分布の違いによる極値データの数値的特徴

志村　隆彰　数理・推論研究系 准教授

1　研究の概略

大きな正規乱数の小数点 1 桁目は多くが 0 であり、9 であることは稀である。このようにランダムな実数値に現れる数字 $\{0,1,2,\ldots,9\}$ の出方には一定の法則がある。極値データとは多くの数値データの内、極端に大きい、或いは小さいものとし、極値データが特定の（確率）分布に従ってランダムに生じたとき、分布の違いが極値データの数値的な特徴として現れることについて述べる。

2　問題の定式化

データを表す（実数値）確率変数を X とするとき、極値データは、大きな x に対して、$X>x$ を満たすもので、この事象の確率は$P(X>x)$ と表され、F を X の分布関数とすれば $1-F(x)$ と書ける。これを分布 F の裾(tail of a distribution F)といい、$\bar{F}(x)=1-F(x)$ で表す。極値データは分布の裾に従うから、$x\to\infty$ のときの裾の（漸近）挙動（減衰の速さ）が重要である。

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

以下、分布 F は上端点が無限、すなわち、いくらでも大きい値を取りうるものとする（下端点は1 としておく）。ここで、実数値の極値データを $[0,1)$ へ移す次の変換を考える。

$$d_1d_2d_3\ldots d_n.d_{n+1}\ldots \text{ in } [10^{n-1},10^n)$$
$$\to 0.d_2d_3\ldots \text{ in } [0,1),$$

ここで、$d_i\in\{0,1,2,\ldots,9\}$(i = 1, 2, 3, ...)、n は自然数である。

この変換は、大きな数値の小数点の位置をずらして、1以上10 未満にスケール変換したのち、整数部分を引き、$[0,1)$ に入るようにするものである。たとえば、$2023.52612\ldots$ は $0.023.52612\ldots$ に写される。この変換で極値データの最初の数字が省かれることに注意する。

極値を表す確率変数 $X=d_1d_2d_3\ldots d_n.d_{n+1}\ldots$($n$桁の自然数＋小数部分) をこの変換で $[0,1)$上の確率変数 $Y=0.d_2d_3\ldots d_nd_{n+1}\ldots$ で写し、削除した最初の数字$d_1=0,1,2,\ldots,9$ で条件付けたものの挙動を考える。自然数値確率変数 N と K を

$$N:10^{N-1}\le X<10^N,\quad K:K10^{N-1}\le X<(K+1)10^{N-1}$$

とすれば、

$$Y=(X-K10^{N-1})/10^{N-1}$$

である。次の条件付確率分布を考える。

$$F^{k,n}(x)=P(Y\le x|K=k,N=n),$$
$$for\ k=1,2,\ldots,9.$$

この分布は極値データの最初の数字を固定したとき、２つ目以降の数字の分布を意味していて、各 $k=1,2,\ldots,9$ に対して、$n\to\infty$ としたときの $F^{k,n}$ の挙動に関心がある。

【裾挙動（減衰の速さ）によって決まる確率分布族】

裾挙動を表すために使う正則変動関数とその関連概念とそれを用いた裾の減衰で定義される分布族を紹介する。

1. 正則変動関数と正則変動をする裾を持つ分布

正値可測関数 $f(x)$ が指数 ρ $(\in\mathbf{R})$ の正則変動関数(regularly varying function $f\in\mathbf{R}_\rho$) とは、任意の $k>0$ に対し、

$$\lim_{x\to\infty} f(kx)/f(x)=k^\rho.$$

となるときをいう。$\rho=0$ のとき、特に緩慢変動（slowly varying）という。正則変動性は、べきのオーダー $f(x)=x^\rho$ を拡張した概念である。コーシー分布、パレート分布、F分布、ジップ分布などの裾は正則変動する（指数は負）。

2. Π変動関数とΠ変動をする裾を持つ分布

$(0,\infty)$ 上の正値可測関数 $f(x)$ は、$(0,\infty)$上の正値関数 $a(x)$ があって、任意の $\lambda>0$ に対し、

$$\lim_{x\to\infty}\frac{f(\lambda x)-f(x)}{a(x)}=\log\lambda.$$

となるとき、Π 変動するという（$f\in\Pi$ 或いは $f\in\Pi(a)$ とかく）。

$f(x)=\log x$ は Π変動する（$a(x)=1$）。Π変動は緩慢変動の一部であり、裾が1/Π変動する分布は非常に重く、対数コーシー分布はその例である。

3. 急速変動関数と急速変動をする裾を持つ分布

正値可測関数 $f(x)$ は任意の $\lambda>1$ に対し、$\lim_{x\to\infty} f(\lambda x)/f(x)=\infty.$ となるとき、（指数無限の）急速変動関数という（$f\in\mathbf{R}_\infty$）。$f(x)=\exp x$ が例となる。$(f\in\mathbf{R}_\infty)$ に対し、$1/f(x)$ を指数マイナス無限の急速変動という（$f\in\mathbf{R}_{-\infty}$）。

正規分布、指数分布、対数正規分布など、多くの分布の裾が急速変動する。

3　結果

$F^{k,n}(x)=P(Y\le x|K=k,N=n)\quad(k=1,2,\ldots,9)$ の $n\to\infty$ の挙動について次が成り立つ。

定理

(i) $\bar{F}(x)\in\mathbf{R}_{-\infty}$ ならば、すべての k に対して、

$$\lim_{n\to\infty}F^{k,n}(x)=1_{\{x\ge1\}},$$

ここで、1_A は集合 A の集合関数である。

つまり、$F^{k,n}$ は $n\to\infty$ のとき δ_0（$\{0\}$ に退化した分布)に収束する。

(ii) $\bar{F}(x)\in\mathbf{R}_{-\alpha}(\alpha>\mathbf{0})$ ならば、$0\le x\le1$ に対し、

$$F^k(x)=\lim_{n\to\infty}F^{k,n}(x)=\frac{1-(1+\frac{x}{k})^{-\alpha}}{1-(1+\frac{1}{k})^{-\alpha}}.$$

(iii) $1/\bar{F}(x)\in\Pi$ ならば、$0\le x\le1$ に対し、

$$F^k(x)=\lim_{n\to\infty}F^{k,n}(x)=\frac{\log(1+\frac{x}{k})}{\log(1+\frac{1}{k})}.$$

(i) は極限分布が退化してしまって、極値データの最初の数字である k が極限分布に現れないが、(ii) とその極限といえる(iii) では、最初の数字でそれに続く数字の出方が変わることがわかる（いずれの場合も小さい数字の方が出る確率が大きい）。

この定理は、分布の違いがその分布から生じるデータの数字に反映されることを示す一例である。この他にも、上限のある分布の上限までの長さの分布や極値データのm桁目$(m=2,3,\ldots)$の分布なども求めることができる。冒頭で触れた大きな正規乱数の例は、乱数の整数部分と小数部分に関して、正規乱数の場合、整数部分が大きくなると小数部分が δ_0 に収束することから導かれることである。指数分布に従う乱数では、小数部分は整数部分とは独立で、ある特定の分布になり、正規乱数の場合とは異なる。パレート分布ではこれらとは違う。詳しいことは、下記参考文献をご覧いただきたい。

【参考文献】

T.Shimura(2011). Limit distribution of a roundoff error, Statistics and Probability Letters **82**, 713-719.

T.Shimura(2014). A numerical characteristic of extreme values, Statistics and its Interface, **7**, 375-379.