統計数理（2004）
第 52 巻 第 1 号 93–116

特集「極値理論」
［原著論文］

上位 r 個の観測値に基づく確率点の推定

高橋 倫也[1] ・渋谷 政昭[2]

（受付 2003 年 7 月 31 日；改訂 2004 年 1 月 27 日）

要　　旨

いくつかの単位領域（または単位期間）のそれぞれで測定したデータを利用して，上側微小確率点を推定する．単位領域ごとの最大値データのみを用いるのが古典的極値解析の手法であった．ここでは，最大値だけでなく上位 r 個のデータを用いることにより，推定の精度がどの程度改善されるか漸近相対効率を用いて明らかにする．また，この手法の実用上の問題点である r の決定について議論する．例として腐食孔の深さを推定する．

キーワード：漸近相対効率，漸近分散，一般極値分布，Gumbel 分布．

1. はじめに

与えられた領域（または期間）での最大値を推定するために，いくつかの単位領域（または単位期間）ごとの最大値（極値）データを利用するのが古典的極値解析の手法である．その推定精度を上げる方法として「上位 r 個のデータ」または「ある閾値以上の全てのデータ」を用いることが提案されている．ここでは前者の場合について，r を増やすと推定精度がどの程度改善するか漸近相対効率を計算する．また，実用上の問題点である r の決定について議論する．

上位 r 個のデータを用いるべき場合として，各年のすべてのまたは十分大きな閾値以上のデータが独立同一分布の条件を満たしているとは見なせないが，上位の何個かのデータを取り出す限りにおいてはそれらは独立同一分布からのものと見なせる場合（Tawn（1988）参照），スポーツや自然災害のデータで各年の最高や最悪の何個かの記録しか残っていない場合等が考えられる．

Weissman（1978）は極値理論に基づく上位 r 個のデータ（一組）を用いる上側微小確率点の推定を初めて議論した．彼は母集団分布が Gumbel 分布の吸引領域に属す場合（Gumbel モデル）を詳しく調べた．Smith（1986）は Gumbel モデルの下で上位 r 個のデータを N 組用いる場合の推定問題を議論した．彼は，r の決定法を提案し，情報量を計算し，位置パラメータに時間依存性を導入してベニスの毎年上位 10 個の潮位データを解析し，この手法の有効性を示した．Dupuis（1997）は Smith（1986）のモデルとデータの下でロバスト推定法について議論している．Tawn（1988）は，Smith（1986）の結果を一般極値分布（GEV モデル）へ拡張し，Lowestoft の潮位データ等の解析を行った．Tsimplis and Blackman（1997）でも潮位データの解析にこの手法を用いている．

この手法の他の分野への応用には次のような研究がある．Robinson and Tawn（1995）と Smith（1997）は 1993 年女子 3000 m で驚異的な記録を出した王軍霞（Wang Junxia）が薬物を使用し

[1] 神戸商船大学 商船学部（現 神戸大学 海事科学部）：〒658–0022 兵庫県神戸市東灘区深江南町 5–1–1
[2] 高千穂大学 経営学部：〒168–8508 東京都杉並区大宮 2–19–1

た可能性があるか，すなわち王の記録はそれまでの競技者の能力をはるかに超えたものかどうかの議論をしている．Robinson and Tawn (1995)は，それまでの各年の上位 5 位の記録を指数的に減少する傾向をもった GEV モデルの下で調べ，さらに 1500 m の上位 5 位の記録(3000 m の記録に含まれない選手に限る)との相対関係を考慮するモデルまでも利用したが，王のデータは 90% 信頼区間におさまり，かなり稀少なデータだが否認できないという結論を述べた．Smith (1997)はこの結論を批判し，非有限事前分布を用い，推測でなく予測を行い，王のデータは明らかに外れ値であると断定した．Strand and Boes (1998)はロードレースに参加した各年齢別の上位 5 位のデータ解析を Gumbel モデルの下で行っている．腐食の分野では，Scarf et al. (1992)は GEV モデルで位置と尺度パラメータに時間依存性を導入して腐食孔の深さデータの解析を行った．また，総合報告の Scarf and Laycock (1996)でも手法が紹介されている．風速に関しては Coles and Walshaw (1994)と総合報告の Palutikof et al. (1999)がある．

上位 r 個の漸近分布に関しては，Nagaraja (1982)と Scarf (1993)が詳しい．

以下，2 節で極値理論から導かれる上位 r 個の順序統計量の漸近同時分布を示しその性質についてまとめる．3 節でパラメータの推定法と r の決定法について述べ，4 節で上位 r 個を用いる有効性について議論する．5 節で実データの解析例を示す．付録で上位 r 個の漸近同時分布の情報量と確率加重モーメント(PWM)法について述べる．

2. 極値理論

一般極値 (generalized extreme value)分布の標準型を

$$G_\xi(z) = \exp[-(1+\xi z)^{-1/\xi}], \quad 1+\xi z > 0 \ \ (\xi \in \boldsymbol{R}) \tag{2.1}$$

とする．ここで G_ξ は，$\xi < 0$ の場合は(負の)Weibull 分布，$\xi = 0$ の場合は $G_0(z) = \lim_{\xi\to 0} G_\xi(z) = \exp(-e^{-z})$ で Gumbel 分布，$\xi > 0$ の場合は Fréchet 分布である．

分布 F からの確率標本 $Y_1, Y_2, \ldots, Y_n$ の順序統計量を $Y_{1:n} \geq Y_{2:n} \geq \cdots \geq Y_{n:n}$ とし，分布 F が一般極値分布 G_ξ の吸引領域に属すと仮定する：すなわち 適当な数列 $a_n > 0$，$b_n \in \boldsymbol{R}$，$n = 1, 2, \ldots$ が存在し

$$\lim_{n\to\infty} P\left(\frac{Y_{1:n} - b_n}{a_n} \leq z\right) = G_\xi(z), \quad \forall z \in \boldsymbol{R}.$$

このとき，上位 r 個の順序統計量の同時分布関数

$$P\left(\frac{Y_{1:n} - b_n}{a_n} \leq z_1, \frac{Y_{2:n} - b_n}{a_n} \leq z_2, \ldots, \frac{Y_{r:n} - b_n}{a_n} \leq z_r\right), \quad z_1 \geq z_2 \geq \cdots \geq z_r$$

は同時密度関数

$$g_{\xi,\,12\cdots r}(z_1,\, z_2, \ldots,\, z_r) = \frac{g_\xi(z_1)\cdots g_\xi(z_{r-1})}{G_\xi(z_1)\cdots G_\xi(z_{r-1})} g_\xi(z_r), \quad g_\xi(z) = dG_\xi(z)/dz$$

を持つ分布関数 $G_{\xi,\,12\cdots r}$ に収束する．David and Nagaraja (2003)の 10.6 節 を参照．

ここで，

$$g_{\xi,\,12\cdots r}(z_1,\, z_2, \ldots,\, z_r) = \begin{cases} \exp\left(-\sum_{j=1}^{r} z_j - e^{-z_r}\right), & \xi = 0 \\ \left(\prod_{j=1}^{r}(1+\xi z_j)^{-1/\xi-1}\right)\exp[-(1+\xi z_r)^{-1/\xi}], & \xi \neq 0 \end{cases} \tag{2.2}$$

である．

確率ベクトル $(Z_1,\, Z_2, \ldots, Z_r)$ **が分布** $G_{\xi,\,12\cdots r}$ **に従うとき，** Z_j**，** $j \geq 1$ **の周辺分布関数** $G_{\xi,j}$ **は** r **によらず**

$$G_{\xi,j}(z) = \begin{cases} \displaystyle\sum_{k=0}^{j-1} \exp(-kz - e^{-z})/k!\,, & \xi = 0 \\ \displaystyle\sum_{k=0}^{j-1} (1+\xi z)^{-k/\xi} \exp[-(1+\xi z)^{-1/\xi}]/k!\,, & \xi \neq 0 \end{cases} \tag{2.3}$$

その周辺密度関数 $g_{\xi,j}$ **は**

$$g_{\xi,j}(z) = \begin{cases} \exp(-jz - e^{-z})/\Gamma(j)\,, & \xi = 0 \\ (1+\xi z)^{-j/\xi-1} \exp[-(1+\xi z)^{-1/\xi}]/\Gamma(j)\,, & \xi \neq 0 \end{cases} \tag{2.4}$$

となる．

定理 1． $(Z_1,\, Z_2, \ldots, Z_r)$ **は次の性質をもつ．ただし，** $W_1, W_2, \ldots$ **は標準指数分布(**Exp **(**1**))** **に従う独立確率変数で** $S_j = \sum_{k=1}^{j} W_k$ **とし，** Γ **はガンマ関数，** ψ **はディ・ガンマ関数，** ψ' **はトリ・ガンマ関数とする．**

(I**)** Gumbel $(\xi = 0)$ **モデルの場合：**

(I.a**)** $$\{Z_j,\ j \geq 1\} \stackrel{d}{=} \{-\log S_j,\ j \geq 1\}\,.$$

(I.b**)** $(Z_1,\, Z_2, \ldots,\, Z_r)$ **は** Exp **(**1**)からの上位** r **個の順序統計量と見なせ，**

$$j(Z_j - Z_{j+1})\,, \quad j = 1, 2, 3, \ldots$$

は互いに独立に Exp(1) **に従う．**

(I.c**)** $$E_0(Z_j) = -\psi(j)\,, \quad V_0(Z_j) = \psi'(j)\,.$$

(I.d**)** $$\mathrm{Cor}_0(Z_j,\, Z_{j+1}) = \sqrt{\psi'(j+1)/\psi'(j)}\,.$$

(II**)** GEV $(\xi \neq 0)$ **モデルの場合：**

(II.a**)** $$\{Z_j,\ j \geq 1\} \stackrel{d}{=} \{(S_j^{-\xi} - 1)/\xi,\ j \geq 1\}.$$

(II.b**)** $(Z_1,\, Z_2, \ldots,\, Z_r)$ **は 形状パラメータ** ξ **の一般** Pareto **分布からの上位** r **個の順序統計量と見なせ，**

$$\frac{j}{\xi} \log \frac{1+\xi Z_j}{1+\xi Z_{j+1}}\,, \quad j = 1, 2, 3, \ldots$$

は互いに独立に Exp(1) **に従う．**

(II.c**)** $$E_\xi(Z_j) = \frac{\Gamma(j-\xi) - \Gamma(j)}{\xi\Gamma(j)}\,, \quad V_\xi(Z_j) = \frac{\Gamma(j-2\xi)\Gamma(j) - \Gamma^2(j-\xi)}{\xi^2\Gamma^2(j)}\,.$$

(II.d**)** $$\mathrm{Cor}_\xi(Z_j,\, Z_{j+1}) = \frac{1}{j-\xi} \sqrt{\frac{\Gamma(j+1-2\xi)\Gamma(j+1) - \Gamma^2(j+1-\xi)}{\Gamma(j-2\xi)\Gamma(j) - \Gamma^2(j-\xi)}}\,.$$

