数理解析研究所講究録 1308 巻 2003 年 61-74

上位 r 個の観測値に基づく確率点の推定

神戸商船大学　高橋 倫也 (Rinya Takahashi)
Kobe University of Mercantile Marine
高千穂大学　渋谷 政昭 (Masaaki Sibuya)
Takachiho University

いくつかの単位期間または単位領域のそれぞれで測定した確率標本データの，各標本上位 r 個を用いる極値解析法について述べる．上側微小確率点推定の精度とその問題点について議論する．

1　はじめに

確率標本より，与えられた期間または領域での最大値を予測するために，いくつかの単位期間または単位領域ごとの最大値（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．その予測精度を上げる方法として「上位 r 個までのデータ」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，「推定精度がどの程度改善するか」と「この方法の問題点」について議論する．

Weissman (1978) により，極値理論に基づく上位 k 個のデータ (一組) を用いる上側微小確率点の推定が初めて議論された．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 Weissman (1978) を発展させ，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，ベニスの毎年上位 10 個の潮位データを位置パラメータに時間依存性を導入して解析し，この手法の有効性を示した．Tawn (1988) は，Smith (1986) の結果を一般極値 (GEV) モデルへ拡張し，Lowestoft の潮位データの解析を行った．Scarf et al. (1992) は，GEV モデルの下で位置と尺度パラメータに時間依存性を導入し，上位 r 個の孔食深さデータの解析を行い手法の有効性を示した．

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を述べる．

2　極値理論

一般極値 (GEV) 分布の標準型を

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

とする．ここで G_ξ は，$\xi < 0$ の場合は 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 GEV 分布 G_ξ の吸引領域に属す ($F \in \mathcal{D}(G_\xi)$) と仮定する：すなわち 適当な数列 $a_n > 0$, $b_n \in \mathbf{R}$, $n = 1, 2, \ldots$ が存在し，

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

このとき，上位 r 個の順序統計量の同時分布関数

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

は同時密度関数

$$g_{\xi,12\cdots r}(z_1, z_2, \ldots, z_r) = \frac{g_\xi(z_1)\cdots g_\xi(z_{r-1})}{G_\xi(z_1)\cdots G_\xi(z_{r-1})} g_\xi(z_r), \quad g_\xi(z) = dG_\xi(z)/dz, \tag{2.2}$$

を持つ分布関数 $G_{\xi,12\cdots r}$ に収束する．

分布 $G_{\xi,12\cdots r}$ に従う確率ベクトルを $(Z_1, Z_2, \ldots, Z_r)$ とする．このとき，Z_j, $j \geq 1$ の周辺分布関数，$G_{\xi,j}$, は r によらず，

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

で，その周辺密度関数，$g_{\xi,j}$, は

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

となる．

定理 1．$(Z_1, Z_2, \ldots, Z_r)$ は次の性質をもつ．ただし $E_1, E_2, \ldots$ を標準指数分布 (Exp(1)) に従う独立確率変数，$S_j = \sum_{k=1}^{j} E_k$ とする．

(A) GEV ($\xi \neq 0$) モデルの場合：

1) $E_\xi(Z_j) = \Big(\Gamma(j-\xi) - \Gamma(j)\Big)\Big/(\xi\Gamma(j))$, $V_\xi(Z_j) = \Big(\Gamma(j-2\xi)\Gamma(j) - \Gamma^2(j-\xi)\Big)\Big/(\xi^2\Gamma^2(j))$.

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

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

(B) Gumbel ($\xi = 0$) モデルの場合：

1) $E_0(Z_j) = -\psi(j)$, $V_0(Z_j) = \psi'(j)$.

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

3) $\{Z_j, j \geq 1\} \stackrel{d}{=} \{-\log S_j, j \geq 1\}$.

ただし，Γ はガンマ関数，ψ はディ・ガンマ関数，ψ' はトリ・ガンマ関数である．

注 1．(連続性) **(A)** で $\xi \to 0$ とすれば **(B)** が得られる．

注2． **(A)** 3)，**(B)** 3) は Nagaraja (1982)，**(B)** 1) は Weissman (1978) 参照．

注3． 変数が正整数の場合，ディ・ガンマ関数とトリ・ガンマ関数の値は，

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

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

注4． $\mathrm{Cor}_0(Z_j, Z_{j+1})$ は j に関して狭義の増加関数で 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．

