Technical Report YK-020
Feb. 18, 2019
Y. Karasawa

極値統計学へのいざない

～想定外の出来事を想定外としないために～

唐沢好男

人生１００年の時代である。これに比べてはるかに長いスケールでしか起きないであろう最悪事態、そのような現象を、たまたま、目の当たりにすることになったとき、それを想定外の出来事と呼ぶであろう。超大型の地震、暴風、豪雨、噴火などによる自然災害（catastrophe）はそういうものである。このような最悪事態の発生、すなわち観測量の最大値を扱う学問が極値統計学である。想定外の出来事を想定外としないための心積もりを与える学問とも言える。極値統計学は、その古典的な学術書[1]や最新の教科書（例えば[2]）からも分かるように、数理統計学に基盤を置いて、体系的な学問として完成されている。しかし、その体系は高度に数学的であり、「例えば、100 年に一回の最悪値はどのくらいだろう」と言うシンプルな問の答えに行き着くまでには、越えるべきハードルが高い。本レポートは、上記答えを簡易に求める方法についてのみに絞り、極値統計学の壺をまとめる。すなわち、難しい理論や高度な応用は極値統計学の専門書にお任せし、実データ解析に役立ちそうな部分をつまみ食いして簡潔にまとめる。計算機シミュレーションデータを用いて、イメージを捕らえやすくしているので、極値統計学のイロハの部分は大体分かったというレベルにはたどり着くと思う（筆者の知識もこのあたりまで)。なお、このテーマをまとめてみようと思いたった動機を以下のコラムに。

前レポート（YK-019 [3]）では、回帰分析の基本である単回帰モデルと信頼区間をまとめた。その続編として、多変量解析法のエッセンスをまとめたいと思い、池袋のジュンク堂に行って、材料集めをしていた。その棚で「極値統計学によろしく：暗黒通信団」[4]という 24 ページの超薄い本を見つけた。本のタイトルや出版社（さらには表紙の漫画）から、何でこんな棚にこんなものがという怪しいイメージを受けたが、ぱらぱら捲ってみるとユーモアを交えつつも至極まともな本で大変感動した（250 円と言う安さも嬉しい)。これに刺激を受け、かつ、この本から有用なイメージをもらったので、予定を極値統計のテーマに切り替え、無線工学分野に役立つように、筆者流のまとめ方をしてみた。

Technical Report YK-020
Feb. 18, 2019
Y. Karasawa

１．順序統計

１．１　順序統計量の確率分布

極値統計とは、確率過程に基づく N 個の標本値（観測値）の最大値、あるいは、最大値を含むいくつかの値の統計的性質を調べたり予測したりするものであり、極値統計学はその理論を与える。本節では、その基になる、順序統計の概要を述べる。

順序統計量とは、N 個の標本値を得る統計において n 番目に小さい値である標本を言う。日本工業規格では「標本の全ての観測値をその大きさの順に小さいほうから並べたもの。また、より一般的には、この並び替えの関数として求められることの全てを指すこともある。」と定義されている。１番目の値が最小値、N 番目の値が最大値である。

標本 $x_1, x_2, \cdots, x_n, \cdots, x_N$ は、連続確率分布（確率密度関数）$f(x)$に従うものであって、ランダムに抽出されたものであるとする。また、標本値を小さい順に並べなおした順序標本値を $x_{(1)}, x_{(2)}, \cdots, x_{(n)}, \cdots, x_{(N)}$とする。

このとき、n 番目の順序統計量 $x_{(n)}$の累積分布関数 $F_N(x;n)$は、元の累積分布関数 $F(x)$を用いて、次式となる。

$$F_N(x;n)=\sum_{k=n}^{N}\binom{N}{k}\left\{P\left(x_{(n)}\le x\right)\right\}^k\left(1-\left\{P\left(x_{(n)}\le x\right)\right\}^{N-k}\right)$$

$$=\sum_{k=n}^{N}\binom{N}{k}F^k(x)\left(1-F(x)\right)^{N-k} \tag{1}$$

$$\text{where}\quad \binom{N}{k}\equiv\frac{N!}{k!(N-k)!}$$

確率密度関数は累積分布の導関数であり、次式となる（導出は、例えば[5]）。

$$f_N(x;n)=N\binom{N-1}{n-1}F^{n-1}(x)\left(1-F(x)\right)^{N-n}\,f(x) \tag{2}$$

最大値 $x_{(N)}$では、累積分布関数と確率密度関数は次式となる。

$$F_N(x;N)=F^N(x) \tag{3a}$$

$$f_N(x;N)=NF^{N-1}(x)\,f(x) \tag{3b}$$

１．２　観測値を累積分布としてプロットする方法

測定や計算機シミュレーションによって、N 個の出力値（$x_1, x_2, \cdots, x_N$）を得た場合、このデータを累積分布としてプロットする場合の方法について述べる。

先ずは、最初に、データを小さい順（あるいは大きい順）に並べ直す（ソートする）。これを、$x_{(1)}, x_{(2)}, \cdots, x_{(N)}$と置く。ここまでは良いであろう。次のステップとして、具体的に累積分布としてグラフ化するとき、悩んだことは無いだろうか？$x_{(n)}$を $F=n/N$ にプロットするのは良いのだろうか。でもそれだと $x_{(1)}$は $F=1/N$ であるので良いとしても、$x_{(N)}$側は $F=1$ となって、明らかに正しくない。それではということで $F=(n-1)/N$ とすると、$x_{(N)}$の問題は解決されるが、$x_{(1)}$側が正しくなくなる。ではどうすれば？合理的な考え方に基づく一つの答えは次式である。

$$x_{(n)} \quad \rightarrow F = \frac{n}{N+1} \tag{4}$$

累積確率の 0～1 の区間を $1/(N+1)$分割し、両端（0 と 1）を除いた N 個の分岐点に順次プロットしてゆくのである。以下は、(4)式の根拠を与える大筋の説明である（詳細は[2]の付録 A.3）。

最初に、一様分布（確率変数 x_u）の順序統計量の期待値を求める。一様分布（0～1 区間での）の確率密度関数 f_u と累積分布関数 F_u は次式である。

$$f_u(x_u) = 1, \;\; F_u(x_u) = x_u \quad (0 \le x_u \le 1) \tag{5a, b}$$

n 番目の値の確率密度関数は(2), (5a)式より

$$f_{Nu}(x_{u(n)}) = N\binom{N-1}{n-1} x_{u(n)}^{n-1}\left(1-x_{u(n)}\right)^{N-n} \tag{6}$$