証明． **(**I**)** **(**I.a**)** $(-\log S_1, \ldots, -\log S_r)$ **は** $(W_1, \ldots, W_r)$ **を変換して得られる事に注意して，その同時密度関数を求めると** $g_{0,12\cdots r}$ **に一致する．**

(I.b) $(Z_1 - Z_2,\, 2(Z_2 - Z_3), \ldots, (r-1)(Z_{r-1} - Z_r))$ の同時密度関数を計算すると

$$e^{-y_1} e^{-y_2} \cdots e^{-y_{r-1}}, \quad y_1, y_2, \ldots, y_{r-1} > 0$$

となり，$j(Z_j - Z_{j-1})$ は互いに独立に Exp(1)に従う．

(I.c) 式 (2.4)から求まる．

(I.d) $(Z_j,\, Z_{j+1})$ の同時密度関数は

$$g_{0,jj+1}(z_j,\, z_{j+1}) = \frac{1}{\Gamma(j)} \exp(-jz_j) g_0(z_{j+1}), \quad z_j \geq z_{j+1}$$

となる．これから

$$E_0(Z_j Z_{j+1}) = \frac{1}{j^2 \Gamma(j)} (j\Gamma''(j+1) - \Gamma'(j+1)) = \psi^2(j+1) + \psi'(j+1) - \frac{1}{i}\psi(j+1)$$

を求め，(I.c)を用いればよい．

(II) 上の (I) の証明と同様の方法で証明ができる．ここでは，(I) の結果を用いた証明を示す．

(II.a) 次の同時分布関数を考える：

$$P\left(\frac{S_j^{-\xi} - 1}{\xi} \leq z_j, \quad j = 1, 2, \ldots, r \right) = P(-\log S_j \leq \log(1 + \xi z_j)^{1/\xi},\ j = 1, 2, \ldots, r).$$

ここで，(I.a)から左辺の同時密度関数は

$$\left(\prod_{j=1}^{r} (1 + \xi z_j)^{-1} \right) g_{0,12\cdots r}(\log(1 + \xi z_1)^{1/\xi}, \ldots, \log(1 + \xi z_r)^{1/\xi})$$

となり，これは $g_{\xi,12\cdots r}(z_1, \ldots, z_r)$ に一致する．

(II.b) 形状パラメータ ξ の標準一般 Pareto 分布

$$P(y) = 1 - (1 + \xi y)^{-1/\xi}, \quad 1 + \xi y > 0$$

($\xi \geq 0$ のとき $y > 0$，$\xi < 0$ のとき $0 < y < -1/\xi$)からの n 個の順序統計量を

$$V_{1:n} \geq V_{2:n} \geq \cdots \geq V_{n:n}$$

とすると，

$$V_{j:n} \stackrel{d}{=} \frac{U_{n-j+1:n}^{-\xi} - 1}{\xi}, \quad j = 1, 2, \ldots, n$$

と表される．ただし，$1 > U_{1:n} \geq U_{2:n} \geq \cdots \geq U_{n:n} > 0$ は一様分布 U(0, 1) からの n 個の順序統計量である．この $V_{j:n}$ の表現と (II.a)から $(Z_1,\, Z_2, \ldots,\, Z_r)$ は形状パラメータ ξ の一般 Pareto 分布からの上位 r 個の順序統計量と見なせる．

(I.a)と (I.b)から，$j \log(S_{j+1}/S_j)$ は互いに独立に Exp(1)に従う．一方，(II.a)から

$$\frac{j}{\xi} \log \frac{1 + \xi Z_j}{1 + \xi Z_{j+1}} \stackrel{d}{=} j \log \frac{S_{j+1}}{S_j}, \quad j = 1, 2, 3, \ldots$$

である．

(II.c) 式 (2.4)から求まる．

(II.d) $(Z_j,\, Z_{j+1})$ の同時密度関数は

$$g_{\xi,jj+1}(z_j,\, z_{j+1}) = \frac{1}{\Gamma(j)} (1 + \xi z_j)^{-j/\xi - 1} g_\xi(z_{j+1}), \quad z_j \geq z_{j+1}$$

となる．これから

$$E_\xi(Z_jZ_{j+1}) = \frac{1}{\Gamma(j+1)}\left\{\frac{1}{\xi^2}[\Gamma(j+1-2\xi) - 2\,\Gamma(j+1-\xi) + \Gamma(j+1)]\right.$$
$$\left. + \frac{1}{\xi(j-\xi)}[\Gamma(j+1-2\xi) - \Gamma(j+1-\xi)]\right\}$$

を求め，(II.c)を用いればよい． □

注 1. (I.a)と(II.a)は Nagaraja (1982)で，(I.b)は Weissman (1978)で，(II.b)は Tawn (1988)で示された．上の証明は彼らのものとは異なる．

注 2. (連続性)(I) の結果は(II) で $\xi \to 0$ とすれば得られる．

注 3. $\mathrm{Cor}_0(Z_j, Z_{j+1})$ は j に関して狭義単調増加関数で 1 に収束する．例えば，$\mathrm{Cor}_0(Z_1,\, Z_2) = 0.626$，$\mathrm{Cor}_0(Z_2,\, Z_3) = 0.783$，$\mathrm{Cor}_0(Z_3,\, Z_4) = 0.848$ である．$\xi \neq 0$ の場合も同様のことが成立すると思うが証明は出来ていない．

注 4. ディ・ガンマ関数とトリ・ガンマ関数の値は，変数が正整数の場合，

$$\psi(n) = -\gamma + \sum_{i=1}^{n-1}\frac{1}{i}\,, \quad \psi'(n) = \frac{\pi^2}{6} - \sum_{i=1}^{n-1}\frac{1}{i^2}, \quad n = 1, 2, \ldots$$

から求まる．ただし，$\gamma = 0.57721566...$ は Euler の定数である．

図 1 は，それぞれ $\xi = -0.4$，0，0.4 の場合の Z_1，Z_2，Z_3 の周辺密度関数である．
図 2，3，4 は，それぞれ $\xi = -0.4$，0，0.4 の場合の $(Z_1,\, Z_2)$ の同時密度関数

$$g_{\xi,12}(z_1, z_2) = \begin{cases} (1+\xi z_1)^{-1/\xi-1}(1+\xi z_2)^{-1/\xi-1}\exp\{-(1+\xi z_2)^{-1/\xi}\}\,, & \xi \neq 0 \\ \exp(-z_1 - z_2 - e^{-z_2})\,, & \xi = 0 \end{cases}$$

$z_1 \geq z_2$，とその等高線である．

3. パラメータ推定

一般極値分布 G_ξ の位置と尺度パラメータをそれぞれ μ，σ とする．ここでは，パラメータ $\boldsymbol{\theta} = (\mu,\, \sigma)$ または $(\mu,\, \sigma,\, \xi)$ と，極値理論で重要な上側微小確率点 T-return level (T 再現水準値)$q(T)$，

$$G_\xi\left(\frac{q(T)-\mu}{\sigma}\right) = 1 - \frac{1}{T}\,,$$

すなわち

$$q(T) = \begin{cases} \mu + \sigma\{-\log\big(-\log(1-1/T)\big)\}\,, & \xi = 0 \\ \mu + \sigma\{\big(-\log(1-1/T)\big)^{-\xi} - 1\}/\xi\,, & \xi \neq 0 \end{cases} \tag{3.1}$$

の最尤推定について述べる．

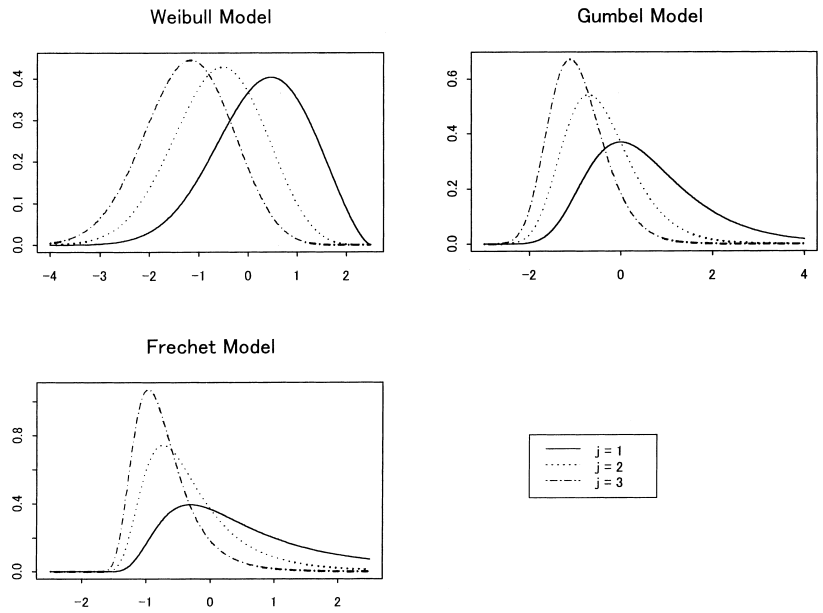

図 1. Weibull ($\xi = -0.4$)**,** Gumbel ($\xi = 0$)**,** Fréchet ($\xi = 0.4$) **モデルの場合の上位** $j(= 1, 2, 3)$ **番目の周辺密度関数.**

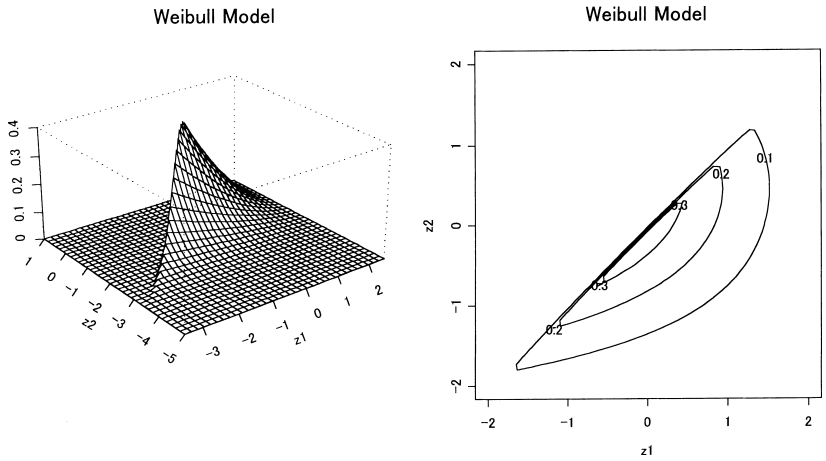

図 2. Weibull **モデル** ($\xi = -0.4$) **の場合の上位** 1**,** 2 **位の同時密度関数とその等高線.**

3.1 モデル

上位 r **個の確率ベクトル** $(X_1, X_2, \ldots, X_r)$ **の従う同時密度関数は,** Gumbel **モデルの場合は**

$$\frac{1}{\sigma^r}\, g_{0,12\cdots r}\left(\frac{x_1-\mu}{\sigma}, \ldots, \frac{x_r-\mu}{\sigma}\right)$$

