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)

2012-10-25 於 松江テルサ

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すればよいという安易な風潮も見受けられるが，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有していない．

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．

本報告では，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，異分布であることが避けられない海洋調査データ（カウントデータ）に対してポアッソン分布あるいはトーマス分布が果たして有効な確率モデルかどうか検証した結果を報告する．

2　ピアソン χ^2 再訪

まず，古典的な場合である，独立同分布な観測 $X_1, X_2, ..., X_n$ の場合から始める．$p_i = P(X_k = x_i), i = 1, 2, ..., m$ 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．以降 $\simeq$ は漸近分布の意味で用いる．ただし，

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

であり，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，あまり注意が払われることが少ないが，$\boldsymbol{Y}_n$ の漸近分布は $\boldsymbol{p}$ に依存するにも関わらず，その２乗ノルムは $\boldsymbol{p}$ に依存しない一定の χ^2 分布に従うのは，漸近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．

2.1 ピアソン χ^2 の一般化

そこで，$\boldsymbol{Y}_n$ に線形変換を施して $\boldsymbol{p}$ によらない漸近分散行列 Σ_m の正規分布に収束させることを考える．しかし $\langle \boldsymbol{Y}_n,\ \sqrt{\boldsymbol{p}} \rangle = 0$ から $\mathrm{Rank}(\Sigma_m) \leq m-1$．したがって $\mathrm{Ker}(\Sigma_m)$ を $||\boldsymbol{r}|| = 1$ で張る一次元部分空間とすれば直交射影 $\Sigma_m = I - \boldsymbol{r}\boldsymbol{r}^T$ が Σ_m の一つの標準形となる．このような漸近共分散を導く線形変換は次の Proposition 1 で与えられる．

Proposition 1

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

あきらかに, $\boldsymbol{r} = \sqrt{\boldsymbol{p}}$ にとれば $\boldsymbol{Z}_n = \boldsymbol{Y}_n$ で $||\boldsymbol{Z}_n||^2$ はピアソンの χ^2 に戻る．

この変換は次の Lemma 1 から得られる．ただし，以下では $\boldsymbol{r}$ の $\boldsymbol{q}$ に直交する成分を $\boldsymbol{r}_{\perp}\boldsymbol{q} = \boldsymbol{r} - < \boldsymbol{q}, \boldsymbol{r} > \boldsymbol{q}$ のような記号を用いて表すことにする．

Lemma 1

$$U = \boldsymbol{r}\boldsymbol{q}^T \ \pm \ \frac{1}{\mu^2} \boldsymbol{q}_{\perp \boldsymbol{r}}\, \boldsymbol{r}_{\perp \boldsymbol{q}}^T$$

はユニタリーで $U\boldsymbol{q} = \boldsymbol{r}$. ただし, $||\boldsymbol{r}|| = ||\boldsymbol{q}|| = 1,\ \mu = ||\boldsymbol{q}_{\perp \boldsymbol{r}}|| = ||\boldsymbol{r}_{\perp \boldsymbol{q}}||$.

実際，部分空間 $L = \mathrm{Span}\{\boldsymbol{q}, \boldsymbol{r}\}$ に対して $V = \mathrm{Proj}_{L^{\perp}} + U$ を定義すれば，$V\boldsymbol{q} = \boldsymbol{r}$, $VV^T = \mathrm{Proj}_{L^{\perp}} + UU^T = I$ であるので，$V(I - \boldsymbol{q}\boldsymbol{q}^T)V^T = I - \boldsymbol{r}\boldsymbol{r}^T$. また，$\boldsymbol{y} \perp \boldsymbol{q}$ に対しては,

$$\begin{aligned}
\mathrm{Proj}_{L^{\perp}} \boldsymbol{y} &= \boldsymbol{y} - \frac{1}{\mu^2} < \boldsymbol{y}, \boldsymbol{r}_{\perp \boldsymbol{q}} > \boldsymbol{r}_{\perp \boldsymbol{q}}, \\
U\boldsymbol{y} &= (\boldsymbol{r}\boldsymbol{q}^T - \frac{1}{\mu^2} \boldsymbol{q}_{\perp \boldsymbol{r}}\, \boldsymbol{r}_{\perp \boldsymbol{q}}^T)\boldsymbol{y} \\
&= -\frac{1}{\mu^2} \boldsymbol{q}_{\perp \boldsymbol{r}} < \boldsymbol{r}_{\perp \boldsymbol{q}}, \boldsymbol{y} > \\
&= -\frac{(\boldsymbol{r} + \boldsymbol{q})(1 - < \boldsymbol{q}, \boldsymbol{r} >) - \boldsymbol{r}_{\perp \boldsymbol{q}}}{\mu^2} < \boldsymbol{r}_{\perp \boldsymbol{q}}, \boldsymbol{y} > \\
&= -\frac{\boldsymbol{r} + \boldsymbol{q}}{1 + < \boldsymbol{q}, \boldsymbol{r} >} < \boldsymbol{r}, \boldsymbol{y} > + \frac{\boldsymbol{r}_{\perp \boldsymbol{q}}}{\mu^2} < \boldsymbol{r}_{\perp \boldsymbol{q}}, \boldsymbol{y} >
\end{aligned}$$

