Stirling の公式を使った多項分布とカイ 2 乗分布の関係

まず Stirling の公式と後に使う対数のマクローリン展開の式を与えておく。

$$n! \sim \sqrt{2\pi n}\left(n/e\right)^n, \quad \log(1+x) \sim x - x^2/2$$

多項分布の確率分布は以下のように与えられる。ここで記号の詳細な説明は省略する。

$$P(n_1, n_2, \cdots, n_k) = \frac{n!}{n_1! n_2! \cdots n_k!} p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k}$$

この式に上の Stirling の公式を用いると以下のように近似される。

$$\begin{aligned} P(n_1, n_2, \cdots, n_k) &\sim \frac{\sqrt{2\pi n}}{\sqrt{(2\pi)^k n_1 n_2 \cdots n_k}} \frac{n^n}{n_1^{n_1} n_2^{n_2} \cdots n_k^{n_k}} p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k} \\ &= \frac{\sqrt{2\pi n}}{\sqrt{(2\pi)^k n_1 n_2 \cdots n_k}} \left(\frac{np_1}{n_1}\right)^{n_1} \left(\frac{np_2}{n_2}\right)^{n_2} \cdots \left(\frac{np_k}{n_k}\right)^{n_k} \\ &= \frac{\sqrt{2\pi n}}{\sqrt{(2\pi)^k n_1 n_2 \cdots n_k}} \left(\frac{n_1}{np_1}\right)^{-n_1} \left(\frac{n_2}{np_2}\right)^{-n_2} \cdots \left(\frac{n_k}{np_k}\right)^{-n_k} \end{aligned}$$

これをさらに変形して、対数のマクローリン展開を利用すると以下となる。

$$\begin{aligned} P(n_1, n_2, \cdots, n_k) &\sim \frac{\sqrt{2\pi n}}{\sqrt{(2\pi)^k n_1 n_2 \cdots n_k}} \left(1 + \frac{n_1 - np_1}{np_1}\right)^{-n_1} \left(1 + \frac{n_2 - np_2}{np_2}\right)^{-n_2} \cdots \left(1 + \frac{n_k - np_k}{np_k}\right)^{-n_k} \\ &= \frac{\sqrt{2\pi n}}{\sqrt{(2\pi)^k n_1 n_2 \cdots n_k}} \exp\left[-\sum_{i=1}^k n_i \log\left(1 + \frac{n_i - np_i}{np_i}\right)\right] \\ &\sim \frac{\sqrt{2\pi n}}{\sqrt{(2\pi)^k n_1 n_2 \cdots n_k}} \exp\left[-\sum_{i=1}^k n_i \left\{\frac{n_i - np_i}{np_i} - \frac{(n_i - np_i)^2}{2(np_i)^2}\right\}\right] \end{aligned}$$

ここで、

$$\begin{aligned} \sum_{i=1}^k n_i \frac{n_i - np_i}{np_i} &= \sum_{i=1}^k (n_i - np_i + np_i) \frac{n_i - np_i}{np_i} = \sum_{i=1}^k \left[\frac{(n_i - np_i)^2}{np_i} + (n_i - np_i)\right] \\ &= \sum_{i=1}^k \frac{(n_i - np_i)^2}{np_i} \end{aligned}$$

を用いて、$(n_i - np_i)/np_i \ll 1$を仮定すると、

$$\begin{aligned} \sum_{i=1}^k n_i \left\{\frac{n_i - np_i}{np_i} - \frac{(n_i - np_i)^2}{2(np_i)^2}\right\} &= \sum_{i=1}^k \frac{(n_i - np_i)^2}{np_i} \left\{1 - \frac{n_i}{2np_i}\right\} \\ &= \sum_{i=1}^k \frac{(n_i - np_i)^2}{np_i} \left\{\frac{np_i - (n_i - np_i)}{2np_i}\right\} \\ &\sim \sum_{i=1}^k \frac{(n_i - np_i)^2}{2np_i} \equiv \frac{\chi^2}{2} \end{aligned}$$