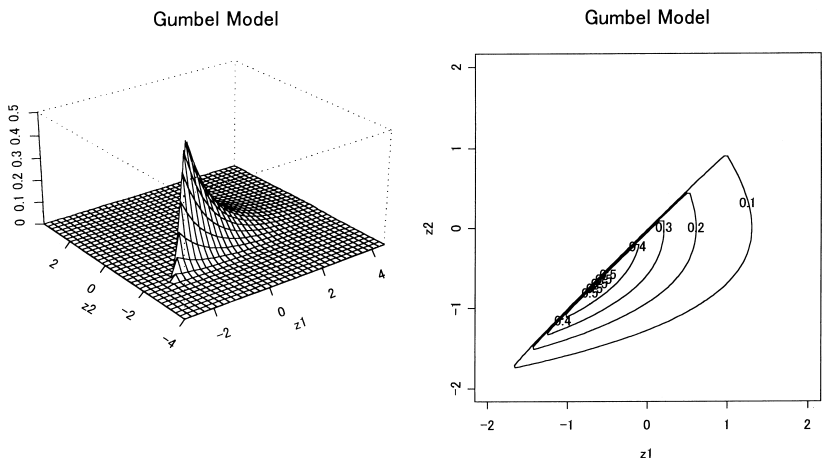

図 3. Gumbel **モデル** ($\xi = 0$) **の場合の上位** 1, 2 **位の同時密度関数とその等高線.**

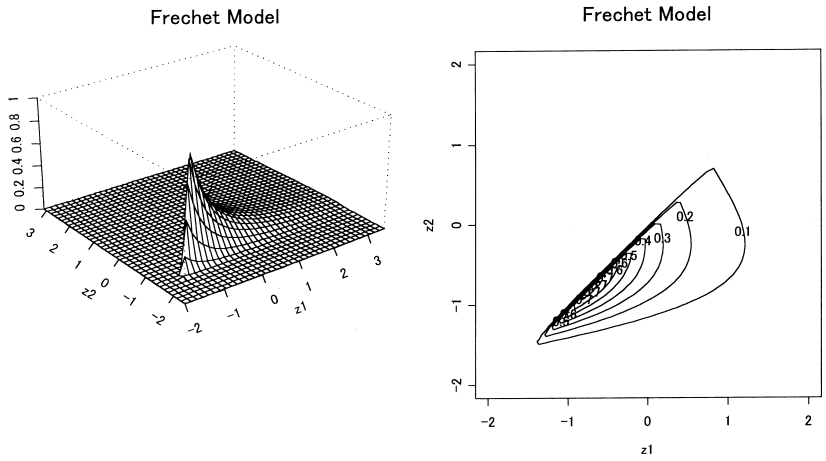

図 4. Fréchet **モデル** ($\xi = 0.4$) **の場合の上位** 1, 2 **位の同時密度関数とその等高線.**

で, GEV **モデルの場合は**

$$\frac{1}{\sigma^r}\, g_{\xi,12\cdots r}\left(\frac{x_1-\mu}{\sigma},\ldots,\frac{x_r-\mu}{\sigma}\right)$$

となる.

データは独立な $x_1 \geq x_2 \geq \cdots \geq x_r$ **が** n **組, すなわち**

$$x_{i1} \geq x_{i2} \geq \cdots \geq x_{ir};\ i = 1,\ldots,n$$

の $n \times r$ **個の数値とする. ただし,** r **は組ごとに異なっても** 3.2 **節,** 3.3 **節の対数尤度を書くことができ, パラメータの最尤推定値を求めることができる.**

3.2 Gumbel モデル

この場合の対数尤度は

$$l(\mu,\,\sigma) = -nr\log\sigma - \sum_{i=1}^{n}\left[\sum_{j=1}^{r}\left(\frac{x_{ij}-\mu}{\sigma}\right) + \exp\left(-\frac{x_{ir}-\mu}{\sigma}\right)\right]$$

となる．パラメータ $\theta = (\mu,\,\sigma)$ を最尤法で推定する．尤度方程式は次の様になる：

$$\begin{cases} \dfrac{\partial}{\partial\mu}l(\mu,\,\sigma) = -\displaystyle\sum_{i=1}^{n}\left[-\frac{r}{\sigma} + \frac{1}{\sigma}\exp\left(-\frac{x_{ir}-\mu}{\sigma}\right)\right] = 0\,, \\ \dfrac{\partial}{\partial\sigma}l(\mu,\,\sigma) = -\dfrac{nr}{\sigma} + \displaystyle\sum_{i=1}^{n}\left[\sum_{j=1}^{r}\left(\frac{x_{ij}-\mu}{\sigma^2}\right) - \frac{x_{ir}-\mu}{\sigma^2}\exp\left(-\frac{x_{ir}-\mu}{\sigma}\right)\right] = 0\,. \end{cases}$$

この連立非線形方程式から σ だけについての方程式

$$h_r(\sigma) := \sum_{i=1}^{n}\left(\frac{x_{ir}-\bar{\bar{x}}_r}{\sigma}+1\right)\exp\left(-\frac{x_{ir}-\bar{\bar{x}}_r}{\sigma}\right) = 0\,,\quad \bar{\bar{x}}_r = \frac{1}{nr}\sum_{i=1}^{n}\sum_{j=1}^{r}x_{ij}$$

が得られる．そこで

$$h_r'(\sigma) = \frac{1}{\sigma}\sum_{i=1}^{n}\left(\frac{x_{ir}-\bar{\bar{x}}_r}{\sigma}\right)^2\exp\left(-\frac{x_{ir}-\bar{\bar{x}}_r}{\sigma}\right)$$

を用いて，ニュートン法で $\widehat{\sigma}_r$ を求める．これから

$$\widehat{\mu}_r = -\widehat{\sigma}_r\,\log\left[\frac{1}{nr}\sum_{i=1}^{n}\exp\left(-\frac{x_{ir}}{\widehat{\sigma}_r}\right)\right]$$

が求まる．この $\widehat{\theta}_r = (\widehat{\mu}_r,\,\widehat{\sigma}_r)$ が，上位 r 個のデータを用いた場合の $\theta = (\mu,\,\sigma)$ の最尤推定値である．T-return level $q(T)$ の推定は

$$\widehat{q}_r(T) = \widehat{\mu}_r + \widehat{\sigma}_r\{-\log(-\log(1-1/T))\} \tag{3.2}$$

とすればよい．また，推定値 $\widehat{\mu}_r$，$\widehat{\sigma}_r$，$\widehat{q}_r(T)$ の標準誤差は付録 A.1 の漸近分散行列等を用いて推定する．

r の決定：ここでは，上位 r 個のデータ $\{(x_{i1},\,x_{i2},\ldots,x_{ir}),\,i=1,2,\ldots,n\}$ が分布 $G_{0,\,12\ldots r}$ に従うと見なせる最大の r の決定法について議論する．

上位 j 番目の確率変数 X_j は周辺分布関数 $G_{0,j}\big((z-\mu)/\sigma\big)$ を持つ．したがって，

$$U_{ij} = G_{0,j}\left(\frac{X_{ij}-\mu}{\sigma}\right)$$

により，X_{ij} を一様分布 $\mathrm{U}(0,\,1)$ からの確率変数 U_{ij} に変換できる．

このことから，次の周辺分布の適合を見る r の決定法が考えられる．

PP plot：r を固定し，上位 r 個のデータから推定値 $\widehat{\mu}_r$，$\widehat{\sigma}_r$ を求める．これらを用いて，上位 $j(=1,2,\ldots,r)$ 番目のデータ $\{x_{1j},\,x_{2j},\ldots,x_{nj}\}$ から $u_{ij} = G_{0,j}((x_{ij}-\widehat{\mu}_r)/\widehat{\sigma}_r),\,i=1,2,\ldots,n$ を求める．この u_{ij} の順序統計量を $u_{(1)j}\geq u_{(2)j}\geq\cdots\geq u_{(n)j}$ とし，

$$\left(1-\frac{i}{n+1},\,u_{(i)j}\right),\quad i=1,2,\ldots,n$$

をプロットし，r 個の PP plot を作成する．そして r を動かし，r 個すべての PP plot でプロットが直線性を示していると見なせる最大の r を決定する．

一方，次の決定法も考えられる．

QQ plot：r を固定し，上位 r 個のデータから推定値 $\widehat{\mu}_r$，$\widehat{\sigma}_r$ を求める．$j(=1,2,\ldots,r)$ に対して確率点 $q_{(i)j}=G_{0,j}^{-1}(1-i/(n+1))$ を求める．上位 j 番目のデータ $\{x_{1j},\, x_{2j},\ldots,x_{nj}\}$ の順序統計量を $x_{(1)j}\geq x_{(2)j}\geq\cdots\geq x_{(n)j}$ とし，

$$(\widehat{\mu}_r+\widehat{\sigma}_r q_{(i)j},\, x_{(i)j})\,,\quad i=1,2,\ldots,n$$

をプロットし，r 個の QQ plot を作成する．そして r を動かし，r 個すべての QQ plot でプロットが直線性を示していると見なせる最大の r を決定する．ここで，$j\geq 2$ に対しては数値計算で確率点 $q_{(i)j}$ を求める必要がある．

これら PP plot，QQ plot による方法では周辺分布の適合しか見ていない．分布の同時性をチェックする方法として，定理 1（I.b）より次のものが考えられる．

指数確率紙：$j=1,2,\ldots$ に対して $\{x_{ij}-x_{ij+1},\ \ i=1,2,\ldots,n\}$ を指数確率紙にプロットする．すべての $j(<r)$ の指数確率紙でプロットが直線性を示していると見なせる最大の r を決定する．

したがって，r として PP plot，QQ plot と指数確率紙から得られた中で最小のものを採用する．

PP plot，QQ plot を用いる方法はそれぞれ Smith（1986），Coles（2001）で提案された．指数確率紙を用いることは Smith（1986）で言及されてはいるがデータ解析には使われていない．

補助変量を含む場合

Smith（1986）はパラメータが補助変量の関数になる場合を議論している．ベニスの潮位データが年 $(i=1,2,\ldots,n)$ とともに増加の傾向があることから，次の線型トレンドモデル：

$$\mu_i=\alpha+\frac{\beta}{n}i\,,\quad \sigma_i=\sigma\,,\quad i=1,2,\ldots,n$$

を提案した．この場合の対数尤度は

$$l(\alpha,\,\beta,\,\sigma)=-nr\log\sigma-\sum_{i=1}^{n}\left[\sum_{j=1}^{r}\left(\frac{x_{ij}-\alpha-\beta\, i/n}{\sigma}\right)+\exp\left(-\frac{x_{ir}-\alpha-\beta\, i/n}{\sigma}\right)\right]$$