が成り立つので,

$$V\boldsymbol{y} = \boldsymbol{y} - \frac{\boldsymbol{r} + \boldsymbol{q}}{1 + < \boldsymbol{q}, \boldsymbol{r} >} < \boldsymbol{r}, \boldsymbol{y} >$$

となり．$\boldsymbol{q} = \sqrt{\boldsymbol{p}}$, $\boldsymbol{y} = \boldsymbol{Y}_n$ にとれば，Proposition1 が従う．

2.2 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，

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

を定義すれば $\boldsymbol{\eta}_k = (\eta_{k1}, \eta_{k2}, ..., \eta_{km})^T$ は独立であり，

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

が成立する．ただし $\sqrt{\boldsymbol{p}_k}^T = (\sqrt{p_{k1}}, \sqrt{p_{k2}}, ..., \sqrt{p_{km}})$ である．ただし，これは右辺の漸近分散が存在すればの話であり，一般的には存在するとは限らず，また射影行列になっていることは必ずしも期待できない．ここに，異分布の場合の適合度検定統計量の導出の困難さがあった．しかし，同分布の場合の一般化を応用すれば

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

の平均の分散共分散行列は $\boldsymbol{r}_k$ 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漸近分布をもつようにできる．

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

したがって，$\|\boldsymbol{W}_n\|^2$ は常に自由度 $m-1$ の漸近 χ^2 分布を持つことになる．図 1 , 2 は乱数実験の結果を示したもので，異分布の場合でも $n = 100$ ぐらいで十分漸近分布に収束していることが見てとれる．図 2 は図 1 に示したような強度 $\lambda_i, i = 1, 2, ..., 100$ で定まるポアッソン分布 $Po(\lambda_i), i = 1, 2, ..., n$ に従う乱数を 50,000 回生成したときの $\|\boldsymbol{W}_n\|^2$ の値の QQ プロットである．ただし，$p_{k6} = P(X_k \geq 5)$ のように裾の確率はまとめ, $m = 6$ に設定している．

3 パラメータ推定を含む場合

3.1 独立同分布標本のとき

確率が一次元パラメータ θ によって既知の関数によって定まる $p_i = p_i(\theta), i = 1, 2, ..., m$ のときを考える．パラメータ既知のときと同じように

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

を定義し，パラメータ推定としては，MDE(Minimum Distance Estimate),

$$\hat{\theta} : ||\boldsymbol{Y}_n(\theta)||^2 \to \min$$

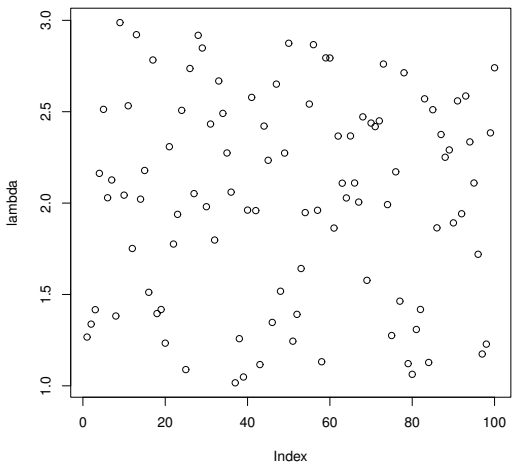

図 1: $\lambda_i, i = 1, 2, ..., 100$

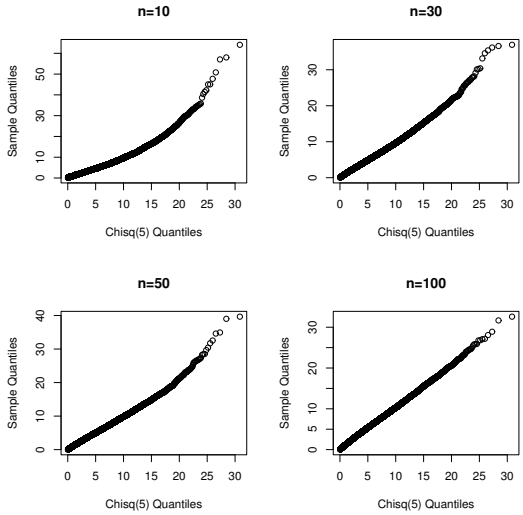

図 2: 自由度 5 の χ^2 に対する QQ プロット

を用いる．ただし $\boldsymbol{Y}_n(\theta) = (Y_{1n}, Y_{2n}, ..., Y_{mn})^T$ である．MDE に対しては漸近的に

