『入門 機械学習による異常検知 — R による実践ガイド —』（コロナ社、2015）の補足

Tsuyoshi Idé (井手 剛) ide@ide-research.net

平成 28 年 2 月 28 日

1　F 分布とカイ 2 乗分布の関係

第 2 章「正規分布に従うデータからの異常検知」において、F 分布の極限でカイ 2 乗分布が得られることを証明なしに言及しました（p.48）。表記もあいまいだったので、問題の形で導出を書きたいと思います。この点をご指摘頂きました高木英明 筑波大学名誉教授に感謝いたします。

1.1　スターリング公式による F 分布の直接近似

問題: 自由度 (m,n) の F 分布に従う確率変数 z があるとします。すなわち、z の分布が次のように与えられるとします。

$$\mathcal{F}(z|m,n) \equiv \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \left(\frac{m}{n}\right)^{\frac{m}{2}} z^{\frac{m}{2}-1} \left(1+\frac{mz}{n}\right)^{-\frac{m+n}{2}} \tag{1}$$

m を有限に保ったまま $n \to \infty$ とするとき、確率変数 mz が、自由度 m、スケール因子 1 のカイ 2 乗分布に従うことを次の順序で示してください。

1. ガンマ関数に関するスターリング近似（Stirling's approximation）

$$\ln\Gamma(c) = c\ln c - c - \frac{1}{2}\ln\left(\frac{c}{2\pi}\right) + O\left(\frac{1}{c}\right) \tag{2}$$

を使うことで（$c > 0$）

$$\ln\frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} = \frac{m}{2}\ln\left(\frac{n}{2}\right) + O\left(\frac{m}{n}\right) \tag{3}$$

を証明してください。ただし、$O\left(\frac{m}{n}\right)$ は、「高々 $\frac{m}{n}$ の微小量」という意味です。

2. 同じく $O\left(\frac{m}{n}\right)$ の誤差は別にして成り立つ式

$$\ln\left\{\left(1+\frac{m}{n}z\right)^{-\frac{m+n}{2}}\right\} = -\frac{mz}{2} + O\left(\frac{m}{n}\right) \tag{4}$$

を示してください。

3. m を有限に保ったまま $n \to \infty$ とするとき、$\mathcal{F}(z|m,n)$ が

$$\frac{m}{2\Gamma\left(\frac{m}{2}\right)}\left(\frac{mz}{2}\right)^{\frac{m}{2}-1}\exp\left(-\frac{mz}{2}\right) \tag{5}$$

と近似できることを示してください。

4. 上記の分布に従う z に対し、$u \equiv mz$ で定義される新しい確率変数 u を考えます。このとき、u の分布が $\chi^2(u \mid m, 1)$ であることを示してください。

解答: 1 について。単にスターリングの式を代入すると

$$\begin{aligned}
\ln\frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} &= \frac{m+n}{2}\ln\left(\frac{m+n}{2}\right) - \frac{m}{2} - \frac{1}{2}\ln\left(\frac{m+n}{n}\right) - \frac{n}{2}\ln\left(\frac{n}{2}\right) \\
&= \frac{m}{2}\ln\left(\frac{m+n}{2}\right) - \frac{m}{2} - \frac{1}{2}\ln\left(\frac{m+n}{n}\right) + \frac{n}{2}\ln\left(\frac{m+n}{n}\right) \\
&= \frac{m}{2}\ln\left\{\frac{n}{2}(1+\epsilon)\right\} - \frac{m}{2} - \frac{1}{2}\ln(1+\epsilon) + \frac{n}{2}\ln(1+\epsilon) \\
&= \frac{m}{2}\ln\left(\frac{n}{2}\right) - \frac{m}{2} + \frac{m}{2}\ln(1+\epsilon) - \frac{1}{2}\ln(1+\epsilon) + \frac{n}{2}\ln(1+\epsilon)
\end{aligned}$$