であるので、その期待値は次式となる。

$$\left\langle x_{u(n)} \right\rangle = \int_0^1 x_u \; f_u(x_u)dx_u$$

$$= N\binom{N-1}{n-1}\int_0^1 x_u^n \left(1-x_u\right)^{N-n} dx_u = \frac{n}{N+1} \tag{7}$$

次に、一様分布と目的とする分布の同累積確率値に着目した対応関係を見る。

$$F_u(x_{u0}) = F(x_0) \tag{8}$$

であれば、

$$x_0 = F^{-1}F_u(x_{u0}) \tag{9}$$

であるので、

$$\left\langle x_{(n)} \right\rangle = F^{-1}F_u\left(\left\langle x_{u(n)} \right\rangle\right) = F^{-1}\left(\frac{n}{N+1}\right) \tag{10}$$

となり、

$$F\left(\left\langle x_{(n)} \right\rangle\right) = \frac{n}{N+1} \tag{11}$$

となる。図１はこの関係をまとめている。

順位付けされた N 個の観測データの n 番目の値を累積確率は、0～1 区間を N+1 分割し、n 番目の位置に配してゆくことは、その期待値が累積確率 $n/(N+1)$になるのであるから、(4)式の対応付けは合理的であるといえる。

なお、信号強度の分布など、我々が日常で行う評価などでは、累積確率の 0 から 1 まで全体を正確に見たいことは稀で、どちらか一方の側だけを正確に見たいと言う場合が多い。そのような場合、N の値が十分大きければ、全体を N 分割して、見たい側を 1/N から置いてゆくことで問題は無い。ただ、原理原則は知っておいてほしいと言う思いと、この関係を３節での最大値推定に利用することの下準備の意味で、この節を設けた。

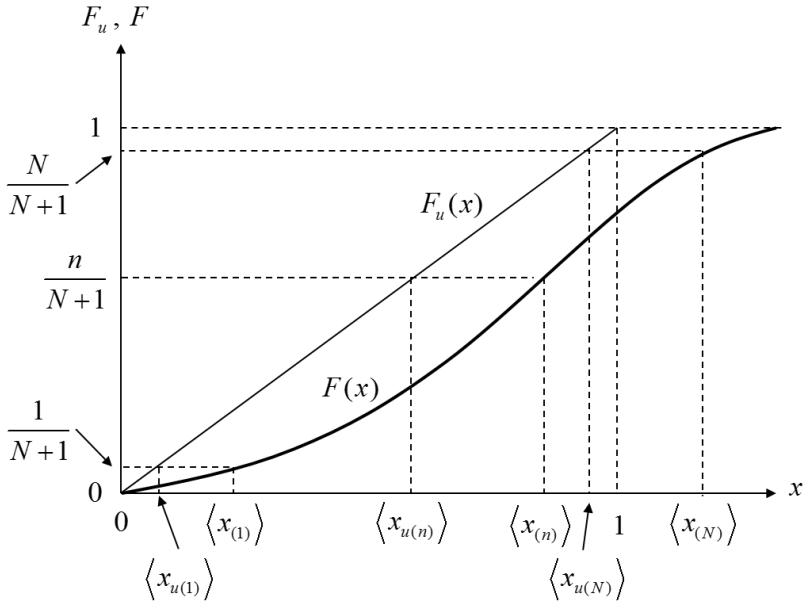


図１　任意の確率分布と一様分布の累積確率の関係

２．最大値の漸近確率分布（極値分布）

極値統計学の専門書や学術論文では、最大値の確率分布に着目し、その普遍化に重きが置かれていて、かつ、理論的にも難易度の高い部分になる。本レポートでは、３節での、最大値そのものの求め方を主題にしたいため、この部分は天下り的に結果式を示すに留める。詳しいところは、専門書（例えば[2]）で学んでほしい。

データ数 N が十分大きくなると、最大値の確率分布の形は、元の分布 $f(x)$に依存しなくなる。このような漸近特性を有する性質は、和の分布に対する中心極限定理（正規分布に収束）にも見られる。

以下の説明では、最大値 $x_{(N)}$を x_{max} で置き換え、さらに、それを以下の式で規格化する。

$$z = \frac{x_{max} - b_N}{a_N} \tag{12}$$

規格化の係数 a_N, b_N は吸引係数とも呼ばれる。元になる変数の確率分布によって異なり、N の関数で与えられる。この吸引係数が適当に定められたときの z の漸近確率分布（N が十分大きいときに漸近する確率分布：極値分布と呼ばれる）は、以下の三つの形のどれかで表すことができる（この三つに限ることが証明されている）。極値分布の累積確率分布 F と確率密度関数 f は以下のとおりである。

１）Gumbel（グンベル）分布

$$F_G(z) = \exp\{-\exp(-z)\} \qquad (-\infty < z < \infty) \tag{13a}$$

$$f_G(z) = e^{-z} F_G(z) = \exp\{-z - \exp(-z)\} \tag{13b}$$

２）Fréchet（フレッシェ）分布

$$F_F(z) = \exp(-z^{-\alpha}) \quad (z \geq 0,\ \alpha > 0) \tag{14a}$$

$$f_F(z) = \alpha z^{-\alpha-1} F_F(z) = \alpha z^{-\alpha-1} \exp(-z^{-\alpha}) \tag{14b}$$

３）（負の）Weibull（ワイブル）分布

$$F_W(z) = \exp\{-(-z)^{\alpha}\} \quad (z \leq 0,\ \alpha > 0) \tag{15a}$$

$$f_W(z) = \alpha(-z)^{\alpha-1} F_W(z) = \alpha(-z)^{\alpha-1} \exp\{-(-z)^{\alpha}\} \tag{15b}$$

個々の確率分布の漸近分布がどの形になるかを機械的に分類することは困難であるが、目安として次の指針が与えられている。

・分布の右袖が指数関数的に減少する連続分布（正規分布、レイリー分布、仲上 m 分布、仲上・ライス分布、指数分布、ガンマ分布、対数正規分布など）では、Gumbel 分布型
・右袖がベキで減少する裾の厚い分布（t 分布、コーシー分布など）は Fréchet 分布型
・上限のある分布（例えば一様分布）は（負の）Weibull 分布型

