# 対数正規変量に対する単純集計問題とトリムド平均

リスク解析戦略研究センター　データ科学研究系
教授　　椿　広計

## 1　問題の発端

### 1.1　官庁統計で頻出する対数正規変量の算術平均

総合エネルギー統計は､需要側でのエネルギー使用量を捕捉するために資源エネルギー庁が実施している。この単純集計は、外れ値の検出基準を変えることで総使用量推計値が大きく変動するとされている。基本的には、ボックスプロットの髭の外側を外れ値とする、あるいは、データが対数正規分布に近い分布をしめすので、対数変換を行った上で外れ値除去処理を行い、その後算術平均を求める処理が行われてきた。これは、「外れ値」の検出問題ではなく、ただの点推定論上の要請と考えることもできる。使用量分布は対数正規分布に「近い」挙動を示し、当該変量に対して、期待値推定のために算術平均処理が行われているのである。データ解析家の立場としては、「対数平均の報告ないしはそれを原変数に戻した幾何平均の報告が意味がある」と言ってしまいそうだが、社会的に関心があるのは「総量」の推定なので、実務的には原変数の単純集計を超えた操作を行う事は極めて困難である。そこで、分布と推論の不整合を解決するために、外れ値除去と称してトリムド平均(Trimmed Mean)により推定値を安定化しているのである。それでは､どのような｢見なし外れ値｣基準を設けたトリムド平均を設計するのが妥当なのだろうか。これは、官庁統計分野では日常的に起きていることと考えられる。

### 1.2　定式化と準備

観測変量$Y$は、互いに独立に同一の対数正規変量　$\log Y_i \sim NID(\mu,\sigma^2)$, $i=1,\ldots,n$　に従うとすると、$E[Y_i]=\exp(\mu+\sigma^2/2)$，$Var[Y_i]=\exp(2\mu+\sigma^2)\{\exp(\sigma^2)-1\}$，$CV^2=\exp(\sigma^2)-1$　となる。算術平均　$AM=\sum y_i/n$ は、期待値の不偏推定量であり、その分散は、 $\exp(2\mu+\sigma^2)\{\exp(\sigma^2)-1\}/n$ となる。算術平均は、$\sigma^2$ が既知ならば非許容的であり、最良不変推定量も存在し、縮小係数を$n/(n+CV^2)=n/\{\exp(\sigma^2)+n-1\}$ とする単純な縮小推定ですら、算術平均の平均二乗誤差を一様に改善可能である。一方、期待値の最尤推定量は、$MLE=GM\ \exp[\sum(\log y_i-\sum\log y_j/n)^2/\{2(n-1)\}]$となる。ここで、$GM$は、幾何平均 $GM=\exp(\sum\log y_i/n)$である。最尤推定量の期待値は、$E[MLE]=\exp\{\mu+\sigma^2/(2n)\}\ \{1-\sigma^2/(n-1)\}^{-(n-1)/2}$であり、漸近分散はAsy.Var[$MLE$]=$\exp(2\mu+\sigma^2)(\sigma^2+\sigma^4/2)/n$となり、$\sigma$が大きくなると$MLE$の分散は$AM$に比べ改善する。しかし、期待値のMLEが実務的に使われることは極めて稀である。

## 2　方便としての外れ値処理とトリムド平均

### 2.1　提案する方法

ここでは、変量の従う分布形から見て好ましい推定量、特に、最尤推定量にトリムド平均によって合わせこみを行うという素朴な方法を提唱し、その性能を調べる。

標本に対して順序統計量を　$Y_{(1)} < Y_{(2)} <\ldots<Y_{(n)}$　とし、トリム平均 を

$$M(L,U)=\{Y_{(L)}+Y_{(L+1)}+\ldots+Y_{(U-1)}+Y_{(U)}\}/(U-L+1)$$

としたとき、$L$, $U$を次の基準(1)が最小化されるように選択する方法を提唱する。

$$R(L,U)=\{M(L,U)-\boldsymbol{MLE}\}^2+\sum_{L\le k\le U}\{Y_{(k)}-M(L,U)\}^2/\{(U-L)\,(U-L+1)\} \qquad (1)$$

(1)式右辺第1項が偏り、第2項が分散に対するペナルティ項である。以下では、簡単のために$L$=1の場合の提案方法の挙動を調べる。

## 2.2　提案する方法の挙動

$\mu$=0, $\sigma$=3の対数正規乱数を100個生成し、基準(1)に基づく上側だけをトリムする平均の挙動を調べた。**Fig.1**に原標本（左列）、対数変換標本（右列）のヒストグラムとボックスプロットを**Fig.2**に、基準(1)と非トリム率との関係（上段）、トリム平均値と非トリム率との関係（中段）、並びにトリム平均値と基準(1)との関係を示した。この例では、最大値がトリムされている。