と書ける．これを数値計算で最大化し最尤推定値 $(\widehat{\alpha}_r,\,\widehat{\beta}_r,\,\widehat{\sigma}_r)$ を求める．推定値の標準誤差は観測情報行列の逆行列または数値微分を用いて求める．

このとき i 年の T-return level は

$$\widehat{\alpha}_r+\frac{\widehat{\beta}_r}{n}i+\widehat{\sigma}_r\{-\log\big(-\log(1-1/T)\big)\}$$

で推定できる．

r の決定は上と同様にできる．ただし，PP plot と QQ plot で $\widehat{\mu}_r$ の代わりに

$$\widehat{\mu}_i=\widehat{\alpha}_r+\frac{\widehat{\beta}_r}{n}i\,,\quad i=1,2,\ldots,n$$

を用いる．

パラメータが補助変量のもっと複雑な関数の場合も同様にできる．

3.3 GEV モデル

この場合の対数尤度は

$$l(\mu,\,\sigma,\,\xi)=-nr\log\sigma-\sum_{i=1}^{n}\left\{\left(\frac{1}{\xi}+1\right)\sum_{j=1}^{r}\log\left[1+\xi\left(\frac{x_{ij}-\mu}{\sigma}\right)\right]+\left[1+\xi\left(\frac{x_{ir}-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}$$

となる．パラメータ $\theta = (\mu, \sigma, \xi)$ を最尤法で推定する．

尤度方程式は簡単にはならない．ニュートン法で連立非線形の尤度方程式を解かなければならないが，初期値としては極値データから PWM 法(付録 A.3 参照) で求めた推定値を用いればよい．非線形最適化のソフトを利用して最尤推定値を求める方法もある．得られた最尤推定値を $\widehat{\theta}_r = (\widehat{\mu}_r, \widehat{\sigma}_r, \widehat{\xi}_r)$ とすると，T-return level $q(T)$ の推定は

$$\widehat{q}_r(T) = \widehat{\mu}_r + \widehat{\sigma}_r\{(-\log(1-1/T))^{-\widehat{\xi}_r} - 1\}/\widehat{\xi}_r$$

とすればよい．また，推定値 $\widehat{\mu}_r$，$\widehat{\sigma}_r$，$\widehat{\xi}_r$，$\widehat{q}_r(T)$ の標準誤差は付録 A.2 の Fisher 情報行列または観測情報行列を用いて推定する．一般に，標準誤差の推定値は観測情報行列の逆行列から求めた方が精度がよいことが知られている．また，信頼区間を求めるにはプロファイル対数尤度から求めた方が精度がよいことも知られている．Coles (2001)参照．

r の決定：ここでも，上位 r 個のデータ $\{(x_{i1}, x_{i2}, \ldots, x_{ir}),\ i = 1, 2, \ldots, n\}$ が分布 $G_{\xi,\, 12\ldots r}$ に従うと見なせる最大の r の決定法について議論する．

上位 j 番目の確率変数 X_j は周辺分布関数 $G_{\xi,j}\big((z-\mu)/\sigma\big)$ を持つ．したがって，

$$U_{ij} = G_{\xi,j}\left(\frac{X_{ij}-\mu}{\sigma}\right)$$

により，X_{ij} を一様分布 U(0, 1) からの確率変数 U_{ij} に変換できる．

このことから，次の周辺分布の適合を見る r の決定法が考えられる．

PP plot：r を固定し，上位 r 個のデータから推定値 $\widehat{\mu}_r$，$\widehat{\sigma}_r$，$\widehat{\xi}_r$ を求める．これらを用いて，上位 $j(= 1, 2, \ldots, r)$ 番目のデータ $\{x_{1j}, x_{2j}, \ldots, x_{nj}\}$ から $u_{ij} = G_{\widehat{\xi}_r,j}((x_{ij} - \widehat{\mu}_r)/\widehat{\sigma}_r),\ i = 1, 2, \ldots, n$ を求める．この u_{ij} の順序統計量を $u_{(1)j} \geq u_{(2)j} \geq \cdots \geq u_{(n)j}$ とし，

$$\left(1 - \frac{i}{n+1},\ u_{(i)j}\right), \quad i = 1, 2, \ldots, n$$

をプロットし，r 個の PP plot を作成する．そして r を動かし，r 個すべての PP plot でプロットが直線性を示していると見なせる最大の r を決定する．

また，次の方法も考えられる．

QQ plot：r を固定し，上位 r 個のデータから推定値 $\widehat{\mu}_r$，$\widehat{\sigma}_r$，$\widehat{\xi}_r$ を求める．$j(= 1, 2, \ldots, r)$ に対して，確率点 $q_{(i)j} = G^{-1}_{\widehat{\xi}_r,j}(1 - i/(n+1))$ を求める．j 番目のデータ $\{x_{1j}, x_{2j}, \ldots, x_{nj}\}$ の順序統計量を $x_{(1)j} \geq x_{(2)j} \geq \cdots \geq x_{(n)j}$ とし，

$$(\widehat{\mu}_r + \widehat{\sigma}_r q_{(i)j},\ x_{(i)j}), \quad i = 1, 2, \ldots, n$$

をプロットし，r 個の QQ plot を作成する．そして r を動かし，r 個すべての QQ plot でプロットが直線性を示していると見なせる最大の r を決定する．ここでも，$j \geq 2$ に対しては数値計算で確率点 $q_{(i)j}$ を求める必要がある．

分布の同時性をチェックする方法として，定理 1 (II.b)より次のものが考えられる．

指数確率紙：r を固定し，上位 r 個のデータから推定値 $\widehat{\mu}_r$，$\widehat{\sigma}_r$，$\widehat{\xi}_r$ を求める．上位 $j(= 1, 2, \ldots, r-1)$ 番目と $j+1$ 番目のデータ $\{(x_{ij},\ x_{ij+1}),\ i = 1, 2, \ldots, n\}$ から

$$\frac{j}{\widehat{\xi}_r} \log \frac{\widehat{\sigma}_r + \widehat{\xi}_r(x_{ij} - \widehat{\mu}_r)}{\widehat{\sigma}_r + \widehat{\xi}_r(x_{ij+1} - \widehat{\mu}_r)}, \quad i = 1, 2, \ldots, n$$

を求めて指数確率紙にプロットし，$r-1$ 個の指数確率紙を作成する．そして r を動かし，$r-1$ 個すべての指数確率紙でプロットが直線性を示していると見なせる最大の r を決定する．

したがって，r として PP plot，QQ plot と指数確率紙から得られた中の最小のものを採用する．

PP plot，QQ plot による方法はそれぞれ Tawn（1988），Coles（2001）で提案された．指数確率紙による方法は Tawn（1988）で言及されてはいるがデータ解析には使われていない．

補助変量を含む場合

Tawn（1988）はパラメータ μ が補助変量の関数になる場合を議論している．この場合は，3.2 節の Gumbel モデルで補助変量を含む場合と同様にできる．

Scarf et al.（1992）は次のような腐食データ：

$$(x_{ij},\, t_i)\,,\quad i=1,\dots,n\,;\quad j=1,\dots,r_i;\ (x_{i1}\geq\cdots\geq x_{ir_i})\,,$$

すなわち，時刻 t_i での上位 r_i 個の腐食孔の深さデータ $\{(x_{i1},\dots,x_{ir_i}),\ i=1,2,\dots,n\}$ の解析法を議論している．彼らは，腐食孔の最大深さが時間 t とともに進行するモデルとして

$$\mu_t=\mu t^{\beta}\,,\quad \sigma_t=\sigma t^{\beta}$$

を考えた．この場合の対数尤度は

$$l(\mu,\,\sigma,\,\beta,\,\xi)=-\sum_{i=1}^{n}\left\{r_i\big(\log\sigma+\beta\log t_i\big)+\left(\frac{1}{\xi}+1\right)\sum_{i=1}^{r_i}\log\left[1+\xi\left(\frac{x_{ij}t_i^{-\beta}-\mu}{\sigma}\right)\right]\right.$$

$$\left.-\left[1+\xi\left(\frac{x_{ir_i}t_i^{-\beta}-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}$$

と書ける．これを数値計算で最大化し最尤推定値 $(\widehat{\mu},\,\widehat{\sigma},\,\widehat{\beta},\,\widehat{\xi})$ を求める．推定値の標準誤差は観測情報行列の逆行列または数値微分を用いて求める．

$r_i\ (i=1,2,\dots,n)$ の決定は次の様にする．$r_0=\min\{r_1,\dots,r_n\}$ として，この r_0 までの r に関しては上の 3 つの方法を使う．ただし，$\widehat{\mu}_r$ と $\widehat{\sigma}_r$ は次のもので置き換える：

$$\widehat{\mu}\,t_i^{\widehat{\beta}}\,,\quad \widehat{\sigma}\,t_i^{\widehat{\beta}}\,,\quad i=1,2,\dots,n\,.$$

この方法で $r=r^*(\leq r_0)$ を決める．r^* を超える r_i に関しては，パラメータの推定値の安定性や定理 1 の性質をデータのプロット等で調べ注意深く決定する．通常の統計学と同様に，r_i $(i=1,\dots,n)$ に極端な違いがあるのは望ましくないと思われる．

4. 有効性

ここでは，パラメータ $(\mu,\,\sigma,\,\xi)$ が補助変量に依存しない場合を議論する．

上位 $r(\geq 2)$ 個のデータを用いる有効性を T-return level $q(T)$ の推定精度が改善されることで示す．そのために，$q(T)$ の推定で極値データのみを用いる場合と，上位 r 個を用いる場合の漸近分散の比，漸近相対効率，を考える．

まず，Gumbel モデルの場合を議論する．

推定量 $\widehat{q}_r(T)$ の漸近分散は，付録 A.1 より

$$AV(\widehat{q}_r(T))=\frac{\sigma^2}{n(rC_r-B_r^2)}(C_r+2g(T)B_r+(g(T))^2r)\,,\quad g(T)=-\log(-\log(1-1/T))$$

となる．したがって，$\widehat{q}_1(T)$ に対する $\widehat{q}_r(T)$ の漸近相対効率は

$$e_0(\widehat{q}_r(T),\,\widehat{q}_1(T))=\frac{AV(\widehat{q}_1(T))}{AV(\widehat{q}_r(T))}=\frac{(rC_r-B_r^2)(C_1+2g(T)B_1+(g(T))^2)}{(C_1-B_1^2)(C_r+2g(T)B_r+(g(T))^2r)}\tag{4.1}$$

と表される．この漸近相対効率は r と T の関数になるから，

$$e_0(r,\, T) = e_0(\widehat{q}_r(T)\,,\ \widehat{q}_1(T)) \tag{4.2}$$

とおく．

$r = 2, 3, \ldots, 10$ と T =100，1,000，10,000，100,000 に対する $e_0(r,\, T)$ を表 1 に示す．表 1 より，例えば $r = 3$，$T = 10,000$ のとき $e_0(3,\, 10,000) = 1.995$ である．すなわち，10,000-return level を推定するとき，上位 3 個の 50 組のデータは 100 個の極値データとほぼ同じ精度，ほぼ等しい漸近分散，を持つと言える．また，漸近相対効率は r に比例しては増加していない．

