新たな離散異分布適合度検定統計量と その海洋調査データへの適用

柴田里程（慶應義塾大学理工学部)

1　離散分布の適合度検定統計量

与えられたデータに対する確率モデルの有効性を主張するには最終的には適合度検定に頼るしかない．しかし最近では，このような古典的な検定を省略し，頭のなかだけで考えたいくつかのモデルを AIC などのモデル選択基準で比較しその中での最良なモデルを最終結論とすればよい，という安易な風潮も見受けられる．モデル選択はあくまでも，どのモデルも適合度に問題はない場合に，すこしでも簡素で理解しやすいモデルを選択したいという欲求から生まれた方法である．つまり，相対的な比較でしかなく，それ自体で有効性を主張する力はもっていない．

すでに適合度検定統計量については十分研究し尽くされているようにも思えるが，それは連続分布の場合であって，離散分布の場合はピアソン χ^2 がほとんど唯一の統計量といってよく，特に，独立であっても同分布でない場合の研究ははほとんど手がつけられていない現状にある．そこで，Victoria University of Wellington の E.Khmaladze との共同研究の過程で発見した，ピアソン χ^2 の一般化を紹介するとともに，これを応用することで独立異分布の場合にも使える適合度検定統計量が容易に導かれることを示すとともに，その有効性を検証した結果を報告した．

2　離散異分布

異分布であることが避けられない一つの実例として，報告者がここ数年オーストラリアの CSIRO との共同研究として取り組んできた NPF(Northern Prawn Fishery) 海洋調査データをとりあげた．この調査では，エビのトロール漁の影響を調べるため，実験用に設定した領域でトロール漁を行う前と後の数回浚渫を行い，捕捉された海洋生物の種ごとにその個体数と重量を計測している．ここでの問題は，浚渫が船からかごを降ろしそれを引っ張りながら海底を底ざらいする形であるため，かごがいっぱいになりそうになるとそこで引き上げてしまい浚渫面積が一定しないところにある．つまり，n 回の浚渫の結果得られたある種の個体数を $N_1, N_2, ..., N_n$，対応する浚渫面積を $\alpha_1, \alpha_2, ..., \alpha_n$ とすればもっとも簡単なポアッソン分布モデルでも $N_i \sim Po(\alpha_i\lambda), i = 1, 2, ..., n$ のように異分布となり，種によっては集落を形成することが多いことも考慮したトーマス分布なら $Tho(\alpha_i\lambda, \phi), i = 1, 2, ..., n$ のような異分布を考える必要がある．このような異分布性は，連続分布なら適当な変換を導入することにより解消できることも多いが，離散分布，特に個数の分布に対してはこのような変換はかえって分布を複雑にするだけでほとんどメリットがないことが多い．

3　離散異分布の扱い

離散異分布を正面から扱う困難を避けるため，よく行われる便宜的な扱いとしては

1. N_i/α_i のように標準化し正規分布 $N(\lambda, \lambda/\alpha_i)$ あるいは $N(\lambda, \sigma^2)$ で近似する．
2. 0 カウントが多いことを反映するため負の二項分布 $NB_N(\alpha_i, p)$ を用いる．
3. $\log E(N_i) = \log\alpha_i + \log\theta$ のような GLIM(Generalized Linear Model) を用いる．

などがあるが，このような解析は現象の本質に迫るサイエンスというよりは単なる技法の適用であり，判明するのはごく表面的な事実だけである．下手をすれば虚像を見ているだけかもしれない．このような場合に解析者はサイエンティストとしてどう責任をとるつもりなのであろうか．

4　ピアソン χ^2 の一般化

まず，古典的な独立同分布な観測 $X_1, X_2, ..., X_n$ の場合を考えてみる．$p_i = P(X_k = x_i), i = 1, 2, ..., m$ としたとき，ピアソンの χ^2 はいうまでもなく $\chi^2 = ||\boldsymbol{Y}_n||^2 \simeq \chi^2(m-1)$ で与えられる．ここで $\simeq$ は漸近分布の意味で用いており，

$$Y_{in} = \frac{\#\{X_k = x_i,\ k = 1, 2, ..., n\} - np_i}{\sqrt{np_i}},\ i = 1, 2, ..., m$$

漸近 χ^2 分布は $\boldsymbol{Y}_n = (Y_{1n}, Y_{2n}, ..., Y_{mn})^T \simeq N\left(\boldsymbol{0},\ I - \sqrt{\boldsymbol{p}}\sqrt{\boldsymbol{p}}^T\right)$ であることにもとづいている．ただし $\sqrt{\boldsymbol{p}}^T = (\sqrt{p_1}, \sqrt{p_2}, ..., \sqrt{p_m})$.