上述の分類から分かるように、電波伝搬や通信信号処理に現れる確率分布では、その大部分が Gumbel 分布型に属する。図２は Gumbel 分布の累積分布と確率密度関数を図示している。N の増加に対するこの分布への収斂は極めて緩慢であることが調べられている。[2]では、正規分布について評価しているが、N=10,000 でもまだ十分な収束には至っていないことが示されている（同文献の図２．７）。代表的な分布に対する係数 a_N, b_N の式は[2]（の表２．４）に与えられているので、これを用いて、ガンマ分布（本資料の式(26a)で、ν=2, β=0.5）の極値分布への収束具合を乱数を用いた計算機シミュレーションで調べてみる（具体的なガンマ分布生成法は４節(3)に）。N=100 と 10,000 に対して、それぞれ 10,000 回の試行を行って、得られた最大値のヒストグラムを求めた。この結果を図３に示す。同図には、理論式による極値分布 f_G も示している。N=100 ではずれが大きく、N=10,000 で漸く近づいてきていることが分かる。このように収斂が緩慢であることが理解できると思う。

理論で求める極値分布は、N が十分大きいと言う条件では正しいのであるが、任意の確率分布に対して、吸引パラメーラ a_N, b_N を理論的に求めることが容易ではなく、かつ、N に対する収斂も遅いため、最大値そのものを推定するには、このアプローチは適していない。そこで、次節では、最大値のみを簡易に求める方法について述べる。

本節での結果は以下に要約できる。

i) 最大値の漸近確率分布（極値分布）は三つのタイプがあり、どれかに分類される
ii) 元の確率分布と極値分布の形の対応付けについては、目安となる指針がある
iii) 元の確率分布から極値分布の吸引係数を理論的に求めるのは容易でなく、かつ、その適用も N の極めて大きい部分（例えば、$N\geq$10,000）に限られるので、この方法で最大値を推定するのは良いアプローチとは言えない
iv) 最大値の推定のみであれば、次節で述べる簡易な方法がある

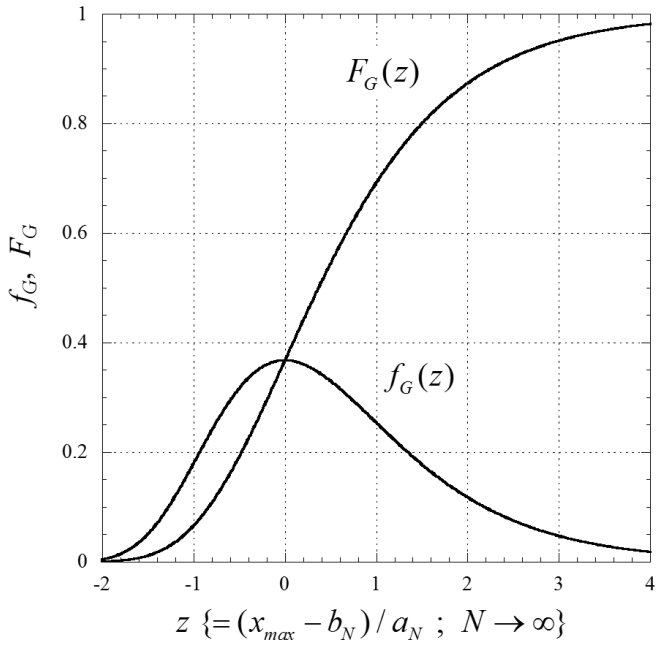
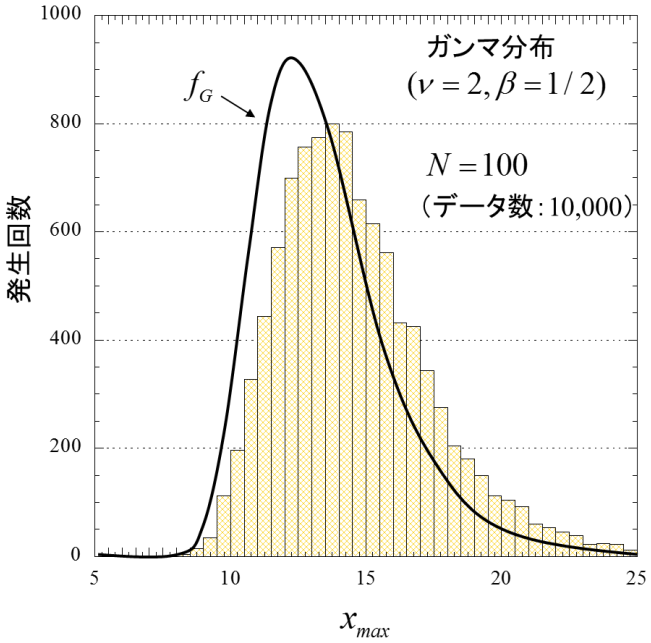
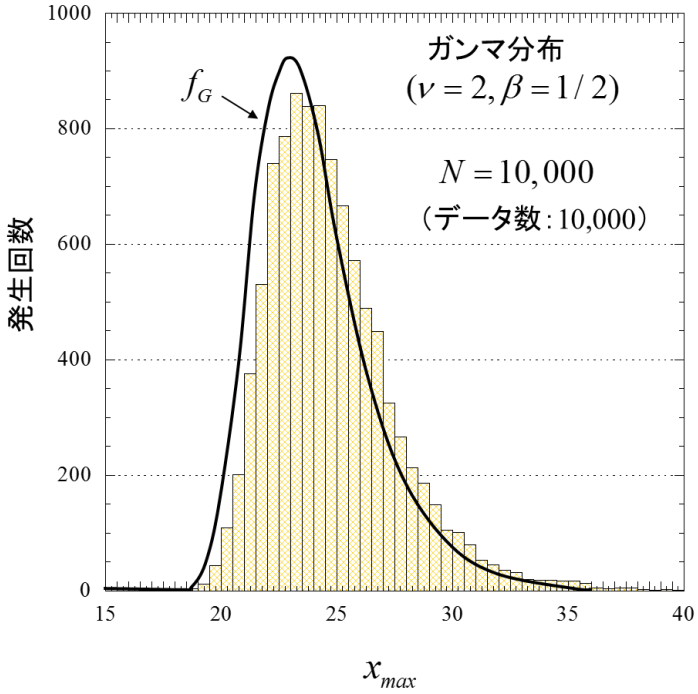


図２　最大値（規格化量）の漸近確率分布（Gumbel 分布型極値分布）



(a) *N*=100（Gumbel 分布からはかなり外れている）



(b) *N*=10,000（Gumbel 分布にだいぶ収斂してきている）

図３　ガンマ分布（ν=2, β=0.5）でのシミュレーションによる最大値 x_{max} のヒストグラム（データ数 10,000）と対応する Gumbel 型極値分布