注5． **(B)** 3) より，Gumbel モデルの下では

$$j(Z_j - Z_{j+1}) = j \log \frac{S_{j+1}}{S_j}, \quad j = 1, 2, 3, \ldots \tag{2.5}$$

は互いに独立に Exp(1) に従う (Weissman, 1978)．従って，GEV ($\xi \neq 0$) モデルの場合は

$$\frac{j}{\xi} \log \frac{1 + \xi Z_j}{1 + \xi Z_{j+1}}, \quad j = 1, 2, 3, \ldots \tag{2.6}$$

が互いに独立に Exp(1) に従う (Tawn, 1988)．

注6． 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

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

とすると，

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

と表される．ただし，$1 \geq U_{1:n} \geq U_{2:n} \geq \cdots \geq U_{n:n} \geq 0$ は一様分布 U(0, 1) からの n 個の順序統計量である．この $Y_{j:n}$ と **(A)** 3) の $Z_j \stackrel{d}{=} (S_j^{-\xi} - 1)/\xi$ 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 (Tawn, 1988)．一方，Gumbel モデルの下では，$(Z_1, Z_2, \ldots, Z_r)$ は指数分布からの上位 r 個の順序統計量と見なせる (Weissman, 1978)．

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．

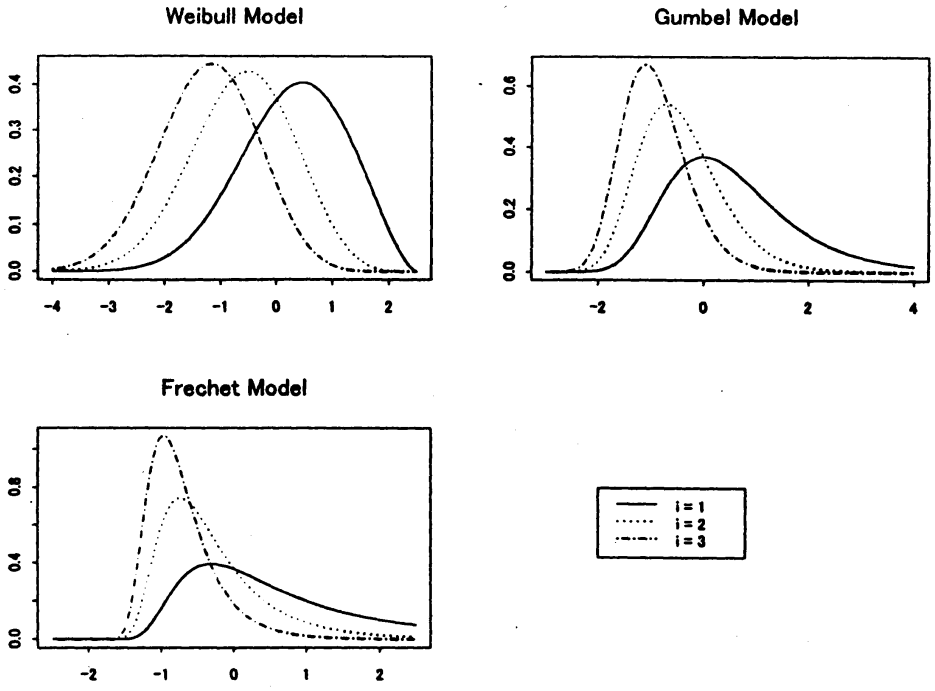


図1．$\xi = -0.4,\ 0,\ 0.4$ の場合の Z_j の周辺密度関数，$j = 1, 2, 3$.