$$\boldsymbol{Y}_n(\hat{\theta}) \approx \boldsymbol{Y}_n(\theta) - \langle \boldsymbol{Y}_n(\theta),\ \boldsymbol{q}_2 \rangle\, \boldsymbol{q}_2$$

が成立する．ただし，

$$\boldsymbol{q}_1 = \sqrt{\boldsymbol{p}(\theta)},\ \ \boldsymbol{q}_2 = \frac{1}{g}\,\frac{\boldsymbol{p}'(\theta)}{\sqrt{\boldsymbol{p}(\theta)}},\ \ g = \left\| \frac{\boldsymbol{p}'(\theta)}{\sqrt{\boldsymbol{p}(\theta)}} \right\|$$

である．ここで，$\boldsymbol{q}_1$ と $\boldsymbol{q}_2$ が直交するだけでなく，上式の右辺もこのいずれとも直交するので，

$$\hat{\boldsymbol{Y}}_n = \boldsymbol{Y}_n(\hat{\theta}) \simeq N\left(\boldsymbol{0},\ I - \boldsymbol{q}_1\boldsymbol{q}_1^T - \boldsymbol{q}_2\boldsymbol{q}_2^T\right),$$

となる．パラメータ既知の場合と同じように

$$\mathrm{Ker}(\Sigma_m) = \mathrm{Span}\{\boldsymbol{r}_1, \boldsymbol{r}_2\},\ \ \boldsymbol{r}_1 \perp \boldsymbol{r}_2,\ \ ||\boldsymbol{r}_1|| = ||\boldsymbol{r}_2|| = 1$$

を選べば，

$$\boldsymbol{Z}_n = \hat{\boldsymbol{Y}}_n - \left\langle \hat{\boldsymbol{Y}}_n, \boldsymbol{\alpha} \right\rangle (\boldsymbol{\alpha} - \boldsymbol{\gamma}) - \left\langle \hat{\boldsymbol{Y}}_n, \boldsymbol{\beta} \right\rangle (\boldsymbol{\beta} - \boldsymbol{\delta}) \simeq N(\boldsymbol{0},\ I - \boldsymbol{r}_1\boldsymbol{r}_1^T - \boldsymbol{r}_2\boldsymbol{r}_2^T)$$

のように，漸近分散が $\boldsymbol{r}_1, \boldsymbol{r}_2$ だけで定まる射影行列となる．ただし

$$\boldsymbol{\alpha} = \frac{(\boldsymbol{r}_1)_\perp \boldsymbol{q}_1\boldsymbol{q}_2}{\left\|(\boldsymbol{r}_1)_\perp \boldsymbol{q}_1\boldsymbol{q}_2\right\|},\ \boldsymbol{\beta} = \frac{(\boldsymbol{r}_2)_\perp \boldsymbol{r}_1\boldsymbol{q}_1\boldsymbol{q}_2}{\left\|(\boldsymbol{r}_2)_\perp \boldsymbol{r}_1\boldsymbol{q}_1\boldsymbol{q}_2\right\|},\ \boldsymbol{\gamma} = \frac{(\boldsymbol{q}_1)_\perp \boldsymbol{r}_1\boldsymbol{r}_2}{\left\|(\boldsymbol{q}_1)_\perp \boldsymbol{r}_1\boldsymbol{r}_2\right\|},\ \boldsymbol{\delta} = \frac{(\boldsymbol{q}_2)_\perp \boldsymbol{q}_1\boldsymbol{r}_1\boldsymbol{r}_2}{\left\|(\boldsymbol{q}_2)_\perp \boldsymbol{q}_1\boldsymbol{r}_1\boldsymbol{r}_2\right\|}$$

であるが，これらは未知パラメータ θ に依存するので，実際には $\hat{\theta}$ をプラグインする．このことによって漸近分布は変化しないことは明らかである．

