最大値吸引領域の離散化に関する話題

志村　隆彰　数理・推論研究系　助教

【極値統計学における離散化の影響】

地震や洪水などの自然災害に代表される、めったに起こらないが、一旦起こると大変大きな影響を及ぼしたり、重要な意味を持つ現象は数多い。このような現象を統計的に扱う場合、日常を表す平均値のような指標ではなく、非日常を表わす最大値のような指標が重要になる。極値統計学はいわば非日常を研究対象にする分野であり、数学的にもっとも簡単で基本的な設定と関心事は$X_1, X_2, \ldots$を共通の確率分布Fに従う実数値独立確率変数列としたときのX_nまでの最大値$M_n = \max\{X_1, \ldots, X_n\}$の$n \to \infty$のときの挙動である。$M_n$が$F$の上端点$x_F = \sup\{x : F(x) < 1\}$（無限と有限の両方がありうる）に収束するのは明らかであるから、定数列$a_n > 0$と$b_n \in (-\infty, \infty)$により正規化された$(M_n - b_n)/a_n$を考える。これが$n \to \infty$のとき非退化分布$G$に分布収束するとき、極限分布$G$を極値分布、$F$は$G$の最大値吸引領域（以下、吸引領域とだけ記す）に属するという。吸引領域は分布Fの裾（確率）

$$\bar{F}(x) = 1 - F(x)$$

の$x \to x_F$のときの漸近挙動で特徴付けされるが、吸引領域への属性と分布Fの連続性の間には密接な関係があり、不連続性は吸引領域への属性を損なうことがある（裾が軽いほど大きな影響を受けることが知られている）。典型的な例は、指数分布と幾何分布である。このふたつの違いは連続と離散であるが、前者は吸引領域に属し、後者は属さない。そして、現実のデータは常に丸められた＝離散化されたものであるから、この種の統計解析を正確に行うためには、離散化の影響を考え、その対策を講じなければならない。

【分布の離散化と連続化】

分布Fの離散化とは、nを整数とするとき、$(n-1, n]$の測度を$\{n\}$に集めて、離散分布（整数値分布）にする操作をいう。離散分布の連続化とは、連続分布でその離散化した分布がもとの離散分布と一致するような連続分布を対応させる操作をいう。離散化と違い、連続化は唯一には決まらないことに注意する。

以下では 離散性に関連したふたつの結果を報告する。最初はポアソン分布の連続化である。

【ポアソン分布の連続版】

[1] では、前述の指数分布と幾何分布の関係を一般化し、吸引領域の分布の離散化にあたる分布の特徴付けをし、更に特定の場合として、次を示した。

定理 1 [1]　離散分布　F　に対し、$\bar{F}(n+1)/\bar{F}(n)$　が単調で$\limsup_{n\to\infty} \bar{F}(n+1)/\bar{F}(n) < 1$かつ

$$\lim_{n\to\infty} \frac{\bar{F}(n+2)\bar{F}(n)}{(\bar{F}(n+1))^2} = 1$$

ならば、ハザード関数が単調な連続化が漸近的に唯一つ存在する。

この定理は、吸引領域に入らない代表例であるポアソン分布に対しても吸引領域に入る自然な連続版が存在することを言っているが、それがどのようなものであるかはわからない。そこでその構成と裾の挙動を考える。

定理2　定理 1 のハザード関数が単調なパラメーターλのポアソン分布の連続化のひとつは次で与えられる。

$$F(x) = 1 - 1 \wedge \sum_{k=0}^{\infty} \frac{e^{-\lambda}\lambda^{x+k+1}}{\Gamma(x+2+k)}.$$

ここで、$\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$ である。

この分布の裾の漸近挙動は

$$\bar{F}(x) \sim \sqrt{\frac{\lambda}{2\pi}} e^{-\frac{3}{2}-\lambda} \left(\frac{e\lambda}{x}\right)^{x+\frac{3}{2}}$$

となる ($\sim$ は両辺の比が $x \to \infty$ で 1 へ収束するの意)。

【２つ目の数字による分布の推定】

ある数値の２つ目の数字とは、2014 であれば 0、3.14... ならば、1 のことである。また、それぞれの数値での 2 や 3 を先頭、或は最初の数字と呼ぶ。さて、標準正規分布表をみればわかるように、正規分布に従う乱数のうち、大きいもの、たとえば 2 以上或は 3 以上のものの２つ目の数字（分布表の範囲ならば小数点一桁目の数字になる）は 0 や 1 のような小さいものほど多く、8 や 9 は少ない（このことは少し考えればすぐにわかる）。この傾向は数値が大きくなるに従って強くなり、やがて２つ目の数字のほとんどは 0 になる。このように極端に大きな（或は）小さいランダムな数値には分布によって一定の傾向があるため、たとえばデータの２つ目の数字だけからでも分布の推定がある程度可能になる。基礎となるのは、一般に重い裾といわれるべきオーダーの裾を持つ分布に対する次の定理である。

定理3[2]　F が指数 $-\alpha(<0)$ の正則変動をする裾を持つと仮定する（任意の $\lambda > 0$ に対し、$\lim_{x\to\infty} \bar{F}(\lambda x)/\bar{F}(x) = \lambda^{-\alpha}$ が成り立つ）。X を分布 F に従う確率変数とし、N を X 桁数、K を X の先頭の数とし、$Y = 10^{-N+1}X - K$ とおく（X の先頭の数字を除き、それ以外を $[0,1)$ に写したもの）。このとき、$0 \le x \le 1$ に対し、

$$\lim_{n\to\infty} P(Y \le x | K = k, N = n), = \frac{1-(1+\frac{x}{k})^{-\alpha}}{1-(1+\frac{1}{k})^{-\alpha}} \quad k = 1, 2, \ldots, 9.$$

さて、定理３では極限分布が先頭の数字 k に依存して、９種類出てくるが、これは十進法であるためで、二進法にすれば、最初の数字は１に限られ、極限分布も一つに定まる。更に、パレート分布の場合は極限を取る必要がない、すなわち数値の絶対的な大きさによらないことに注意する。以下では、このことを利用して、パレート分布のパラメーター推定を試みたい。

パレート分布に従う確率変数の数値を二進法で表した場合、２つ目の数字 0 と 1 それぞれの確率は次で与えられる。

$$P(\{0\}) = \frac{1-\left(\frac{2}{3}\right)^{\alpha}}{1-\left(\frac{1}{2}\right)^{\alpha}}, \quad P(\{1\}) = \frac{\left(\frac{2}{3}\right)^{\alpha}-\left(\frac{1}{2}\right)^{\alpha}}{1-\left(\frac{1}{2}\right)^{\alpha}}.$$

従って、

$$r(\alpha) = \frac{1-\left(\frac{2}{3}\right)^{\alpha}}{1-\left(\frac{1}{2}\right)^{\alpha}}$$

とおき、n 個のデータのうち m 個の２つ目の数字が 0 であれば、$r(\alpha) = \frac{m}{n}$ の解として、パレート分布のパラメーターα の推定量 $\hat{\alpha}$ が得られる。

参考資料

[1] Discretization of distributions in the maximum domain of attraction, Extremes, 15 (2012) 299-317.

[2] A numerical characteristic of extreme values, to appear in Statistics and its Interface.

http://www.ism.ac.jp/ shimura/