漸近相対効率 $e_0(r,\, T)$ に関して次の命題が成り立つ．

命題 1. (I) T を固定する．

(I.a) $e_0(r,\, T)$ は r の狭義単調増加関数で ∞ に発散する．

(I.b) $g(T) > 0$ のとき，$e_0(r,\, T)/r$ は r の狭義単調減少関数．

(II) r を固定する．T が十分大きいとき，$e_0(r,\, T)$ は T の狭義単調増加関数で r に依存する定数

$$\frac{r\big(\psi'(r+1)+1\big)}{\psi'(2)+1}$$

に収束する．

証明. 以下，狭義単調増加(減少)関数を増加(減少)関数と言う．

(I) (I.a) 付録 A.1 より $\Sigma_{r-1}(\boldsymbol{\theta}) - \Sigma_r(\boldsymbol{\theta})$ が正定値だから，

$$AV(\widehat{q}_r(T)) < AV(\widehat{q}_{r-1}(T))\,.$$

したがって

$$e_0(r,\, T) > e_0(r-1,\, T)\,, \quad r = 2, 3, \ldots$$

一方，(4.1)の分子分母を r^2 で割り $r \to \infty$ とすると，

$$\text{分子} = c_1(\psi'(r+1)+1) \to c_1\,,$$

$$\text{分母} = c_2\left\{\frac{C_r}{r^2} + 2g(T)\frac{B_r}{r^2} + \frac{(g(T))^2}{r}\right\} \to 0\,,$$

ただし，c_1，c_2 は正の定数．これから，

$$\lim_{r\to\infty} e_0(r,\, T) = \infty\,.$$

表 1. $\widehat{q}_1(T)$ に対する $\widehat{q}_r(T)$ の漸近相対効率： $e_0(r,\, T)$.

T	$r=2$	3	4	5	6	7	8	9	10
100	1.429	1.769	2.073	2.357	2.627	2.886	3.137	3.382	3.620
1,000	1.498	1.909	2.284	2.638	2.978	3.307	3.627	3.940	4.246
10,000	1.539	1.995	2.417	2.818	3.205	3.581	3.949	4.310	4.665
100,000	1.566	2.053	2.507	2.941	3.362	3.773	4.176	4.572	4.963
∞	1.696	2.341	2.970	3.591	4.208	4.822	5.435	6.047	6.658

（I.b） ここで

$$\frac{r}{\sigma^2}\Sigma_r(\boldsymbol{\theta}) = \frac{1}{\psi'(r+1)+1}\begin{bmatrix} \psi^2(r+1)+\psi'(r+1)+1 & \psi(r+1) \\ \psi(r+1) & 1 \end{bmatrix}$$

より，

$$\frac{1}{\sigma^2}\{r\Sigma_r(\boldsymbol{\theta}) - (r-1)\Sigma_{r-1}(\boldsymbol{\theta})\} = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$$

とおくと，

$$a = \frac{\psi^2(r+1)+\psi'(r+1)+1}{\psi'(r+1)+1} - \frac{\psi^2(r)+\psi'(r)+1}{\psi'(r)+1}\,,$$

$$b = \frac{\psi(r+1)}{\psi'(r+1)+1} - \frac{\psi(r)}{\psi'(r)+1}\,, \quad c = \frac{1}{\psi'(r+1)+1} - \frac{1}{\psi'(r)+1}$$

である．ψ は増加関数で ψ' は減少関数だから $b,\ c > 0$．一方，$a > 0$ は a の式の右辺の分子どうしの差が正より示せる．

よって，$g(T) > 0$ のとき

$$\begin{bmatrix} 1 \\ g(T) \end{bmatrix}' \{r\Sigma_r(\boldsymbol{\theta}) - (r-1)\Sigma_{r-1}(\boldsymbol{\theta})\} \begin{bmatrix} 1 \\ g(T) \end{bmatrix} = \sigma^2\{a + 2b\,g(T) + c\,g^2(T)\} > 0\,.$$

したがって

$$r\,AV(\widehat{q}_r(T)) > (r-1)AV(\widehat{q}_{r-1}(T))$$

より

$$\frac{e_0(r,\,T)}{r} < \frac{e_0(r-1,\,T)}{r-1}\,, \quad r = 2, 3, \ldots$$

（II） r を固定する．$g(T)$ は T の増加関数だから，次の g の関数について考えれば良い：

$$e(g) = c\,\frac{g^2 + 2B_1 g + C_1}{rg^2 + 2B_r g + C_r}\,, \quad c\text{：正定数}.$$

g が十分大のとき，$e(g)$ が増加関数であることを証明する．

$e(g)$ を g で微分し，正の定数倍を無視した分子のみを考え，それを $n(g)$ とすると

$$n(g) = (B_r - rB_1)g^2 + (C_r - rC_1)g + (B_1C_r - B_rC_1)$$

となる．ここで，$r \geq 2$ のとき g^2 と g の係数は 次より共に正である：

$$B_r - rB_1 = r\{\psi(r+1) - \psi(2)\} > 0\,,$$

$$\begin{aligned} C_r - rC_1 &= r\{\psi^2(r+1)+\psi'(r+1)+1\} - r\{\psi^2(2)+\psi'(2)+1\} \\ &= r\{\psi^2(r+1) - \psi^2(2) + \psi'(r+1) - \psi'(2)\} \\ &= r\left\{\left(\psi(2) + \sum_{i=2}^{r}\frac{1}{i}\right)^2 - \psi^2(2) - \sum_{i=2}^{r}\frac{1}{i^2}\right\} > 0\,. \end{aligned}$$

したがって，g が十分大きいとき $n(g) > 0$ となり $de(g)/dg > 0$ である．このことから，$e_0(r,\,T)$ は T の増加関数になる．（特に，表 1 での r を固定したとき，$T(\geq 100)$ に関して $e_0(r,\,T)$ は増加関数になっている．）

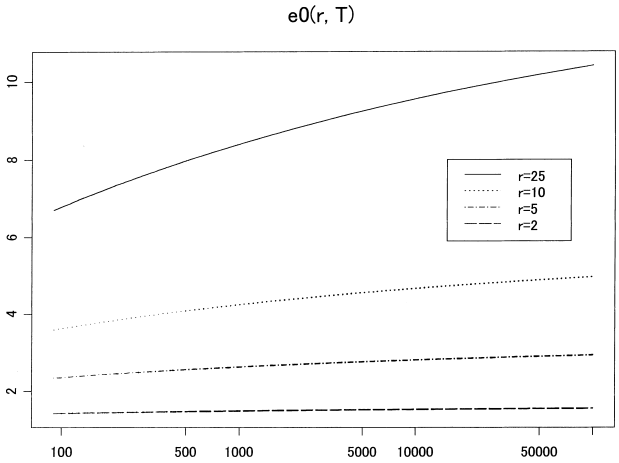

図 5. 漸近相対効率 $e_0(r,\,T) = e_0(\widehat{q}_r(T),\,\widehat{q}_1(T))$.

一方, (4.1) **より**

$$\lim_{T\to\infty} e_0(r,\,T) = \frac{rC_r - B_r^2}{r(C_1 - B_1^2)} = \frac{r\big(\psi'(r+1)+1\big)}{\psi'(2)+1}$$

が示される. □

図 5 **は相対漸近効率** $e_0(r,\,T)$ **の図である.** r **を固定したとき,** $e_0(r,\,T)$ **は** T **に関して非常に緩やかに増加している.**

GEV **モデルの下での有効性の議論は難しい.**

この場合の, $\widehat{q}_1(T)$ **に対する** $\widehat{q}_r(T)$ **の漸近相対効率は付録** A.2 **より,** ξ **と** r **と** T **の関数になるから**

$$e_\xi(r,\,T) = e_\xi(\widehat{q}_r(T),\,\widehat{q}_1(T)) \tag{4.3}$$

と表すことにする.

表 2**,** 3**,** 4**,** 5 **は,それぞれ** $\xi = -0.3, -0.1,\,0.1,\,0.3$ **の場合に表** 1 **と同じ** r**,** T **について** $e_\xi(r,\,T)$ **を数値計算したものである.命題** 1 **の** $e_0(r,\,T)$ **と同様の性質が見て取れる.また,** $\xi = 0$ **の場合と比べて漸近相対効率は良くない.**

この節での議論は,「上位 r **個のデータが正確に想定した漸近同時分布からのものである」と仮定している事に注意する必要がある.**

5. 腐食データの解析

ここでは,給湯用銅管の腐食を実験室で 9**ヶ月間加速再現したサンプルの腐食孔の深さデータの解析を行う.データは,一つの銅管の面積が等しい** 30 **個の領域内において上位** 3 **個の腐食孔の深さ(単位** mm**)の測定値からなる.図** 6 **は,データの(**1 **位,** 2 **位),(**2 **位,** 3 **位)の散布図である.相関係数はそれぞれ** 0.813 **と** 0.851 **となった.**

以下, Gumbel **と** GEV **モデルの下で解析した結果を述べるが,データの取得状況からパラメータ** $(\mu,\,\sigma)$ **が補助変量の関数になると考える必要はない.**

表 2. 漸近相対効率：$e_{-0.3}(r,\,T)$.

T	$r=2$	3	4	5	6	7	8	9	10
100	1.018	1.030	1.042	1.054	1.065	1.077	1.087	1.098	1.108
1,000	1.080	1.138	1.186	1.229	1.267	1.302	1.334	1.364	1.393
10,000	1.135	1.232	1.311	1.379	1.440	1.495	1.546	1.593	1.638
100,000	1.170	1.291	1.390	1.475	1.551	1.621	1.685	1.744	1.800

表 3. 漸近相対効率：$e_{-0.1}(r,\,T)$.

T	$r=2$	3	4	5	6	7	8	9	10
100	1.083	1.146	1.202	1.253	1.300	1.344	1.386	1.426	1.464
1,000	1.190	1.334	1.456	1.566	1.666	1.760	1.848	1.932	2.012
10,000	1.264	1.467	1.642	1.800	1.946	2.082	2.211	2.334	2.452
100,000	1.313	1.559	1.773	1.968	2.148	2.318	2.480	2.635	2.784

表 4. 漸近相対効率：$e_{0.1}(r,\,T)$.

T	$r=2$	3	4	5	6	7	8	9	10
100	1.152	1.267	1.368	1.460	1.545	1.625	1.701	1.774	1.844
1,000	1.268	1.478	1.661	1.829	1.985	2.132	2.272	2.407	2.537
10,000	1.342	1.618	1.863	2.089	2.301	2.503	2.697	2.885	3.067
100,000	1.391	1.713	2.003	2.273	2.528	2.773	3.009	3.239	3.462

表 5. 漸近相対効率：$e_{0.3}(r,\,T)$.