３．最大値の推定

２節では、最大値の漸近確率分布の三つの形を述べたが、N 標本中の最大値 x_{max} の代表値を具体的に求めることには触れていない。確率密度関数 $f(x)$から極値分布 f_N を導き（すなわち、２節の方法を厳密に適用し）、そこから各種代表値（期待値、中央値、最頻値など）を求めることも可能である。しかし、それには、式の展開や数値計算に手間がかかり、大まかな x_{max} を簡単に推定したいと言う本レポートの目的には向いていない。ここでは、N 標本の中の最大値 x_{max} $(=x_{(N)})$の代表値を簡易に推定する以下の三つの方法をまとめる。

・最頻値（モード）の推定
・期待値の上限の推定
・累積分布関数の逆関数からの期待値の推定

以下、順次この方法を説明する。

（１）最頻値（モード）の推定

最大値 x_{max} の最頻値を求める方法である。最頻値はモードとも呼ばれ、x_{max} の確率密度関数のピーク値（最大値）を与えるものである。最頻値とするメリットは三点あり、一つは、解析的に求めやすいこと、二つ目は、平均値が存在しないような分布であっても求めることができる点、三つ目は、結果式が、元になる確率分布の極値付近のみの性質によって決まり、誤差を与えやすい確率分布全体の形を使わないことである。

最大値の累積分布関数とその確率密度関数は、式(3a, b)で与えられており、これを再掲する（ただし、変数 x を x_{max} で置き換えている）。

$$F_N(x_{max}) = F^N(x_{max}) \qquad \text{((3)の再掲)} \qquad (16a)$$

$$f_N(x_{max}) = NF^{N-1}(x_{max}) f(x_{max}) \qquad (16b)$$

最頻値は $df/dx_{max}=0$ を与える x_{max} であり、これを x_{mode} とすると、以下の関係になる。

$$(N-1) f^2(x_{mode}) + F(x_{mode}) f'(x_{mode}) = 0 \qquad (17)$$

$F(x_{mode})$は、(4)式より、$N/(N+1)$で近似でき、さらに N が大きい場合、途中計算に現れる $(N+1)(N-1)/N$ が N に近似できるので、上式は次式のように整理される。

$$Nf^2(x_{mode}) + f'(x_{mode}) = 0 \qquad (18)$$

これより、最大値の最頻値の推定値 x_{mode} は、N の関数として、次式で定められる。

$$N = \frac{-f'(x_{mode})}{f^2(x_{mode})} \equiv \Phi(x_{mode}) \;\rightarrow\; x_{mode} = \Phi^{-1}(N) \tag{19}$$

(18), (19)式の形を見ると、上に挙げた三つの利点のうち、特に、一点目と三点目が確認できるであろう。個別の確率分布に対しては、４節において、他の方法とも比較して、推定値の特徴を調べる。

（２）期待値の上限の推定

x_{max} の期待値<x_{max}>の上限を推定する理論がある。この導出の説明は割愛する（詳しくは[4],[6]）。結果のみ示すと

$$\langle x_{\max} \rangle \le \inf_{t>0} \left\{ \frac{1}{t} \log\left(NG(t) \right) \right\} \tag{20a}$$

$$G(t) \equiv \int_{-\infty}^{\infty} e^{tx} f(x) dx \tag{20b}$$

ここで、記号 inf は t を指定範囲で変化させたときの下限、$G(t)$は確率密度関数 $f(x)$の積率母関数（モーメント母関数）である。

具体的にこれで求めてみようとすると、積率母関数が簡単に求められるかどうか、また、t を変化させて下限値を求める演算などに煩雑さがある。また、推定は(20a)式のように不等式になっていて、値が特定されない。そのようにして求めた値もさらには、式(20b)から分かるように分布形全体を使って推定する方法なので、本来求めたい値に影響しない極値から離れた部分の分布の形まで、すなわち、分布全体の形が正確でないと推定値にその影響が現れる。故に、(1)と(2)の方法を比較した場合、(1)の方法に優位性があると考えてよいであろう。（本レポートでは、この方法にはこれ以上触れない）

（３）累積分布関数の逆関数からの期待値の推定

１節の(11)式で説明したように、N 標本での最大値の累積確率の期待値は $N/(N+1)$であった。ということは、この方法で求める最大値の代表値を x_{inv} とすると、累積分布関数 $F(x)$に対して、

$$F(x_{inv}) = \frac{N}{N+1} \quad \rightarrow x_{inv} = F^{-1}\left(\frac{N}{N+1} \right) \tag{21}$$

として求められる。あるいは、

$$N = \frac{F(x_{inv})}{1 - F(x_{inv})} \equiv \Psi(x_{inv}) \quad \rightarrow \quad x_{inv} = \Psi^{-1}\left(N \right) \tag{22}$$

のような形に整理でき、x_{inv} と N が、より直接的に関係付けられる。

この方法は、累積分布関数の逆関数 F^{-1} が式として表されていれば、代入するだけなので、簡単に求められ、非常に便利な方法である。確率密度関数やその累積分布関数は、閉形式で表現できるものが多いが、その逆関数は、一般的には閉形式にならない分布形が多い。その場合でも、(22)式のように x_{inv} と N の関係式に整理すれば、数値計算による対応付けは可能である。

４．各種確率分布の最大値

３節で述べた三つの方法のうち、具体的な推定が容易な（１）最頻値推定と（３）累積分布関数の逆関数からの期待値の推定の二つの方法について、計算結果を、乱数を用いたシミュレーション結果と比較し、推定法の特徴や N と最大値との関係を調べる。

理論計算においては、最頻値推定では、(19)式より、$f(x)$ と $f'(x)$ を用い、また、逆関数からの期待値推定では、(22)式より $F(x)$ を用いて求める。確率分布は以下の５つを取り上げる。

・標準正規分布
・レイリー分布
・ガンマ分布
・対数正規分布
・標準コーシー分布

コーシー分布を除く４つの分布については、その物理的意味を[7], [8]にまとめている。

理論推定値の確認として、乱数を用いたシミュレーションを行う。指定の確率分布に従う N_{data}=1,000,000 個の実現値を用い、N=10, 100, 1,000, 10,000, 100,000 での最大値を調べる。それぞれの N に対する最大値（x_{max}）の数は N_{data}/N である。このデータから、コーシー分布以外の分布に対しては算術平均 $\overline{x}_{max}$ と標準偏差を求める。対数正規分布では相乗平均も算出する。また、コーシー分布では平均値が意味を持たないので、平均値の代わりに中央値（メディアン）を求める。