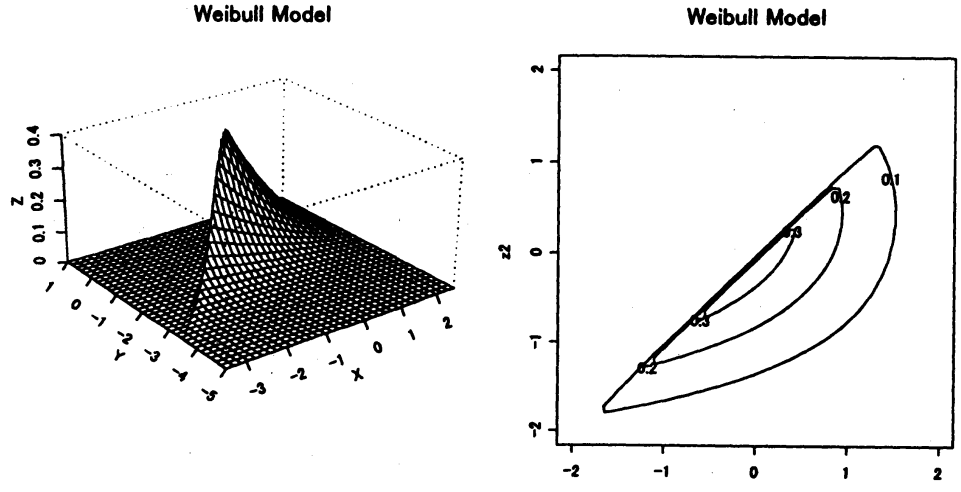


図2．$\xi = -0.4$ の場合の (Z_1, Z_2) の同時密度関数とその等高線.

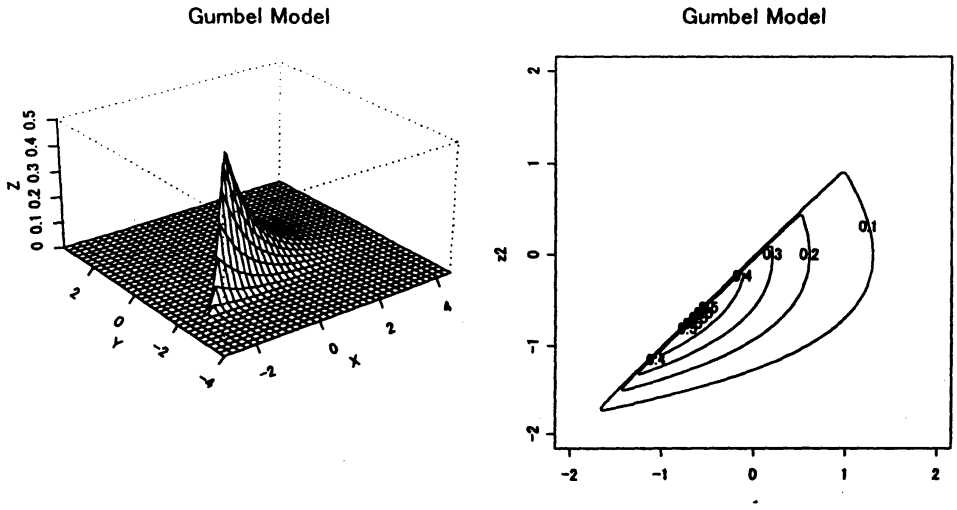


図3．$\xi = 0$ の場合の (Z_1, Z_2) の同時密度関数とその等高線.

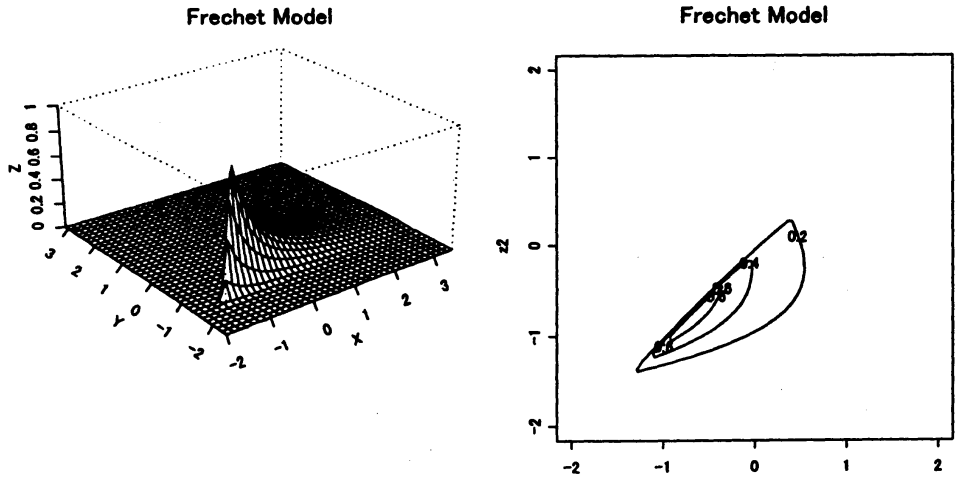


図4．$\xi = 0.4$ の場合の (Z_1, Z_2) の同時密度関数とその等高線.