T	$r=2$	3	4	5	6	7	8	9	10
100	1.205	1.362	1.500	1.625	1.742	1.853	1.959	2.060	2.158
1,000	1.315	1.566	1.789	1.994	2.186	2.370	2.545	2.715	2.880
10,000	1.380	1.692	1.973	2.234	2.481	2.718	2.947	3.169	3.385
100,000	1.421	1.773	2.093	2.394	2.680	2.956	3.223	3.484	3.738

5.1 Gumbel モデル

先ず，極値データ ($r=1$) が Gumbel 分布に従うと見なしてよいかを調べた．すなわち，一般極値分布の形状パラメータ ξ について，検定 $\mathrm{H}_0:\xi=0$，$\mathrm{H}_1:\xi\neq 0$ を ξ の PWM 推定量に基づく方法（付録 A.3 参照）で行った．ξ の PWM 推定値は $\widehat{\xi}=-0.191$，検定統計量の値は 1.395 で p 値は 0.163 となった．

上位 $r=1,\,2,\,3$ 個のデータを用いた場合の μ と σ の最尤推定値は次の様になった：

$$
\begin{aligned}
r=1: &\quad \widehat{\mu}_1=0.420\,, \quad \widehat{\sigma}_1=0.037\,,\\
r=2: &\quad \widehat{\mu}_2=0.423\,, \quad \widehat{\sigma}_2=0.036\,,\\
r=3: &\quad \widehat{\mu}_3=0.431\,, \quad \widehat{\sigma}_3=0.045\,.
\end{aligned}
$$

ここで，r の決定を考える．PP plot，QQ plot を書かせたのが図 7，8 で，指数確率紙が図

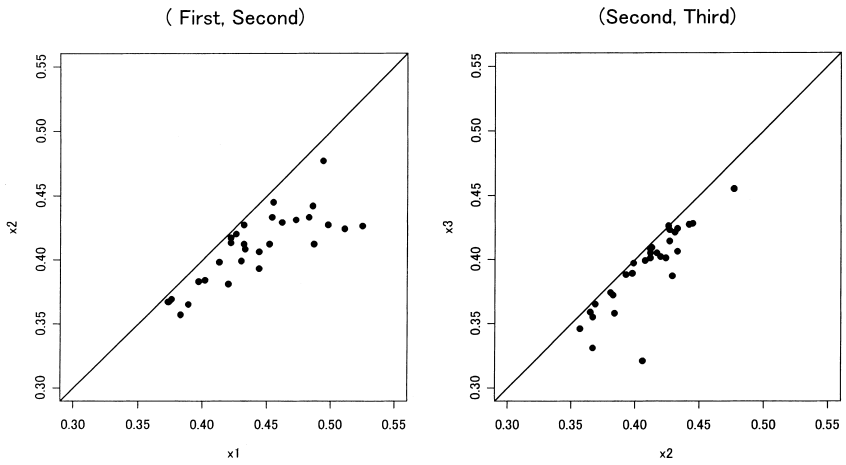

図 6.　(1 位, 2 位) と (2 位, 3 位) のデータの組の散布図.

9 である．PP plot と QQ plot では適合の度合いがずいぶん違う．QQ plot によると他の PP plot や指数確率紙と違い $r = 2$ も怪しくなる．したがって，Gumbel モデルの下では極値データのみを用いて推定を行うことになる．

この場合の T-return level の推定は

$$\widehat{q}_1(T) = 0.420 + 0.037\{-\log(-\log(1 - 1/T))\}$$

となる．

5.2　GEV モデル

このモデルの下で，上位 $r = 1,\ 2,\ 3$ 個のデータを用いた場合の最尤推定値は次の様になった：

$$\begin{aligned} r = 1: &\quad \widehat{\mu}_1 = 0.425\,, \quad \widehat{\sigma}_1 = 0.040\,, \quad \widehat{\xi}_1 = -0.253\,, \\ r = 2: &\quad \widehat{\mu}_2 = 0.425\,, \quad \widehat{\sigma}_2 = 0.034\,, \quad \widehat{\xi}_2 = -0.156\,, \\ r = 3: &\quad \widehat{\mu}_3 = 0.432\,, \quad \widehat{\sigma}_3 = 0.036\,, \quad \widehat{\xi}_3 = -0.280\,. \end{aligned}$$

r の決定のために PP plot，QQ plot を書かせたのが図 10，11 で，指数確率紙が図 12 である．これらから，このモデルの下では上位 2 個までのデータが使えると考えられる．

したがって，この場合の T-return level の推定は

$$\widehat{q}_2(T) = 0.425 + 0.034\{(-\log(1 - 1/T))^{0.156} - 1\}/(-0.156)$$

とすればよい．また，最大腐食孔の深さの上限値は 0.646 と推定される．

推定値から，$\xi < 0$ の可能性が強く，最大値の分布は上限のある Weibull 分布の可能性が高い．しかし，$r = 2$ の場合の形状パラメータ ξ の推定値 $\widehat{\xi}_2 = -0.156$ の標準誤差は 0.106 で 95% 信頼区間は正の値を含む．

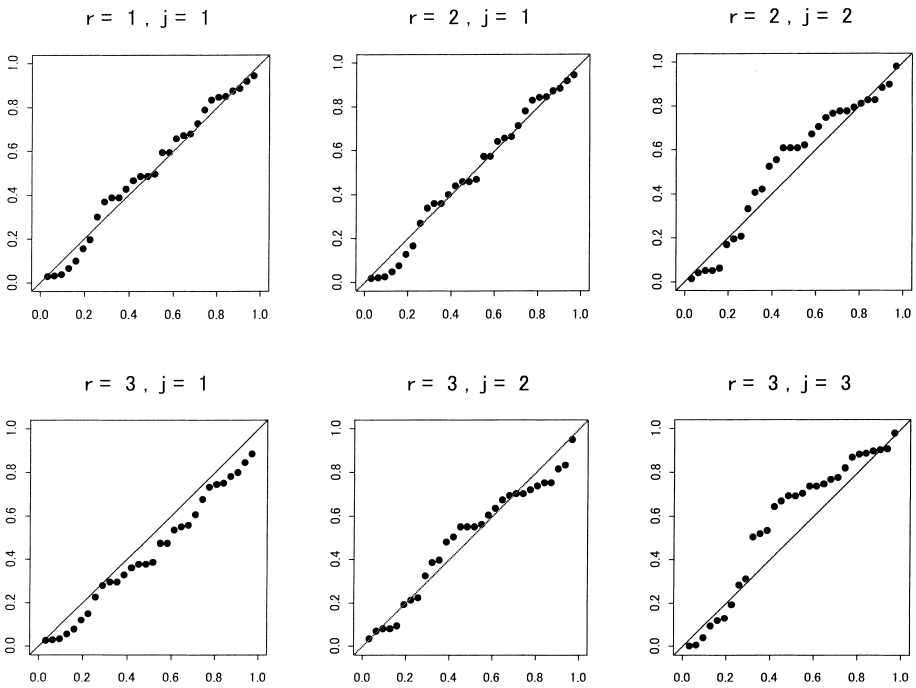

図 7. Gumbel **モデルで，上位** r **個のデータを用いた場合の上位** j **番目のデータの** PP plot**.**

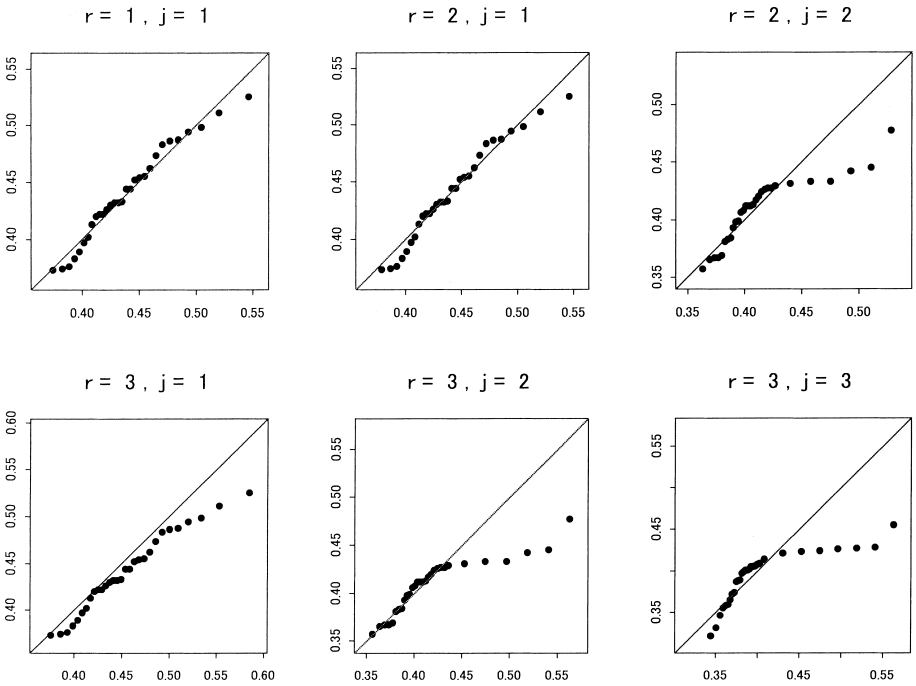

図 8. Gumbel **モデルで，上位** r **個のデータを用いた場合の上位** j **番目のデータの** QQ plot**.**

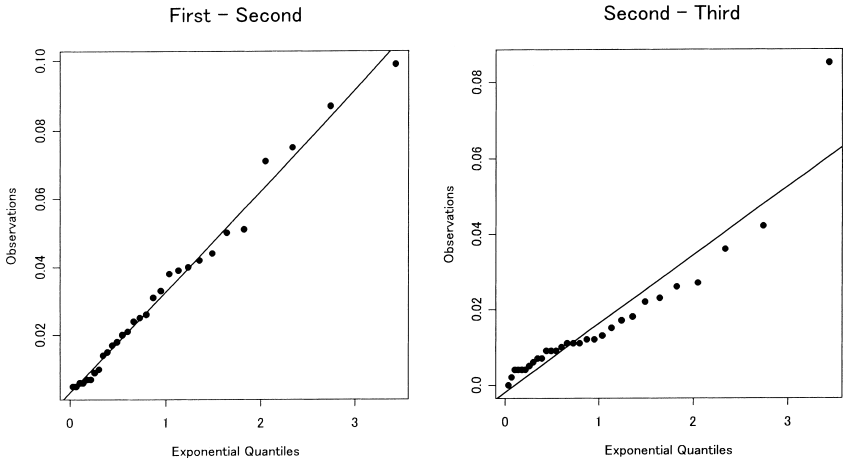

図 9. Gumbel モデルの下で上位 r 個のデータを用いた場合の（1 位, 2 位）と（2 位, 3 位）データによる指数確率プロットと回帰直線.