ピアソンの χ^2 は余りにも有名であるのであまり注意も払われないが，$\boldsymbol{Y}_n$ の漸近分布が $\boldsymbol{p}$ に依存するにも関わらず，その２乗ノルムが $\boldsymbol{p}$ に依存しない一定の χ^2 分布に従うのは，漸近分散共分散が射影行列になっているという極めてラッキーな状況をうまく利用しているからである．したがって，$\boldsymbol{Y}_n$ の一部の要素にもとづいた統計量などはもはやこのようなきれいな性質は持たない．しかし，すこし発想を変えれば漸近分散共分散を射影行列であることは保ちながら自由に変化させることができる．

$$\boldsymbol{Z}_n = \boldsymbol{Y}_n - \langle \boldsymbol{Y}_n, \boldsymbol{r} \rangle \frac{\boldsymbol{r} + \sqrt{\boldsymbol{p}}}{1 + \left\langle \sqrt{\boldsymbol{p}}, \boldsymbol{r} \right\rangle} \ \simeq\ N(0,\ I - \boldsymbol{r}\boldsymbol{r}^T).$$

あきらかに，常に $||\boldsymbol{Z}_n||^2 \simeq \chi^2(m-1)$ であり $\boldsymbol{r} = \sqrt{\boldsymbol{p}}$ にとれば $\boldsymbol{Z}_n = \boldsymbol{Y}_n$ で，ピアソンの χ^2 に戻る．

5　異分布の場合への応用

前節の結果は，パラメータ $\boldsymbol{p}$ によらない漸近分布をもつように変換できるだけでなく，$\boldsymbol{r}$ を自由に選択できることから，検定の幅を広げるのにも役立つと思われるが，異分布の場合にはこのアイディアが本質的な役割を果たす．$p_{ki} = P(X_k = x_i), i = 1, 2, ..., m$ とし，

$$\eta_{ki} = \frac{I_{(X_k = x_i)} - p_{ki}}{\sqrt{p_{ki}}},\ i = 1, 2, ..., m,\ k = 1, 2, ..., n$$

を標本ごとに定義すれば

$$\boldsymbol{Z}_k = \boldsymbol{\eta}_k - \langle \boldsymbol{\eta}_k, \boldsymbol{r}_k \rangle \frac{\boldsymbol{r}_k + \sqrt{\boldsymbol{p}_k}}{1 + \left\langle \sqrt{\boldsymbol{p}}_k, \boldsymbol{r}_k \right\rangle},\ k = 1, 2, ..., n$$

の算術平均の漸近分散共分散は $\boldsymbol{r}_k$ を選ぶことでかなり自由に定められ，

$$\boldsymbol{W}_n = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \boldsymbol{Z}_k \ \simeq\ N\left(\boldsymbol{0},\ I - \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \boldsymbol{r}_k \boldsymbol{r}_k^T\right)$$

が成立する．特に $\boldsymbol{r}_k = \boldsymbol{r}$ のように同じベクトルにとれば，独立同分布の場合とまったく同じように

$$\boldsymbol{W}_n = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \boldsymbol{Z}_k \ \simeq\ N\left(\boldsymbol{0},\ I - \boldsymbol{r}\boldsymbol{r}^T\right).$$

となり，$\|\boldsymbol{W}_n\|^2$ は自由度 $m-1$ の漸近 χ^2 分布を持つ．パラメータ推定を含む場合も，MDE(Minimum Distance Estimate) を用いる限り同じような変換で射影行列となる漸近分散共分散を自由に選ぶことができることも報告したが，ここではその詳細を省略する．