以上を用いると以下の式を得る。ここに χ^2 は適合度検定のカイ 2 乗統計量である。

$$P(n_1,n_2,\cdots,n_k)\sim\frac{\sqrt{2\pi n}}{\sqrt{(2\pi)^k n_1 n_2\cdots n_k}}\exp\left[-\frac{1}{2}\sum_{i=1}^{k}\frac{(n_i-np_i)^2}{np_i}\right]$$
$$\sim\frac{\sqrt{2\pi n}}{\sqrt{(2\pi)^k np_1 np_2\cdots np_k}}\exp\left[-\frac{1}{2}\sum_{i=1}^{k}\frac{(n_i-np_i)^2}{np_i}\right]$$

次に確率の和について考える。n_i の増分 Δn_i は 1 なので、

$$P\left(\chi^2\le\chi_0^2\right)\sim\sum_{\chi^2\le\chi_0^2}\frac{\sqrt{2\pi n}}{\sqrt{(2\pi)^k np_1 np_2\cdots np_k}}\exp\left[-\frac{1}{2}\sum_{i=1}^{k}\frac{(n_i-np_i)^2}{np_i}\right]$$
$$=\sum_{\chi^2\le\chi_0^2}\frac{\sqrt{2\pi n}}{\sqrt{(2\pi)^k np_1 np_2\cdots np_k}}\exp\left[-\frac{1}{2}\sum_{i=1}^{k}\frac{(n_i-np_i)^2}{np_i}\right]\Delta n_1\cdots\Delta n_{k-1}$$

これを以下の積分に置き換える。

$$P\left(\chi^2\le\chi_0^2\right)\sim\frac{\sqrt{2\pi n}}{\sqrt{(2\pi)^k np_1 np_2\cdots np_k}}\int_{\chi^2\le\chi_0^2}\exp\left[-\frac{1}{2}\sum_{i=1}^{k}\frac{(n_i-np_i)^2}{np_i}\right]dn_1\cdots dn_{k-1}$$

変数変換で、

$$\sum_{i=1}^{k}\frac{(n_i-np_i)^2}{np_i}=\sum_{i=1}^{k-1}\frac{(n_i-np_i)^2}{np_i}+\frac{(n_k-np_k)^2}{np_k}=\sum_{i=1}^{k-1}z_i^2$$

とするには、

$$\sum_{i=1}^{k-1}\frac{(n_i-np_i)^2}{np_i}+\frac{(n_k-np_k)^2}{np_k}=\sum_{i=1}^{k-1}\frac{(n_i-np_i)^2}{np_i}+\frac{1}{np_k}\left(\sum_{i=1}^{k-1}(n_i-np_i)\right)^2$$
$$=\sum_{i=1}^{k-1}\sum_{j=1}^{k-1}\left(\frac{n_i-np_i}{\sqrt{np_i}}\frac{n_j-np_j}{\sqrt{np_j}}\delta_{ij}+\frac{n_i-np_i}{\sqrt{np_k}}\frac{n_j-np_j}{\sqrt{np_k}}\right)$$
$$=\sum_{i=1}^{k-1}\sum_{j=1}^{k-1}(n_i-np_i)\left(\frac{1}{\sqrt{np_i}}\frac{1}{\sqrt{np_j}}\delta_{ij}+\frac{1}{np_k}\right)(n_j-np_j)$$

より、以下の行列

$$a_{ij}=\frac{1}{\sqrt{np_i}}\frac{1}{\sqrt{np_j}}\delta_{ij}+\frac{1}{np_k}$$

の固有値 λ_i を使って、以下のような変数変換を行えばよい。

$$z_i=\sqrt{\lambda_i}(n_i-np_i)$$

即ち、

$$\sum_{i=1}^{k-1}\sum_{j=1}^{k-1}(n_i-np_i)\left(\frac{1}{\sqrt{np_i}}\frac{1}{\sqrt{np_j}}\delta_{ij}+\frac{1}{np_k}\right)(n_j-np_j)=\sum_{i=1}^{k-1}\left(\sqrt{\lambda_i}(n_i-np_i)\right)^2$$