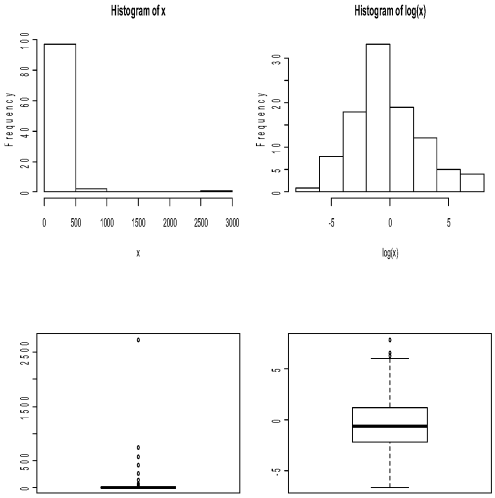


**Fig.1 乱数で発生させた標本の分布**

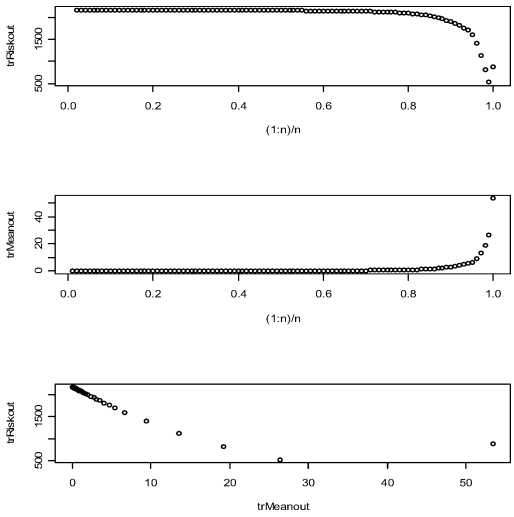


**Fig.2　トリムされる個数と基準(1)**

1996年のわが国上場企業2091社の売上高(単位100万円)データから200社をランダムサンプリングすることを想定し、算術平均、対数正規分布を想定したMLE、提案方法をBootstrap分布で比較した。提案方法を検討する**Fig.2** に相当するプロットを**Fig.3**（この例では、上位2社がトリム）に、3種の推定量のBootstrap分布を**Fig.4**に示す（左列:算術平均、中央列:MLE、右列:MLEに合わせこんだトリム平均）。

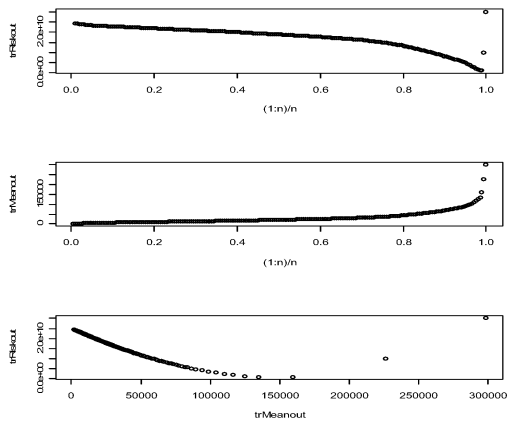


**Fig. 3　実データでトリムされる個数と基準(1)**

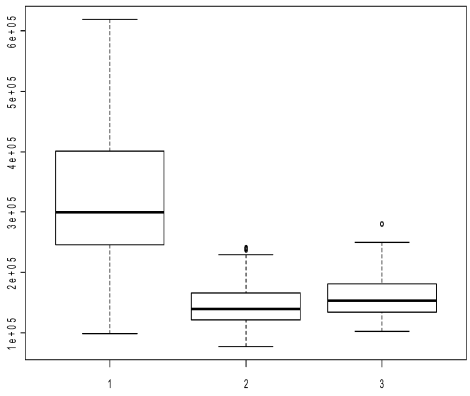

**Fig. 4 3推定方式のBootstrap分布**

このデータでは算術平均は、Bootstrap分布を眺める限りMLEに比べて不安定となっていることが分かる。提案方法は少なくともMLEにはよく合わせこまれ、安定的な推定量となっていることも分かる。この方法自体は、算術平均よりは外れ値に対して頑健であることは予期されるが、本当に外れ値が存在する場合の頑健推定量を与えたわけではないことには注意が必要である。