（１）標準正規分布

正規分布は、中心極限定理に見られるように、観測値が種々の要因の和で表されるような確率過程（加法性確率過程）に現れ、最も基本となる確率分布である。その中でも、平均値 0、分散 1 の正規分布は標準正規分布（$N(0,1)$で表記される）と呼ばれ、基本中の基本である。標準正規分布の確率密度関数 f、累積分布関数 F、確率密度関数の導関数 f'はそれぞれ以下で表される。

$$f(x)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right) \tag{23a}$$

$$F(x)=\frac{1}{2}\left\{1+\mathrm{erf}\left(\frac{x}{\sqrt{2}}\right)\right\} \quad \left(\mathrm{erf}(x)\equiv\frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt\right) \tag{23b}$$

$$f'(x)=\frac{-x}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right) \tag{23c}$$

(23)式と、(19), (22)式より、x_{mode}, x_{inv} と N の関係が得られる。x_{mode} については、

$$N=\sqrt{2\pi}x_{mode}\exp\left(\frac{x_{mode}^2}{2}\right) \tag{24a}$$

であるので、これより、

$$\ln N=\ln\left(\sqrt{2\pi}x_{mode}\right)+x_{mode}^2/2\approx x_{mode}^2/2 \ \ (N>>1) \tag{24b}$$

$$\rightarrow \quad x_{mode}\approx\sqrt{2\ln N} \quad (N>> \tag{24c}$$

となり、最大値 x_{mode} は N の値が十分大きくなると x_{mode} も大きな値になるので、(24b)式右辺のように近似でき、$\sqrt{\ln N}$ に比例することが分かる。この性質は、正規分布ばかりでなく、分布に $\exp(-x^2)$が積の形で入る分布、すなわち、仲上m分布系（$x^a\exp(-bx^2)$の形[7],[8]）に共通である。（なお、(24c)式の近時の精度については、ガンマ分布の項で、もう少し定量的に議論している。）

図４は推定値 x_{mode}, x_{inv} を N の関数で示している。同図より、最頻値（x_{mode}）と逆関数による推定値（x_{inv}）は極めてよく一致していることがわかる。シミュレーション値もこの結果をよくフォローしている。シミュレーション値（平均値）が二つの理論推定値に比べてやや大きくなっているのは、図２の分布形に見られるように、最頻値が小さい側に寄っている非対称性による。しかしその差は小さいので、理論推定値を用いることで十分であろう。正規分布はいわゆる裾の軽い分布の代表であるため、N の増加に対して最大値の増加は極めて緩慢である。

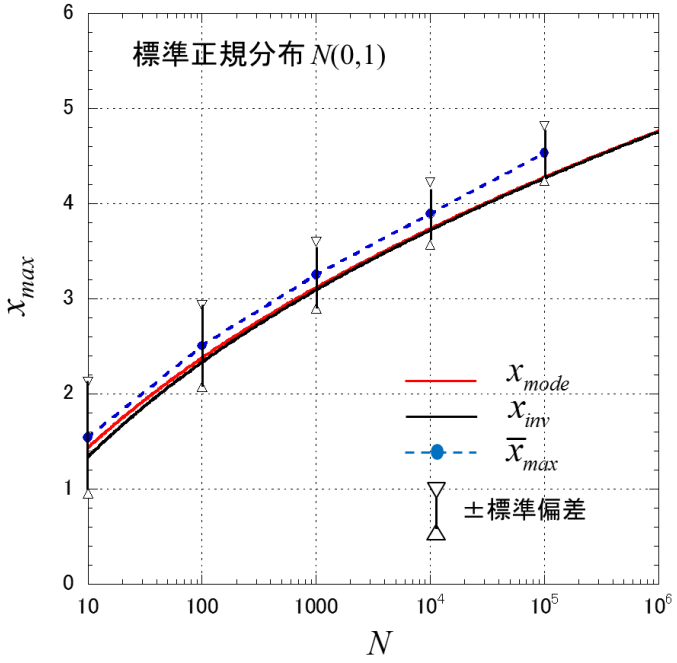


図４　N 標本の最大値の推定：標準正規分布（$N(0,1)$）

（２）レイリー分布

レイリー分布は、電波伝搬分野ではマルチパスフェージングにおける振幅変動の確率分布が代表的である。レイリー分布の確率密度関数 f、累積分布関数 F、確率密度関数の導関数 f'は、一つのパラメータσを用いて、それぞれ以下で表される。

$$f(x)=\frac{x}{\sigma^2}\exp\left(-\frac{x^2}{2\sigma^2}\right) \tag{24a}$$

$$F(x)=1-\exp\left(-\frac{x^2}{2\sigma^2}\right) \tag{24b}$$

$$f'(x)=\frac{1}{\sigma^2}\left(1-\frac{x^2}{\sigma^2}\right)\exp\left(-\frac{x^2}{2\sigma^2}\right) \tag{24c}$$

x_{inv} は閉形式の解が得られ次式となる。

$$x_{inv}=\sqrt{2\ln(N+1)}\,\sigma\approx\sqrt{2\ln N}\,\sigma\ (N>>1) \tag{25}$$

図５はσ=1 として求めた推定値 x_{mode}, x_{inv} を N の関数で示している。計算機シミュレーション結果も示している。結論としては、標準正規分布で述べたことと同じである。レイリー分布も仲上 m 分布に含まれる分布であることから、前述の結論と同じになるのは当然であろう。