3　パラメータ推定

パラメータ $\boldsymbol{\theta} = (\mu, \sigma)$ または (μ, σ, ξ) と，極値理論で重要な上側微少確率点 T-return level $q(T)$,

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

すなわち,

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

の推定法について述べる．これらのパラメータが補助変量の関数である場合に拡張することもで

3.1 モデル

上位 r 個の確率ベクトル $(X_1, X_2, \ldots, X_r)$ の従う結合密度関数は Gumbel モデルの場合は，$\boldsymbol{\theta} = (\mu, \sigma)$ として，

$$g(x_1, x_2, \ldots, x_r; \boldsymbol{\theta}) = \frac{1}{\sigma^r} \exp\left[-\sum_{j=1}^{r}\left(\frac{x_j - \mu}{\sigma}\right) - \exp\left(-\frac{x_r - \mu}{\sigma}\right)\right], \tag{3.2}$$

$$x_1 \geq x_2 \geq \cdots \geq x_r,$$

で，GEV モデルの場合は $\boldsymbol{\theta} = (\mu, \sigma, \xi)$ として，

$$g(x_1, \ldots, x_r; \boldsymbol{\theta}) = \frac{1}{\sigma^r} \exp\left\{-\left(\frac{1}{\xi}+1\right)\sum_{j=1}^{r}\log\left[1+\xi\left(\frac{x_j-\mu}{\sigma}\right)\right] - \left[1+\xi\left(\frac{x_r-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}, \tag{3.3}$$

$$x_1 \geq \cdots \geq x_r, \quad 1+\xi(x_j-\mu)/\sigma > 0, \ j = 1, \ldots, r,$$

となる．

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

$$\begin{cases} x_{11} \geq x_{12} \geq \cdots \geq x_{1r}, \\ x_{21} \geq x_{22} \geq \cdots \geq x_{2r}, \\ \qquad \cdots \\ x_{i1} \geq x_{i2} \geq \cdots \geq x_{ir}, \\ \qquad \cdots \\ x_{n1} \geq x_{n2} \geq \cdots \geq x_{nr}, \end{cases}$$

の $r \times n$ 個の数値とする．

3.2 Gumbel モデル

Gumbel モデルの場合の対数尤度は

$$l(\mu, \sigma) = -nr\log\sigma - \sum_{i=1}^{n}\left[\sum_{j=1}^{r}\left(\frac{x_{ij}-\mu}{\sigma}\right) + \exp\left(-\frac{x_{ir}-\mu}{\sigma}\right)\right], \tag{3.4}$$

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

$$\begin{cases} \dfrac{\partial}{\partial\mu} l(\mu, \sigma) &= -\displaystyle\sum_{i=1}^{n}\left[-\frac{r}{\sigma} + \frac{1}{\sigma}\exp\left(-\frac{x_{ir}-\mu}{\sigma}\right)\right] = 0, \\ \dfrac{\partial}{\partial\sigma} l(\mu, \sigma) &= -\dfrac{nr}{\sigma} + \displaystyle\sum_{i=1}^{n}\left[\sum_{j=1}^{r}\left(\frac{x_{ij}-\mu}{\sigma^2}\right) - \frac{x_{ir}-\mu}{\sigma^2}\exp\left(-\frac{x_{ir}-\mu}{\sigma}\right)\right] = 0. \end{cases} \tag{3.5}$$

この連立非線形方程式から，

$$h_r(\sigma) := \sum_{i=1}^{n}\left(\frac{x_{ir}-\bar{\bar{x}}_r}{\sigma}+1\right)\exp\left(-\frac{x_{ir}-\bar{\bar{x}}_r}{\sigma}\right) = 0, \quad \bar{\bar{x}}_r = \frac{1}{nr}\sum_{i=1}^{n}\sum_{j=1}^{r} x_{ij} \tag{3.6}$$

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 $\hat{\sigma}_r$ を求める．これから，

$$\hat{\mu}_r = -\hat{\sigma}_r \log\left[\frac{1}{nr}\sum_{i=1}^{n}\exp\left(-\frac{x_{ir}}{\hat{\sigma}_r}\right)\right] \tag{3.7}$$

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

$$\hat{q}_r(T) = \hat{\mu}_r + \hat{\sigma}_r\Big\{-\log(-\log(1-1/T))\Big\}, \tag{3.8}$$

とすればよい．また，推定値 $\hat{\mu}_r$，$\hat{\sigma}_r$，$\hat{q}_r(T)$ の標準誤差は付録 A1 の漸近分散行列を用いて推定する．