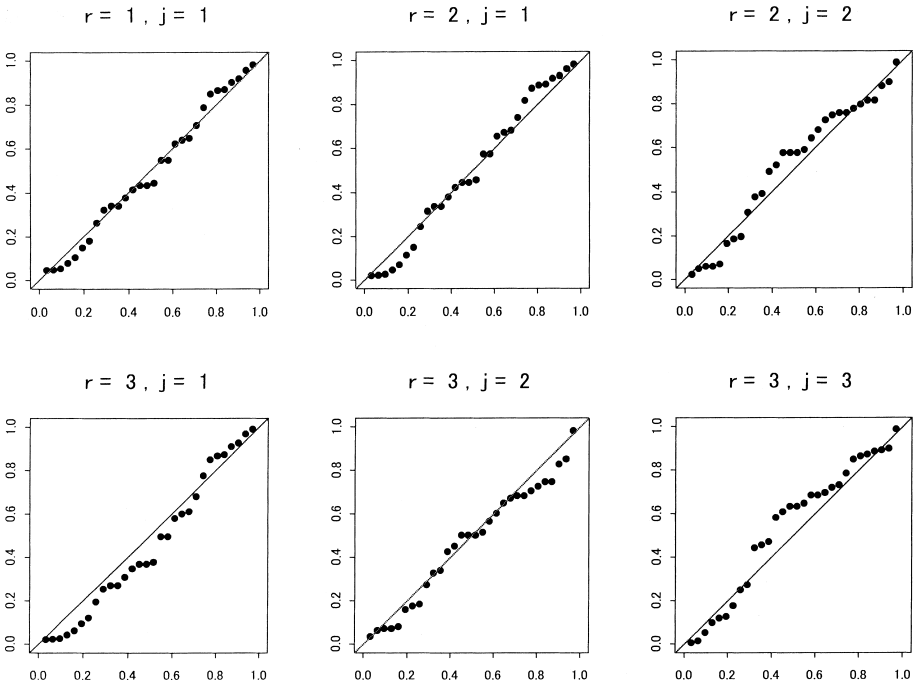

図 10. GEV モデルで，上位 r 個のデータを用いた場合の上位 j 番目のデータの PP plot.

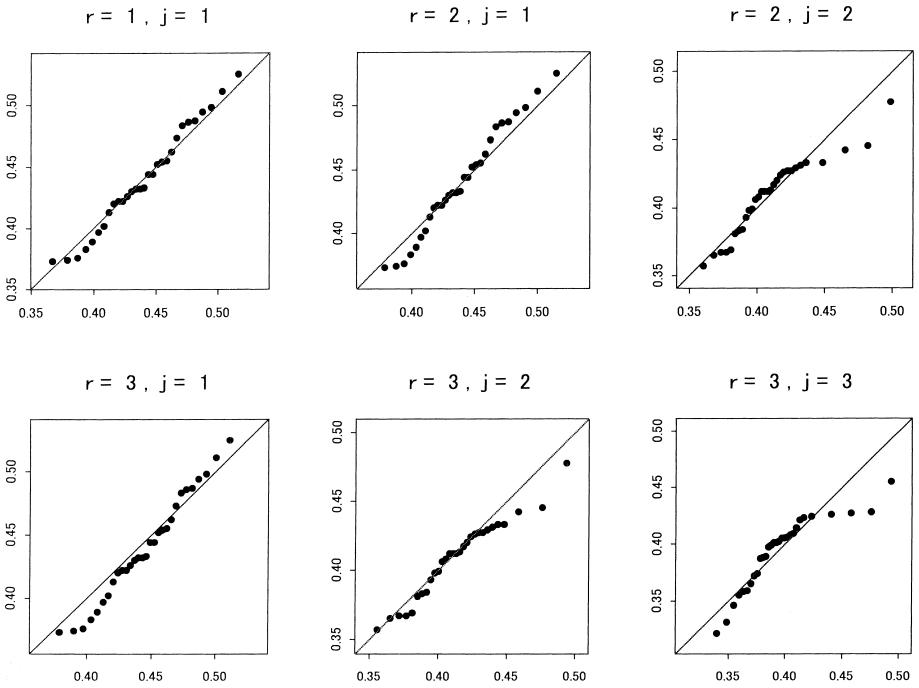

図 11. GEV **モデルで，上位** r **個のデータを用いた場合の上位** j **番目のデータの** QQ plot**.**

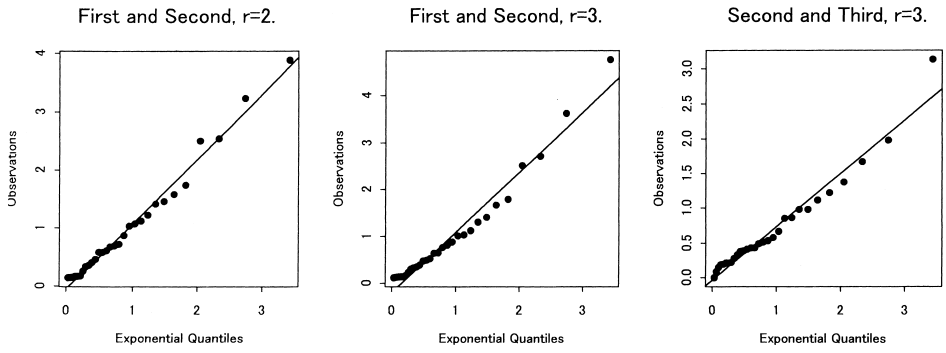

図 12. GEV **モデルの下で上位** r **個のデータを用いた場合の(1 位, 2 位)と(2 位, 3 位)データによる指数確率プロットと回帰直線.**

付 録

A.1 Gumbel モデルの場合の Fisher 情報量

Smith (1986) より，上位 r 個の同時分布の Fisher 情報行列は

$$I_r(\boldsymbol{\theta}) = \frac{1}{\sigma^2}\begin{bmatrix} r & -B_r \\ -B_r & C_r \end{bmatrix} \tag{A.1}$$

となる．ただし，$\boldsymbol{\theta} = (\mu,\, \sigma)$,

$$B_r = r\psi(r+1), \quad C_r = r(\psi^2(r+1) + \psi'(r+1) + 1). \tag{A.2}$$

逆行列は

$$\Sigma_r(\boldsymbol{\theta}) = \frac{\sigma^2}{rC_r - B_r^2}\begin{bmatrix} C_r & B_r \\ B_r & r \end{bmatrix} \tag{A.3}$$

と表される．

ここで

$$I_r(\boldsymbol{\theta}) - I_{r-1}(\boldsymbol{\theta}) = \frac{1}{\sigma^2}\begin{bmatrix} 1 & B_{r-1} - B_r \\ B_{r-1} - B_r & C_r - C_{r-1} \end{bmatrix} = \frac{1}{\sigma^2}\begin{bmatrix} 1 & -\psi(r) - 1 \\ -\psi(r) - 1 & (\psi(r)+1)^2 + \psi'(r) \end{bmatrix}$$

より，任意の $[x,\, y]' \neq \mathbf{0}$ に対して

$$\begin{bmatrix} x \\ y \end{bmatrix}' (I_r(\boldsymbol{\theta}) - I_{r-1}(\boldsymbol{\theta})) \begin{bmatrix} x \\ y \end{bmatrix} = \frac{1}{\sigma^2}\{(x - (\psi(r)+1)y)^2 + \psi'(r)y^2\} > 0\,.$$

したがって，$I_r(\boldsymbol{\theta}) - I_{r-1}(\boldsymbol{\theta})$ は正定値である．このことから，$\Sigma_{r-1}(\boldsymbol{\theta}) - \Sigma_r(\boldsymbol{\theta})$ も正定値となる．

サイズ n の標本による最尤推定量 $\widehat{\boldsymbol{\theta}}_r = (\widehat{\mu}_r,\, \widehat{\sigma}_r)$ の漸近分散行列は $\Sigma_r(\boldsymbol{\theta})/n$ になる．

T-return level $q(T)$ の最尤推定量は

$$\widehat{q}_r(T) = \widehat{\mu}_r + \widehat{\sigma}_r g(T)\,, \quad g(T) = -\log(-\log(1 - 1/T))$$

で，その漸近分散はデルタ法より

$$AV(\widehat{q}_r(T)) = \begin{bmatrix} 1 \\ g(T) \end{bmatrix}' \frac{1}{n}\Sigma_r(\boldsymbol{\theta}) \begin{bmatrix} 1 \\ g(T) \end{bmatrix} = \frac{\sigma^2}{n(rC_r - B_r^2)}(C_r + 2g(T)B_r + (g(T))^2 r) \tag{A.4}$$

となる．

A.2 GEV モデルの場合の Fisher 情報量

GEV モデルの場合の Fisher 情報行列はかなり複雑になる．

上位 r 個の同時分布の Fisher 情報行列を

$$I_r(\boldsymbol{\theta}) = \begin{bmatrix} E_\theta\left\{-\dfrac{\partial^2 l}{\partial\mu^2}\right\} & E_\theta\left\{-\dfrac{\partial^2 l}{\partial\mu\partial\sigma}\right\} & E_\theta\left\{-\dfrac{\partial^2 l}{\partial\mu\partial\xi}\right\} \\ E_\theta\left\{-\dfrac{\partial^2 l}{\partial\mu\partial\sigma}\right\} & E_\theta\left\{-\dfrac{\partial^2 l}{\partial\sigma^2}\right\} & E_\theta\left\{-\dfrac{\partial^2 l}{\partial\sigma\partial\xi}\right\} \\ E_\theta\left\{-\dfrac{\partial^2 l}{\partial\mu\partial\xi}\right\} & E_\theta\left\{-\dfrac{\partial^2 l}{\partial\sigma\partial\xi}\right\} & E_\theta\left\{-\dfrac{\partial^2 l}{\partial\xi^2}\right\} \end{bmatrix} \tag{A.5}$$