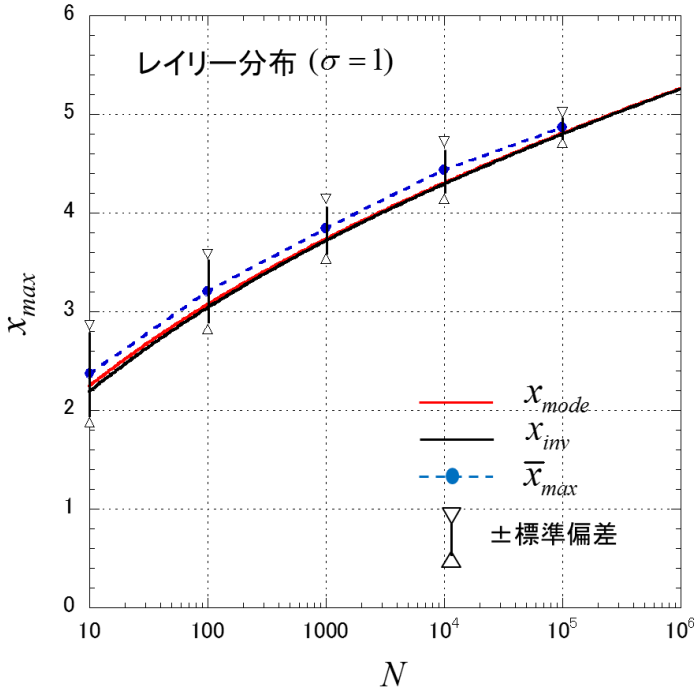


図５　N 標本の最大値の推定：レイリー分布（σ=1）

（３）ガンマ分布

ガンマ分布は、無線通信においては、ダイバーシチ合成後の電力分布（あるいは SN 比の分布）、あるいは、降雨強度や降雨減衰の分布など、種々の場面に現れる。二つのパラメータ（ν: 形状パラメータ、β: スケールパラメータ（正確にはθ=1/βと置いたθが）で表されるため、種々の現象の確率分布に対して柔軟性がある。ガンマ分布の確率密度関数 f、累積分布関数 F、確率密度関数の導関数 f'は、それぞれ以下で表される。

$$f(x)=\frac{1}{\Gamma(\nu)}\beta^{\nu}x^{\nu-1}\exp(-\beta x) \tag{26a}$$

$$F(x)=\gamma(\nu,\beta x)/\Gamma(\nu) \tag{26b}$$

where $\gamma(\alpha, x) \equiv \int_0^x t^{\alpha-1} e^{-t} dt$ （第一種不完全ガンマ関数）

$$f'(x) = \frac{\beta^\nu}{\Gamma(\nu)} (\nu - 1 - \beta x) x^{\nu-2} \exp(-\beta x) \tag{26c}$$

x_{mode} と N の関係では、(19)式より、

$$N = \Gamma(\nu)(\beta x_{mode} - \nu + 1)(\beta x_{mode}) \exp(\beta x_{mode}) \tag{27a}$$

となり、N>>1 では、x_{mode} も大きな値になるので、

$$\ln N = \ln\left\{\Gamma(\nu)\beta^{-\nu}(\beta x_{mode} - \nu + 1)\right\} - \nu \ln x_{mode} + \beta x_{mode} \tag{27b}$$

$$\approx \beta x_{mode} \quad (N >> 1) \tag{27c}$$

$$\rightarrow \; x_{mode} \approx (1/\beta) \ln N \quad (N >> 1) \tag{27d}$$

となって、$\ln N$ に比例する。ただし、この近似は非常に大雑把なものであり、N の全範囲を通して、無視している項（式(27b)の右辺第１項と第２項）の分は $\ln N$ の誤差として固定的に残る。図６は ν=2, β=0.5 の場合の式(27b, c)の違い（黒線：式(27b)、赤線：式(27c)）を示している。N が非常に大きい値（図では 10^{10} 以上）のときの

$$\beta = \lim_{N \to \infty} \frac{d \ln N}{d x_{mode}} \tag{27e}$$

を表していると考えておいてほしい。なお、式(27b, c)の違いの大小はパラメータ ν, β の値に依存するので、式(27d)の近似式を使いたい場合には、その都度、精度の確認が必要である。

最大値の評価では、図６でも示した ν =2, β =0.5 に設定する。これは独立な $N(0,1)$の４つの変数のそれぞれの二乗和の分布（自由度 4 の χ^2（カイ二乗）分布に同じ）であり、シミュレーションでもそのようにしてデータ生成を行う。

図７は ν=2, β=0.5 として求めた推定値 x_{mode}, x_{inv} を N の関数で示している。図より、二つの理論カーブ（x_{mode} と x_{inv}）はぴったりと重なっていること、最大値は $\ln N$ と線形関係であることがわかる。

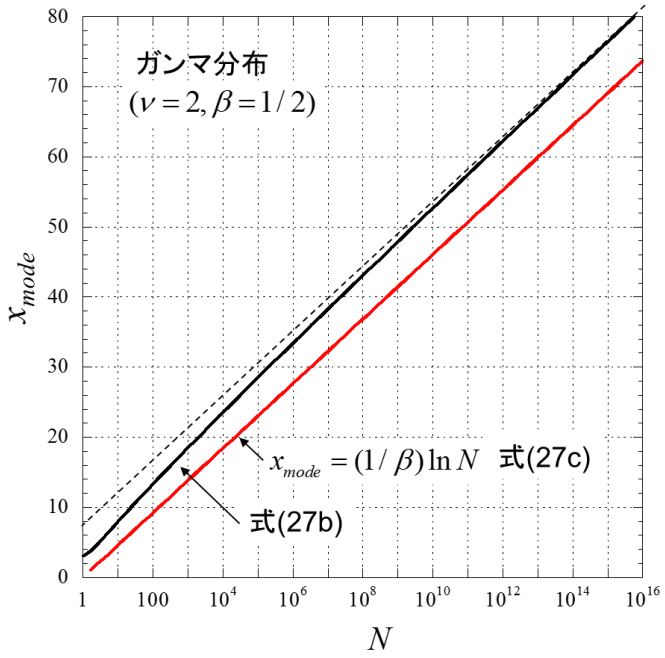


図６　ガンマ分布（ν=2, β=0.5）での、式(27b)と(27c)の計算結果の比較

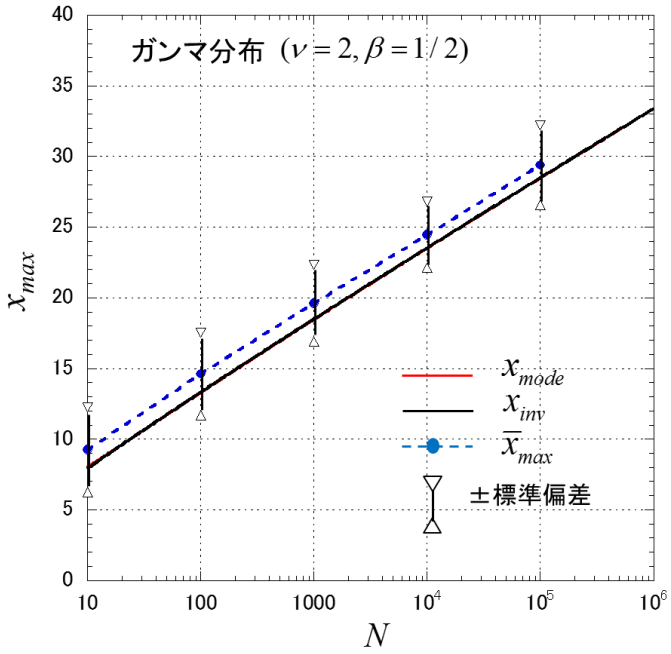


図７　N 標本の最大値の推定：ガンマ分布（ν=2, β=0.5）（x_{mode} と x_{inv} は重なっている）

（４）対数正規分布

対数正規分布は確率変数の対数値が正規分布する分布である。多数の掛け算の結果として表されるような確率過程（乗法性確率過程）に対数正規分布が現れる。変動幅の目安を倍半分のような言い方で表す物理量である。対数正規分布の確率密度関数 f、累積分布関数 F、確率密度関数の導関数 f'は、二つのパラメータ m, σ（m: $\ln x$ の平均、σ:標準偏差）を用いて、それぞれ以下で表される

