# 極値データへの丸め誤差の影響

志村　隆彰　　数理・推論研究系　助教

2010年7月9日　統計数理研究所　オープンハウス

## 【極値統計学とは】

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、中心極限定理に倣い、定数列 $a_n > 0, b_n$ で正規化（スケール変換）したものを考える。$F$ が適当な条件を満たすならば、(非退化な)極限分布が存在する（$\mathcal{L}$ は分布の意)。

$$\mathcal{L}\left(\frac{M_n - b_n}{a_n}\right) \to G \quad (n \to \infty).$$

このときの極限分布$G$を極値分布といい、フレシェ分布、(逆)ワイブル分布、グンベル分布の3種類があり、$F$ は $G$ に吸引されるといい、$G$ に吸引される分布全体を $G$ の吸引領域と呼ぶ（$D(G)$ と書く）。$a_n, b_n$ は規格化定数と呼ばれ、$b_n$ は$F$ の上端点に収束し、$a_n$ は無限、有限正定数、0 へ行く外、理論上は振動することもあり、その挙動は様々である。吸引領域は分布 $F$ の裾 $\bar{F}(x) = 1 - F(x)$ が $F$ の上端点 $x_F$ に近づくときの漸近挙動で特徴付けられる。
ほとんどの連続分布はいずれかの極値分布の吸引領域に属することが知られている。一方で、ほとんどの離散分布はこれに含まれない。従って、連続の場合とは違って極限定理が成り立たず、離散の場合の扱いは困難である。

## 【丸め誤差の極限分布】

観測値に誤差が加わった状況を考える。誤差が加わると分布が変わってしまい、吸引領域から外れ、結果、極限定理が成り立たなくなる心配がある。西山・志村（2009）(統計数理57巻)では極限に影響を及ぼさない誤差の程度について考察したが、今回は誤差として丸め誤差を取り上げる。つまり、観測値が一定の精度までしか得られない状況を考察する。以下、$F$ の上端点が無限の場合のみを考え、分布 $F$ に対して、丸め誤差が加わった分布を $F_1$ と書く。確率変数でいえば、一般性を失うことなく、(連続型）確率変数 $X$ に対し、その整数部分である（離散型）確率変数 $[X]$ を対応させることに相当する。小数部分を $\{X\} = X - [X]$ と書く。$[X]$ が観測値、$\{X\}$ が丸め誤差を意味する。
$F \in D(G)$ であるとき、$F_1 \in D(G)$ であるための必要十分条件は$F \in \mathcal{L}$ : $\lim_{x\to\infty} \bar{F}(x+1)/\bar{F}(x) = 1$ であることが知られている(Shimura(2011))。パレート分布、コーシー分布などがこれに当たる。一方、指数分布や正規分布といった裾の重くない分布では、$F_1 \notin D(G)$ となってしまう。つまり、$X$ が大きい場合を考えているにもかかわらず、$[X]$ だけではなくて、$\{X\}$ が $D(G)$ への属性に大きく影響するのである。

この $[X]$ の値が大きい場合の $\{X\}$ の条件付き分布

$$F_n(x) = P(\{X\} \le x | [X] = n)$$

の $n \to \infty$ のときの極限を考える。

結果を示す前に、裾の挙動（重さの程度）により分布族を定義する。$\mathcal{L}(\infty)$ は、各$k > 0$ に対し、$\lim_{x\to\infty} \bar{F}(x+k)/\bar{F}(x) = 0$ となる分布全体からなる分布族とする。正規分布、レイリー分布、指数が１を超えるワイブル分布などがこれに含まれ、裾が軽い分布の族である。$\mathcal{L}(\gamma)(\gamma > 0)$ は各 $k \in \mathbf{R}$ に対して、$\lim_{x\to\infty} \bar{F}(x+k)/\bar{F}(x) = e^{-\gamma k}$ が成り立つ分布全体からなる分布族とする。$\mathcal{L}(\gamma)$ の分布は指数的裾を持つといわれる。指数分布、ガンマ分布、カイ二乗分布、一般化逆ガウス分布などが $\mathcal{L}(\gamma)$ に含まれる。これは裾が中程度の分布の族である。$\mathcal{L}(\mathcal{L}(0))$ は既に定義したが、これに属する分布を long-tailed という。最後に、$F$ が $\mathcal{L}_0$ に属するとは、$F \in \mathcal{L}$ かつ そのハザード関数 $h(t)$ が各 $k \in \mathbf{R}$ に対して、$\lim_{\to\infty} h(t+k)/h(t) = 1$ を満たすときをいう。コーシー分布、対数正規分布、パレート分布、F分布などが$\mathcal{L}_0$ に入り、裾が重い分布の族である。