とする，ここで $\boldsymbol{\theta} = (\mu,\, \sigma,\, \xi)$．このとき，Tawn (1988) より

$$E_{\boldsymbol{\theta}}\left\{-\frac{\partial^2 l}{\partial\mu^2}\right\}=\frac{(1+\xi)^2}{\sigma^2(1+2\xi)}G(2)\,,$$

$$E_{\boldsymbol{\theta}}\left\{-\frac{\partial^2 l}{\partial\mu\partial\sigma}\right\}=\frac{1}{\sigma^2\xi(1+2\xi)}\{(1+2\xi)G(1)-(1+\xi)^2G(2)\}\,,$$

$$E_{\boldsymbol{\theta}}\left\{-\frac{\partial^2 l}{\partial\mu\partial\xi}\right\}=\frac{1}{\sigma\xi^2(1+2\xi)}\left[(1+\xi)^2G(2)-(1+2\xi)G(1)\left\{\xi\psi(r+1+\xi)+\frac{1+\xi+\xi^2}{1+\xi}\right\}\right],$$

$$E_{\boldsymbol{\theta}}\left\{-\frac{\partial^2 l}{\partial\sigma^2}\right\}=\frac{1}{\sigma^2\xi^2(1+2\xi)}\left\{r(1+2\xi)-2(1+2\xi)G(1)+(1+\xi)^2G(2)\right\},$$

$$E_{\boldsymbol{\theta}}\left\{-\frac{\partial^2 l}{\partial\sigma\partial\xi}\right\}=\frac{1}{\sigma\xi^3(1+2\xi)}\left[(1+2\xi)G(1)\left\{\xi\psi(r+1+\xi)+\frac{1+(1+\xi)^2}{1+\xi}\right\}-r\xi(1+2\xi)\psi(r+1)\right.$$
$$\left.-(1+\xi)^2G(2)-r(1+2\xi)\right],$$

$$E_{\boldsymbol{\theta}}\left\{-\frac{\partial^2 l}{\partial\xi^2}\right\}=\frac{1}{\xi^4(1+2\xi)}\left[(1+\xi)^2G(2)-2(1+2\xi)G(1)\left\{\xi\psi(r+1+\xi)+\frac{1+\xi+\xi^2}{1+\xi}\right\}\right.$$
$$\left.+r(1+2\xi)\{1+2\xi\psi(r+1)+\xi^2[1+\psi'(r+1)+\psi^2(r+1)]\}\right],$$

ただし， $G(j)=G(j;\,r,\,\xi)=\Gamma(r+j\xi+1)/\Gamma(r)$**，** $j=1,2$**.**

この情報行列の各成分は ξ**，** r**，** σ **の関数で** μ **は含まれていない.**

サイズ n **の標本による最尤推定量** $\widehat{\boldsymbol{\theta}}_r=(\widehat{\mu}_r,\,\widehat{\sigma}_r,\,\widehat{\xi}_r)$ **の漸近分散行列は** $I_r^{-1}(\boldsymbol{\theta})/n$ **になる.**

T-return level $q(T)$ **の最尤推定量は**

$$\widehat{q}_r(T)=\widehat{\mu}_r+\widehat{\sigma}_r(y(T)^{-\widehat{\xi}_r}-1)/\widehat{\xi}_r\,,\quad y(T)=-\log(1-1/T)$$

で，その漸近分散はデルタ法より

$$AV(\widehat{q}_r(T))=(\nabla q)'\,\frac{1}{n}I_r^{-1}(\boldsymbol{\theta})\,\nabla q\,,$$

$$(\nabla q)'=[\,1,\,(y(T)^{-\xi}-1)/\xi,\,-\sigma(y(T)^{-\xi}-1)/\xi^2-\sigma(y(T)^{-\xi}\log y(T))/\xi\,]$$

となる．これは， ξ**，** r**，** T**，** σ**，** n **の関数になるから，**

$$AV(\widehat{q}_r(T))=V_\xi(r,\,T,\,\sigma)/n$$

とおく.

ここで， σ **に注意して** $I_r(\boldsymbol{\theta})$ **の逆行列を求め，** $AV(\widehat{q}_r(T))$ **を計算すると，**

$$AV(\widehat{q}_r(T))=\sigma^2V_\xi(r,\,T,\,1)/n \tag{A.6}$$

となることがわかる.

A.3 確率加重モーメント法

ここでは，Hosking et al.**(**1985**)の確率加重モーメント(**probability-weighted moment**，**PWM**)法による一般極値分布のパラメータ** $\boldsymbol{\theta}=(\mu,\,\sigma,\,\xi)$ **の推定について紹介する.**

一般に，確率変数 X **の分布関数が** $G(x)$ **のとき，その** PWM **は実数** p**，** r**，** s **にたいして**

$$M_{p,r,s}=E[X^p\{G(X)\}^r\{1-G(X)\}^s] \tag{A.7}$$

で定義される．これは， G **の逆関数を** G^{-1} **とするとき**

$$M_{p,r,s}=\int_0^1\{G^{-1}(u)\}^p u^r(1-u)^s du \tag{A.8}$$

と表される．

ここで，一般極値分布のパラメータの推定に便利な $\beta_r = M_{1,r,0} = E[X\{G(X)\}^r]$ $(r = 0, 1, 2)$ を考える．PWM β_r の推定は次の様にする．一般に，分布 G からの順序統計量を $X_{1:n} \geq X_{2:n} \geq \cdots \geq X_{n:n}$ とするとき，

$$\widehat{\beta}_r = \frac{1}{n}\sum_{i=1}^{n} \frac{(i-1)(i-2)\ldots(i-r)}{(n-1)(n-2)\cdots(n-r)} X_{n+1-i:n}$$

は β_r の不偏推定量になる．ところで，一般極値分布の場合はシミュレーション実験によると

$$b_r = \frac{1}{n}\sum_{i=1}^{n} \left(\frac{i-0.35}{n}\right)^r X_{n+1-i:n}$$

を用いた方がパラメータ θ の推定精度は良い．

一般極値分布で $\xi \neq 0$ のとき，

$$\beta_r = \frac{\mu + \sigma[(r+1)^{\xi}\,\Gamma(1-\xi) - 1]/\xi}{r+1}\,, \quad \xi < 1$$

となる．これから，一般極値分布のパラメータ μ，σ，ξ に関する次の連立方程式を得る：

$$\begin{cases} \beta_0 = \mu + \sigma\{\Gamma(1-\xi) - 1\}/\xi\,, \\ 2\beta_1 - \beta_0 = \sigma(2^{\xi} - 1)\Gamma(1-\xi)/\xi\,, \\ (3\beta_2 - \beta_0)/(2\beta_1 - \beta_0) = (3^{\xi} - 1)/(2^{\xi} - 1)\,. \end{cases}$$

PWM 推定量 $\widehat{\mu}$，$\widehat{\sigma}$，$\widehat{\xi}$ は β_r をその推定量 b_r で置き換えたときの方程式の解である．$\widehat{\xi}$ を求めるためには，

$$(3b_2 - b_0)/(2b_1 - b_0) = (3^{\xi} - 1)/(2^{\xi} - 1)$$

を解けばよいが，$-1/2 < \widehat{\xi} < 1/2$ の場合は

$$\widehat{\xi} = 7.8590c - 2.9554c^2\,, \quad c = \frac{\log 2}{\log 3} - \frac{2b_1 - b_0}{3b_2 - b_0}$$

とすれば，その誤差は 0.0009 以下である．この $\widehat{\xi}$ を用いて，

$$\widehat{\sigma} = \frac{(2b_1 - b_0)\widehat{\xi}}{(2^{\widehat{\xi}} - 1)\,\Gamma(1-\widehat{\xi})}\,, \quad \widehat{\mu} = b_0 - \widehat{\sigma}\{\Gamma(1-\widehat{\xi}) - 1\}/\widehat{\xi}$$

で PWM 推定量が求まる．

シミュレーション実験によると，PWM 推定量は小標本では良い精度を示す．

PWM 推定量 $\widehat{\xi}$ は $\xi = 0$ のとき漸近的に $\mathrm{N}(0,\ 0.5633/n)$ に従う．したがって，統計量 $Z = \widehat{\xi}\sqrt{n/0.5633}$ が近似的に標準正規分布 $\mathrm{N}(0,\ 1)$ に従う事を用いて，仮説 $\mathrm{H}_0 : \xi = 0$ の検定を行うことが出来る．シミュレーション実験によると，この統計量による検出力は良好である．

謝　辞

本研究の一部は高橋倫也が統計数理研究所客員教授として行ったものである．解析に用いた腐食データは東京ガス技術研究所より提供していただいた．ここに記して感謝いたします．

参 考 文 献

Coles, S. G. (2001). *An Introduction to Statistical Modeling of Extreme Values*, Springer-Verlag, London.

Coles, S. G. and Walshaw, D. (1994). Directional modelling of extreme wind speeds, *Applied Statistics*, **43**, 139–157.

David, H. A. and Nagaraja, H. N. (2003). *Order Statistics*, 3rd ed., John Wiley, Hoboken, New Jersey.

Dupuis, D. J. (1997). Extreme value theory based on the r largest annual events: A robust approach, *Journal of Hydrology*, **200**, 295–306.

Hosking, J. R. M., Wallis, J. R. and Wood, E. F. (1985). Estimation of the generalized extreme-value distribution by the method of probability-weighted moments, *Technometrics*, **27**, 251–261.

Nagaraja, H. N. (1982). Record values and extreme value distributions, *Journal of Applied Probability*, **19**, 233–239.

Palutikof, J. P., Brabson, B. B., Lister, D. H. and Adcock, S. T. (1999). A review of methods to calculate extreme wind speeds, *Meteorological Applications*, **6**, 119–132.

Robinson, M. E. and Tawn, J. A. (1995). Statistics for exceptional athletics records, *Applied Statistics*, **44**, 499–511.

Scarf, P. A. (1993). On the limiting joint distribution of the r deepest pit depths, *Applied Stochastic Models & Data Analysis*, **9**, 267–278.

Scarf, P. A. and Laycock, P. J. (1996). Estimation of extremes in corrosion engineering, *Journal of Applied Statistics*, **23**, 621–643.

Scarf, P. A., Cottis, R. A. and Laycock, P. J. (1992). Extrapolation of extreme pit depths in space and time using the r deepest pit depths, *Journal of the Electrochemical Society*, **139**, 2621–2627.

Smith, R. L. (1986). Extreme value theory based on the r largest annual events, *Journal of Hydrology*, **86**, 27–43.

Smith, R. L. (1997). Statistics for exceptional athletics records, Letter to the editors, *Applied Statistics*, **46**, 123–128.

Strand, M. and Boes, D. (1998). Modeling road racing times of competitive recreational runners using extreme value theory, *The American Statistician*, **52**, 205-210.

Tawn, J. A. (1988). An extreme-value theory model for dependent observations, *Journal of Hydrology*, **101**, 227–250.

Tsimplis, M. N. and Blackman, D. (1997). Extreme sea-level distribution and return periods in the Aegean and Ionian Seas, *Estuarine, Coastal and Shelf Science*, **44**, 79–89.

Weissman, I. (1978). Estimation of parameters and large quantiles based on the k largest observations, *Journal of American Statistical Association*, **73**, 812–815.

116 Proceedings of the Institute of Statistical Mathematics Vol. 52, No. 1, 93–116 (2004)

Estimation of Larger Quantiles Based on the r Largest Observations

Rinya Takahashi

(Kobe University of Mercantile Marine)

Masaaki Sibuya

(Takachiho University)

Assume that larger values are observed in n unit areas or intervals. To estimate quantiles of small upper tail probability, r largest values of n datasets are used. The asymptotic efficiency of the maximum likelihood estimates relative to that of $r = 1$, are shown in Tables. The depth of small pits caused by corrosion are analyzed along the discussions of the paper.

Key words: Asymptotic relative efficiency, asymptotic variance, generalized extreme value distribution, Gumbel distribution.