3.2 独立異分布標本のとき

$p_{ki} = p_{ki}(\theta)$ のとき，

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

から $\boldsymbol{\eta}_k(\theta) = (\eta_{k1}(\theta), ..., \eta_{km}(\theta))^T$ をつくれば，MDE 推定は

$$\hat{\theta}: \ \| \sum_{k=1}^{n} \boldsymbol{\eta}_k(\theta) \|^2 \to \min$$

で与えられ

$$\frac{1}{\sqrt{n}} \sum_{k=1}^{n} \boldsymbol{\eta}_k(\hat{\theta}) \approx \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \boldsymbol{\eta}_k(\theta) - \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \langle \boldsymbol{\eta}_k(\theta), \ \boldsymbol{q}_{2k} \rangle \, \boldsymbol{q}_{2k}$$

が成り立つ. ただし

$$\boldsymbol{q}_{1k} = \sqrt{\boldsymbol{p}_k(\theta)}, \ \ \boldsymbol{q}_{2k} = \frac{1}{g_k} \frac{\boldsymbol{p}_k'(\theta)}{\sqrt{\boldsymbol{p}_k(\theta)}}, \ \ g_k = \left\| \frac{\boldsymbol{p}_k'(\theta)}{\sqrt{\boldsymbol{p}_k(\theta)}} \right\|$$

である．したがって，同分布の場合と同じような理由で

$$\frac{1}{\sqrt{n}} \sum_{k=1}^{n} \boldsymbol{\eta}_k(\hat{\theta}) \simeq N\left(\mathbf{0}, \ I - \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} (\boldsymbol{q}_{1k} \boldsymbol{q}_{1k}^T + \boldsymbol{q}_{2k} \boldsymbol{q}_{2k}^T) \right)$$

となるが，漸近分散はかなり厄介である． しかし

$$\boldsymbol{Z}_k = \boldsymbol{\eta}_k(\hat{\theta}) - \left\langle \boldsymbol{\eta}_k(\hat{\theta}), \boldsymbol{\alpha}_k \right\rangle (\boldsymbol{\alpha}_k - \boldsymbol{\gamma}_k) - \left\langle \boldsymbol{\eta}_k(\hat{\theta}), \boldsymbol{\beta}_k \right\rangle (\boldsymbol{\beta}_k - \boldsymbol{\delta}_k)$$

のような変換を施すことで

$$\boldsymbol{W}_n = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \boldsymbol{Z}_k \ \simeq \ N\left(\mathbf{0}, \ I - \ \boldsymbol{r}_1 \boldsymbol{r}_1^T - \boldsymbol{r}_2 \boldsymbol{r}_2^T \right)$$

のように，一定の共分散をもつ正規分布に収束する．ただし，

$$\boldsymbol{\alpha}_k = \frac{(\boldsymbol{r}_1)_{\perp \boldsymbol{q}_{1k} \boldsymbol{q}_{2k}}}{\left\| (\boldsymbol{r}_1)_{\perp \boldsymbol{q}_{1k} \boldsymbol{q}_{2k}} \right\|}, \ \boldsymbol{\beta}_k = \frac{(\boldsymbol{r}_2)_{\perp \boldsymbol{r}_1 \boldsymbol{q}_{1k} \boldsymbol{q}_{2k}}}{\left\| (\boldsymbol{r}_2)_{\perp \boldsymbol{r}_1 \boldsymbol{q}_{1k} \boldsymbol{q}_{2k}} \right\|}, \ \boldsymbol{\gamma}_k = \frac{(\boldsymbol{q}_{1k})_{\perp \boldsymbol{r}_1 \boldsymbol{r}_2}}{\left\| (\boldsymbol{q}_{1k})_{\perp \boldsymbol{r}_1 \boldsymbol{r}_2} \right\|}, \ \boldsymbol{\delta}_k = \frac{(\boldsymbol{q}_{2k})_{\perp \boldsymbol{q}_{1k} \boldsymbol{r}_1 \boldsymbol{r}_2}}{\left\| (\boldsymbol{q}_{2k})_{\perp \boldsymbol{q}_{1k} \boldsymbol{r}_1 \boldsymbol{r}_2} \right\|}.$$

$\boldsymbol{r}_1, \boldsymbol{r}_2$ としては直交するベクトルなら何をとってもよいが，たとえば，m が偶数の場合

$$\boldsymbol{r}_1 = \left(\frac{1}{\sqrt{m}}, \frac{1}{\sqrt{m}}, ..., \frac{1}{\sqrt{m}} \right)^T, \ \boldsymbol{r}_2 = \left(\frac{1}{\sqrt{m}}, \frac{-1}{\sqrt{m}}, \frac{1}{\sqrt{m}}, \frac{-1}{\sqrt{m}}, ..., \frac{1}{\sqrt{m}} \right)^T$$

のように設定するのが一例である.