分布のそれと一致することが分かる。

系 1 *For all $(n, \mu, \boldsymbol{q})$, $\mathtt{CCP}(n, \pi_1, \pi_2, \ldots, \pi_J, \mu, \boldsymbol{q})$ has the same correlation matrix.*

また命題 1 の式 (2) を見ると、セル母数と他の母数から算出される係数 $f(\mu, \boldsymbol{q}, n)$ を分離出来ることが分かる。疑似多項分布の周辺分布を第二種の疑似二項分布と呼ぶが、その分散についてそのような表現は知られていなかった。今回得た表現は以下の通りである。

系 2 *Let $q_i = (\ell i)^{i-1} \exp(-\ell i)/i!, i \in \mathbb{N}, 0 \leq \ell < 1$. Suppose that $\boldsymbol{F}_J \sim \perp\!\!\!\perp_{j=1}^J \mathtt{CP}(\lambda_j, \boldsymbol{q}), \lambda_j > 0$. Write $\beta := \ell / \sum \lambda_j \geq 0$. Then for $\beta > 0$ and $n \in \mathbb{N}$,*

$$V(F_j | N = n) = n\pi_j(1 - \pi_j) f(n, \beta), \tag{3}$$

where

$$f(n, \beta) = \{n - (n-1)\alpha_{n-2}(n + \beta^{-1}) e^{n + \beta^{-1}} \frac{1}{\beta}\},$$

and

$$\alpha_n(z) = n! z^{-n-1} e^{-z} (1 + z + \frac{z^2}{2!} + \cdots + \frac{z^n}{n!}).$$