その際の変数変換のヤコビアンは、

$$\frac{\partial(n_1,\cdots,n_{k-1})}{\partial(z_1,\cdots,z_{k-1})}=\left[\frac{\partial(z_1,\cdots,z_{k-1})}{\partial(n_1,\cdots,n_{k-1})}\right]^{-1}=1\Big/\sqrt{\lambda_1\cdots\lambda_{k-1}}$$

であるが、この値を求めてみよう。簡単のために、例として 3 行 3 列では、

$$\sqrt{\lambda_1\cdots\lambda_{k-1}}=\frac{\partial(z_1,z_2,z_3)}{\partial(n_1,n_2,n_3)}=\begin{vmatrix}1/np_1+1/np_4 & 1/np_4 & 1/np_4\\ 1/np_4 & 1/np_2+1/np_4 & 1/np_4\\ 1/np_4 & 1/np_4 & 1/np_3+1/np_4\end{vmatrix}$$

$$=\begin{vmatrix}1/np_1+1/np_4 & 1/np_4 & 1/np_4\\ -1/np_1 & 1/np_2 & 0\\ -1/np_1 & 0 & 1/np_3+1/np_4\end{vmatrix}=\begin{vmatrix}1/np_1+1/np_4 & 1/np_4 & 1/np_4\\ -1/np_1 & 1/np_2 & 0\\ -1/np_1 & 0 & 1/np_3\end{vmatrix}$$

$$=\left(\frac{1}{np_1}+\frac{1}{np_4}\right)\frac{1}{np_2}\frac{1}{np_3}+\frac{1}{np_4}\left(\frac{1}{np_1}\frac{1}{np_2}+\frac{1}{np_3}\frac{1}{np_1}\right)$$

$$=\frac{1}{np_1np_2np_3}\left(1+\frac{p_1+p_2+p_3}{p_k}\right)=\frac{n}{np_1np_2np_3np_4}$$

固有値の積が、上のように与えられることから、

$$dn_1dn_2dn_3=\frac{1}{\sqrt{\lambda_1\lambda_2\lambda_3}}dz_1dz_2dz_3=\sqrt{\frac{np_1np_2np_3np_4}{n}}dz_1dz_2dz_3$$

一般的には以下になる。

$$dn_1\cdots dn_{k-1}=\sqrt{\frac{np_1\cdots np_{k-1}np_k}{n}}dz_1\cdots dz_{k-1}$$

この関係を用いると、

$$P\left(\chi^2\le\chi_0^2\right)\sim\frac{\sqrt{2\pi n}}{\sqrt{(2\pi)^k np_1np_2\cdots np_k}}\int_{\chi^2\le\chi_0^2}\exp\left[-\frac{1}{2}\sum_{i=1}^{k}\frac{(n_i-np_i)^2}{np_i}\right]dn_1\cdots dn_{k-1}$$

$$=\frac{\sqrt{2\pi n}}{\sqrt{(2\pi)^k np_1np_2\cdots np_k}}\int_{\chi^2\le\chi_0^2}\exp\left[-\frac{1}{2}\sum_{i=1}^{k-1}z_i^2\right]\sqrt{\frac{np_1\cdots np_{k-1}np_k}{n}}dz_1\cdots dz_{k-1}$$

$$\simeq\frac{1}{\sqrt{(2\pi)^{k-1}}}\int_{\chi^2\le\chi_0^2}\exp\left[-\frac{1}{2}\sum_{i=1}^{k-1}z_i^2\right]dz_1\cdots dz_{k-1}$$

$$=\frac{1}{\sqrt{(2\pi)^{k-1}}}\int_{\chi^2\le\chi_0^2}\exp\left[-\frac{1}{2}\chi^2\right]dz_1\cdots dz_{k-1}$$