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma x}\exp\left\{-\frac{(\ln x-m)^2}{2\sigma^2}\right\} \tag{28a}$$

$$F(x)=\frac{1}{2}\left\{1+\mathrm{erf}\left(\frac{\ln x-m}{\sqrt{2}\sigma}\right)\right\} \tag{28b}$$

$$f'(x)=\frac{-1}{\sqrt{2\pi}\sigma x^2}\left(1+\frac{\ln x-m}{\sigma^2}\right)\exp\left\{-\frac{(\ln x-m)^2}{2\sigma^2}\right\} \tag{28c}$$

x_{mode} と N の関係では、(19)式より、

$$N=\sqrt{2\pi}\sigma\left(1+\frac{\ln x_{mode}-m}{\sigma^2}\right)\exp\left(\frac{\left(\ln x_{mode}-m\right)^2}{2\sigma^2}\right) \tag{29a}$$

となり、N>>1 では、

$$\ln N=\ln\left\{\sqrt{2\pi}\sigma\left(1+\frac{\ln x_{mode}-m}{\sigma^2}\right)\right\}+\frac{\left(\ln x_{mode}-m\right)^2}{2\sigma^2} \tag{29b}$$

$$\approx\frac{\left(\ln x_{mode}-m\right)^2}{2\sigma^2}\quad(N>>1) \tag{29c}$$

$$\rightarrow\ x_{mode}\approx\exp\sqrt{2\sigma^2\ln N}\quad(N>>1,\ m=0) \tag{29d}$$

となって、$\exp\sqrt{2\sigma^2\ln N}$ に比例する。

評価では、m =0, σ =1 とする。シミュレーションでは、N(0,1)のデータ n_i を exp(n_i)に変換して生成した。

図８は m =0, σ =1 として求めた推定値 x_{mode}, x_{inv} を N の関数で示している。シミュレーションでは、x_{max} の算術平均と相乗平均の両方を示しているが両者の違いは小さい。対数正規分布の極値分布も Gumbel 型になるが、その分布の非対称性の強さから、理論推定値（x_{mode}, x_{inv}）とシミュレーション値（平均値）の差が大きくなっている。

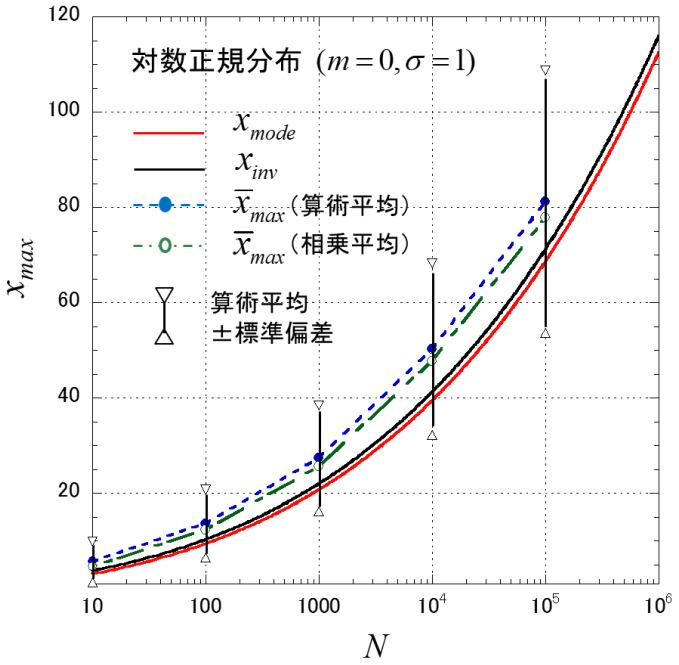


図８　N 標本の最大値の推定：対数正規分布（m =0, σ=1）

（５）標準コーシー分布

水平面上に棒を立てる。東の水平線から現れた太陽が天頂を通って西に沈む。このときの影の長さの確率分布がコーシー分布である。日の出と日没付近で、影の長さは無限大に向かって極端に伸びる。コーシー分布は確率分布を扱う専門書では平均や分散が存在しない分布の例としてよく取り上げられる。影の例をイメージすれば、平均値が意味を成さないことは感覚的にも分かる。無線通信に関する諸現象の中で物理量がコーシー分布となる現象を筆者は知らないが、極端な性質を持つ分布として、ここでは取り上げる。

標準コーシー分布の確率密度関数 f、累積分布関数 F、確率密度関数の導関数 f'は、それぞれ以下で表される

$$f(x)=\frac{1}{\pi(1+x^2)} \tag{30a}$$

$$F(x)=\frac{1}{\pi}\arctan x \tag{30b}$$

$$f'(x) = -\frac{2x}{\pi(1+x^2)^2} \tag{30c}$$

標準コーシー分布の x_{mode} と x_{inv} は(19), (22)式より解析的に求められて次式となる。

$$x_{mode} = \frac{N}{2\pi} \tag{31}$$

$$x_{inv} = \tan\left(\frac{\pi}{2} - \frac{\pi}{N+1}\right) \approx \tan\left(\frac{\pi}{2} - \frac{\pi}{N}\right) \quad (N >> 1) \tag{32}$$

なお、標準コーシー分布の乱数は、影の長さの例を使い、標準一様乱数（0～1 区間）u_i を用いて、$x_i = \tan(\pi u_i)$により生成する。

図 9 は標準コーシー分布での推定値 x_{mode}, x_{inv} を N の関数で示している。また、計算機シミュレーション結果では、平均値が意味を持たないため、最大値の中央値を示す。