となります。ただし、$O\left(\frac{m}{n}\right)$ を書くのを省略し、後半の式では

$$\epsilon \equiv \frac{m}{n}$$

とおきました。ϵ のゼロ次まで正確になるように、対数関数を ϵ についてテイラー展開します。右辺第 5 項については ϵ の 1 次まで含める必要がありますが（なぜなら対数の前に大きな数 $n = m\epsilon^{-1}$ があるので）、右辺第 3 項と 4 項については 0 次で十分です。すなわち

$$\frac{m}{2}\ln(1+\epsilon) \approx 0, \quad -\frac{1}{2}\ln(1+\epsilon) \approx 0, \quad \frac{n}{2}\ln(1+\epsilon) \approx \frac{n}{2}\epsilon = \frac{m}{2}$$

を代入することで

$$\ln\frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} = \frac{m}{2}\ln\left(\frac{n}{2}\right) + O\left(\frac{m}{n}\right)$$

がわかります。

2 について。自明な式変形から

$$\begin{aligned}\ln\left\{\left(1+\frac{m}{n}z\right)^{-\frac{m+n}{2}}\right\} &= -\frac{m+n}{2}\ln\left(1+\frac{mz}{n}\right)\\ &= -\frac{m}{2}\ln\left(1+\epsilon z\right)-\frac{n}{2}\ln\left(1+\epsilon z\right)\end{aligned}$$

のようになります。再び、右辺第 1 項の対数では ϵ の 0 次、右辺第 2 項の対数については（係数 $n = m\epsilon^{-1}$ のために）ϵ の 1 次まで含めて、

$$\begin{aligned}\ln\left\{\left(1+\frac{m}{n}z\right)^{-\frac{m+n}{2}}\right\} &= -0-\frac{n}{2}\epsilon z+O\left(\epsilon\right)\\ &= -\frac{m}{2}z+O\left(\frac{m}{n}\right)\end{aligned}$$

が出てきます。

3 について。以上の結果を使うと、$O\left(\frac{m}{n}\right)$ の誤差を許す近似において

$$\begin{aligned}&\ln\mathcal{F}(z\mid m,n)\\ &\quad\approx -\ln\Gamma\left(\frac{m}{2}\right)+\frac{m}{2}\ln\left(\frac{n}{2}\right)-\frac{m}{2}z+\frac{m}{2}\ln\left(\frac{m}{n}\right)+\left(\frac{m}{2}-1\right)\ln z\\ &\quad= -\ln\Gamma\left(\frac{m}{2}\right)+\frac{m}{2}\ln\left(\frac{m}{2}\right)-\frac{m}{2}z+\left(\frac{m}{2}-1\right)\ln z\\ &\quad= \ln\left\{\frac{1}{\Gamma\left(\frac{m}{2}\right)}\left(\frac{m}{2}\right)^{\frac{m}{2}}z^{\frac{m}{2}-1}\exp\left(-\frac{mz}{2}\right)\right\}\end{aligned}$$

したがって、

$$\mathcal{F}(z\mid m,n)\approx\frac{m}{2\Gamma\left(\frac{m}{2}\right)}\left(\frac{mz}{2}\right)^{\frac{m}{2}-1}\exp\left(-\frac{mz}{2}\right)\tag{6}$$

がわかります。これは式 (5) と一致します。

4 について。これは単なる確率変数の変数変換の問題です。カイ 2 乗分布の定義式（p.21）を眺めると、式 (6) は

$$z\sim\chi^2(z\mid m,1/m)$$

を意味することがわかります。ここで $u = mz$ に移ると

$$\begin{aligned}\chi^2\left(\frac{u}{m}\middle|m,\frac{1}{m}\right)\frac{\mathrm{d}z}{\mathrm{d}u} &= \frac{m}{2\Gamma\left(\frac{m}{2}\right)}\left(\frac{u}{2}\right)^{\frac{m}{2}-1}\exp\left(-\frac{u}{2}\right)\times\frac{1}{m}\\ &= \frac{1}{2\Gamma\left(\frac{m}{2}\right)}\left(\frac{u}{2}\right)^{\frac{m}{2}-1}\exp\left(-\frac{u}{2}\right)\\ &= \chi^2(u\mid m,1)\end{aligned}$$