被積分関数が $\chi^2=\sum_{i=1}^{k-1}z_i^2$ と同型方向の座標の組だけで表されることから、積分を回転方向と同型方向に分けて考える。そのため、$k-1$次元球の体積の問題を考える。

$$V_{k-1}(\chi^2)=\frac{\pi^{(k-1)/2}}{\Gamma((k-1)/2+1)}\chi^{k-1}=\frac{\pi^{(k-1)/2}}{((k-1)/2)\Gamma((k-1)/2)}(\chi^2)^{(k-1)/2}\qquad\text{より、}$$

表面積は、変数を χ^2 として、

$$S_{k-1}(\chi^2)=\frac{d}{d\chi^2}V_{k-1}(\chi^2)=\frac{\pi^{(k-1)/2}}{\Gamma((k-1)/2)}(\chi^2)^{(k-1)/2-1}$$

以上を用いると、

$$P\left(\chi^2 \le \chi_0^2\right) = \frac{1}{\sqrt{(2\pi)^{k-1}}} \int_{\chi^2 \le \chi_0^2} \exp\left[-\frac{1}{2}\chi^2\right] \frac{\pi^{(k-1)/2}}{\Gamma((k-1)/2)} (\chi^2)^{(k-1)/2-1} d\chi^2$$

$$= \frac{1}{2^{(k-1)/2}\Gamma\left((k-1)/2\right)} \int_{\chi^2 \le \chi_0^2} (\chi^2)^{(k-1)/2-1} e^{-\chi^2/2} d\chi^2$$

ここで自由度 k のカイ 2 乗分布の密度関数は以下で与えられので、

$$f(\chi^2;k)d\chi^2 = \frac{1}{2^{k/2}\Gamma(k/2)} (\chi^2)^{k/2-1} e^{-\chi^2/2} d\chi^2$$

上の分布は自由度 $k-1$ のカイ 2 乗分布となる。

イェーツ補正についての考え方

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$

についての積分領域を以下のように変える。

$n_i - np_i \ge 0$ のときは、 $n_i \to n_i - 1/2$

$n_i - np_i < 0$ のときは、 $n_i \to n_i + 1/2$

つまり、

$$P\left(\chi^2 > \chi_0^2\right) = \int_{\chi^2 > \chi_0^2 - \delta} P(n_1, n_2, \cdots, n_k) dn_1 \cdots dn_{k-1}$$

ここに δ は積分領域が変化した差を表すものとする。

この積分領域の変換についての変数変換は、

$n_i - np_i \ge 0$ のときは、 $n_i' = n_i + 1/2$

$n_i - np_i < 0$ のときは、 $n_i' = n_i - 1/2$

として、元の積分領域を取ればよい。ここで

$$n_k = n - \sum_{i=1}^{k-1} n_i = n - \sum_{n_i \ge np_i} (n_i' - 1/2) - \sum_{n_i < np_i} (n_i' + 1/2)$$

$$= n - \sum_{i=1}^{k-1} n_i' + \sum_{n_i \ge np_i} 1/2 - \sum_{n_i < np_i} 1/2 = n_k' + \sum_{n_i \ge np_i} 1/2 - \sum_{n_i < np_i} 1/2$$

であるが、不確定のため、他の n_i と同様に扱う。

最終的に

$$\chi'^2 = \sum_{i=1}^{k} \frac{(|n_i' - np_i| - 1/2)^2}{np_i}$$

として、

$$P\left(\chi^2 > \chi_0^2\right) \sim \frac{\sqrt{2\pi n}}{\sqrt{(2\pi)^k np_1 np_2 \cdots np_k}} \int_{\chi'^2 > \chi_0^2} \exp\left[-\frac{1}{2}\sum_{i=1}^{k} \frac{(|n_i' - np_i| - 1/2)^2}{np_i}\right] dn_1' \cdots dn_{k-1}'$$

以後の計算は前と同様に行える。