コーシー分布では x_{max} は N に比例して増加すると言う異常性（恐ろしいことがおきやすい）を示していて、極値を議論できるような分布ではない。それでも、シミュレーションでの中央値は x_{inv} に近い値を示している。

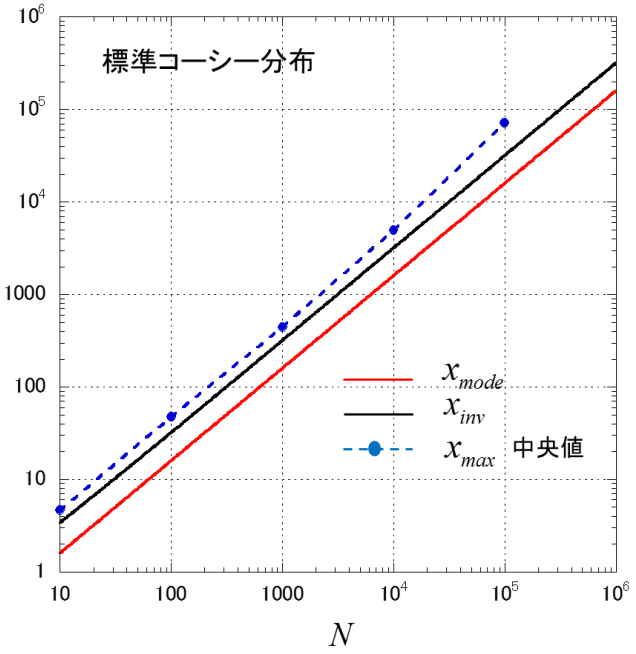


図 9　N 標本の最大値の推定：コーシー分布

比較結果のまとめ

代表的な５つの分布で調べた結果を以下にまとめる。

i) N の増加に対する最大値 x_{max} の増加の割合は、正規分布・レイリー分布（仲上 m 分布系）では $\sqrt{\log N}$ のオーダ、ガンマ分布では $\log N$、対数正規分布では $\log N$ より強く、コーシー分布では N に比例する。ゆえに、最大値の N に対する伸び率の大きさは、

コーシー分布（N）>> 対数正規分布（$\exp\sqrt{\alpha \ln N}$ ） >ガンマ分布（$\log N$）>仲上 m 分布系（$\sqrt{\log N}$ ）

の順である

ii) コーシー分布を除く他の分布では、最大値の理論推定値 x_{mode} と x_{inv} は非常に近い値になる。分布の式を見て求めやすいほうを使えばよい。

iii) 最大値の平均値の理論推定値を求める方法は複雑で、具体的な値を得ることが困難である（３節）。ゆえに､計算機シミュレーションでその値を求めたが、どの分布でも、x_{mode} や x_{inv} に比べてやや多き目の値になった。これは、極値分布の形状が、最頻値が平均より低いほうに現れる非対称の形になっていることによる。

コーシー分布で支配される世界では､その最大値は観測数に比例するものとなるため､極めて恐ろしいこと（=かつて経験したことが無いようなこと）が当たり前のように起きる。コーシー分布は平均値や分散が定まらない分布の代表として、確率分布の教科書ではよく取り上げられるが、幸い、筆者は、日常世界に現れる物理量（自然災害などの）としてそれを見たことが無い。そういう意味では､その世界には縁が無いと言うことで安心してよさそうである。我々の分野では、対数正規分布に支配される世界に危なさが残り、自然災害の要因になる種々の物理量のマグニチュード（単位が対数）は、これに相当するであろう。

５．さらに理解を深めるために

極値統計学の大筋を駆け足のようにまとめた。４節で提示している最大値の推定結果やシミュレーション結果は、類書にもあまり無いようなので(*)、参考になると思う。[*: [4]では、同様の結果がまとめられている（と言うか、筆者がこのまとめ方を参考にさせてもらっている）が、取り上げている分布が一部違っており、かつ、同じものについての示されている結果も幾分違うが、本資料の結果図は、この範囲で正しいと確信する]。

極値（最大値）の推定において、裾の軽い分布の代表である正規分布（仲上 m 分布系）では､データが増えることによって現れる最大値の大きさは抑制的であるが､裾の重い分布（例えば対数正規分布）では、大きい値が現れやすい。すなわち冒頭に述べたような想定外のことが起きる危険が高い。しかしこのような想定外のことでも､極値統計学の壺を押さえておけば､長い目で見れば想定内であると受け止めることができ､その予防対策や気持ちの安心に活かすことができる。

ここでは､一つの確率分布内の現象として扱ってきたので､何が起きても想定内という理屈に行き着くが､現実の自然災害などではどうであろうか？確率分布そのもの（パラメータ

値であったり分布形状であったり）が時間と共に変化してゆくような場合には、もう一段上の理論が必要になる。地球温暖化問題と異常気象の関係などはこれに当たるであろう。発生数が少なすぎて統計的性質が見えないものに対しての予測も深い考察が必要になる。これらの検討のためにも、本資料を手がかりにして極値統計学の基礎を是非学んでほしい。

このレポートを作成するに際して筆者が参考にした文献（本文中でも引用）を以下に挙げる。

参考文献

[1] E.J. *Gumbel, Statistics of Extremes*, Univ. Press, New York, 1958; 翻訳版：河田竜夫, 岩井茂久, 加瀬滋男（監訳）, *極値統計学*, 生産技術センター新社（再販版）.

[2] 高橋倫也, 志村隆彰, *極値統計学*, IMS シリーズ：進化する統計数理５, 近代科学社, 2016.

[3] 唐沢好男, “回帰分析と信頼区間,” Tech. Rep. YK-019（私報）, 2019.01., http://www.radio3.ee.uec.ac.jp/ronbun/YK-019_Kukan_Suitei.pdf

[4] 蒼馬竜, *極値統計によろしく*, （同人集団）暗黒通信団, 2018.

[5] 順序統計量, ウィキペディア, https://ja.wikipedia.org/wiki/%E9%A0%86%E5%BA%8F%E7%B5%B1%E8%A8%88%E9%87%8F

[6] 最大値の期待値・裾確率の不等式の応用例と証明；数理工学のススメ https://math-eng.com/maximum-expectation-tail-inequality

[7] 唐沢好男, “伝搬モデルに現れる確率分布,” Tech. Rep. YK-005, 2017.12. http://www.radio3.ee.uec.ac.jp/ronbun/TR_YK_005_Probability_distribution.pdf

[8] 唐沢好男, *改訂 ディジタル移動通信の電波伝搬基礎*, コロナ社, 2016.