上端点が無限の連続分布のほとんどはこれらの分布族のいずれかに入るといっても過言ではない。そして、$F_n(x)$ の極限分布は$F$ がどの分布族に属するかによって決まる。

**定理１**

(i) $F \in \mathcal{L}(\infty)$ ならば、$F_n$ は$n \to \infty$ のとき $\delta_0$（$\{0\}$ に集中した分布）に収束する。

(ii) $F \in \mathcal{L}(\gamma)(\gamma > 0)$ ならば、$F_n$ は $n \to \infty$ のとき 平均 $\gamma^{-1}$ の指数分布の小数部分の分布（$Fe(\gamma)$ と記す）に収束する。

(iii) $F \in \mathcal{L}_0$ ならば、$F_n$ は $n \to \infty$ のとき $U(0,1)$（$[0,1]$ 上の一様分布）に収束する。

この結果は、丸め誤差が観測値と必ずしも独立ではなく、観測値が大きい場合には分布の裾挙動と密接に関連して３通りの分布型に分類できることを主張している。(iii) は他と違う形になっているが、実際次が成り立つ。

**定理２**　任意の $F \in \mathcal{L}$ と任意の $[0,1]$ の分布 $G$ に対して、$\lim_{x\to\infty} \bar{F}_0(x)/\bar{F}(x) = 1$ かつ $F_0$ の小数部分の分布が $G$ であるような $F_0 \in \mathcal{L}$ が存在する。

ハザード関数へ緩い条件を付けると逆が言える。

**定理３**　$F$ のハザード関数が収束するか無限に行くものとする。

(i) $F \in \mathcal{L}(\infty)$ と $F_n \to \delta_0 (n \to \infty)$ とは同値である。

(ii) $F \in \mathcal{L}(\gamma)$ と $F_n \to Fe(\gamma) (n \to \infty)$ とは同値である。

(iii) $F \in \mathcal{L}$ と $F_n \to U(0,1) (n \to \infty)$ とは同値である。

**注意**　よく知られた分布のほとんどは、ハザード関数があるところから先で単調になるので、定理３の条件を満たしている。

離散分布 $F_1$ が指数 $\gamma$ の幾何的分布であるとは、$\lim_{n\to\infty} \bar{F}_1(n+1)/\bar{F}_1(n) = e^{-\gamma}$ を満たすものをいう。これは指数的裾を持つ分布を持つ分布の整数部分の分布で吸引領域には属さないが、適当な補足をすることで吸引領域に属させることが出来る。

**定理４**　$F_1$ をパラメーター$\gamma(>0)$ の幾何的分布、$G$ を $[0,1]$ 上の分布とするとき、$F_1 * G \in \mathcal{L}(\gamma)$ と $G$ が $Fe(\gamma)$ であることは同値である。

## 【丸め誤差の吸引領域への影響】

丸め誤差の吸引領域への影響をまとめる。

$D(\Phi_\alpha)(\alpha > 0)$ と $D(\Lambda)$ でフレシェ分布及びグンベル分布の吸引領域を表わす。$D(\Phi_\alpha) \subset \mathcal{L}$、$D(\Lambda) \cap \mathcal{L} \neq \emptyset$、$\mathcal{L}(\gamma) \subset D(\Lambda)$、$D(\Lambda) \cap \mathcal{L}(\infty) \neq \emptyset$ が知られている。$F \in \mathcal{L}$ ならば、$F_1$ も同じ吸引領域に属する。$F \in \mathcal{L}(\gamma)$ のとき、$F \in D(\Lambda)$ だが、$F_1$ は $D(\Lambda)$ から外れる。しかし、$F_1$ は Fe 分布との合成積をとることで $D(\Lambda)$ に復帰する。この性質を持つのは Fe 分布だけである。$\mathcal{L}(\infty)$ の分布に対しては、観測値の絶対的な精度は観測値が大きくなるにつれて高まるにもかかわらず、吸引領域に属することはない。

### 参考資料

http://www.ism.ac.jp/ shimura/ をご覧ください。