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 は次の周辺分布関数

$$G_{0,j}(z) = P\left\{\frac{X_j-\mu}{\sigma} \le z\right\} = \sum_{k=0}^{j-1}\exp(-kz - e^{-z})/k!$$

を持つ．従って，

$$U_{ij} = G_{0,j}\left(\frac{X_{ij}-\mu}{\sigma}\right) \tag{3.9}$$

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 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 $\hat{\mu}_r$，$\hat{\sigma}_r$ を求める．これらを用いて，上位 j 番目のデータ $\{x_{1j}, x_{2j}, \ldots, x_{nj}\}$ ($j = 1, 2, \ldots, r$) から $u_{ij} = G_{0,j}((x_{ij}-\hat{\mu}_r)/\hat{\sigma}_r)$, $i = 1, 2, \ldots, n$ を求める．u_{ij} の順序統計量を $u_{(1)j} \ge u_{(2)j} \ge \cdots \ge u_{(n)j}$ 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．ここで，r 個すべての PP plot が直線性を示していると見なせる最大の r を決定する．(Smith, 1986 参照．)

一方，次の決定法も考えられる．

QQ plot：r を固定する．$j = 1$ の場合は通常の Gumbel 確率紙を考える．$j = 2, 3, \ldots, r$ の場合，上位 j 番目のデータ $\{x_{1j}, x_{2j}, \ldots, x_{nj}\}$ の順序統計量を $x_{(1)j} \ge x_{(2)j} \ge \cdots \ge x_{(n)j}$ とする．数値計算で確率点 $q_{(i)j} = G_{0,j}^{-1}(1 - i/(n+1))$ を求め，

$$(x_{(i)j},\ q_{(i)j}), \quad i = 1, 2, \ldots, n,$$

をプロットし r 個の QQ plot を作成する．すべての r 個の QQ plot が直線性を示していると見なせる最大の r を決定する．

これら PP，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，2 節の注 5 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 $j(<r)$ の指数確率紙で直線性を示していると見なせる最大の r を決定する．

従って，PP または QQ と指数確率紙をプロットし得られた r の中で最小のものを採用すればよい．

3.3 GEV モデル

GEV モデルの場合の対数尤度は

$$l(\mu,\sigma,\xi)=-nr\log\sigma-\sum_{i=1}^{n}\left\{\left(\frac{1}{\xi}+1\right)\sum_{j=1}^{r}\log\left[1+\xi\left(\frac{x_{ij}-\mu}{\sigma}\right)\right]+\left[1+\xi\left(\frac{x_{ir}-\mu}{\sigma}\right)\right]^{-1/\xi}\right\} \tag{3.10}$$

となる．

パラメータ $\boldsymbol{\theta}=(\mu,\sigma,\xi)$ を最尤法で推定する．尤度方程式は簡単にははならない．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 (Hosking et al., 1985) で求めた推定値を用いればよい．得られた最尤推定値を $\hat{\boldsymbol{\theta}}_r=(\hat{\mu}_r,\hat{\sigma}_r,\hat{\xi}_r)$ とすると T-return level $q(T)$ の推定は，

$$\hat{q}_r(T)=\hat{\mu}_r+\hat{\sigma}_r\Big\{(-\log(1-1/T))^{-\hat{\xi}_r}-1\Big\}\Big/\hat{\xi}_r \tag{3.11}$$

とすればよい．また，推定値 $\hat{\mu}_r$，$\hat{\sigma}_r$，$\hat{\xi}_r$，$\hat{q}_r(T)$ の標準誤差は付録 A2 の Fisher 情報行列を用いて推定する．

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．

GEV モデルの下で，X_j は次の周辺分布関数

$$G_{\xi,j}(z)=P\left\{\frac{X_j-\mu}{\sigma}\le z\right\}=\sum_{k=0}^{j-1}(1+\xi z)^{-k/\xi}\exp\{-(1+\xi z)^{-1/\xi}\}/k!$$

を持つ．従って，

$$U_{ij}=G_{\xi,j}\left(\frac{X_{ij}-\mu}{\sigma}\right) \tag{3.12}$$

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 $\hat{\mu}_r$，$\hat{\sigma}_r$，$\hat{\xi}_r$ を求める．これらを用いて，上位 j 番目のデータ $\{x_{1j},x_{2j},\ldots,x_{nj}\}$ $(j=1,2,\ldots,r)$ から $u_{ij}=G_{\hat{\xi}_r,j}((x_{ij}-\hat{\mu}_r)/\hat{\sigma}_r)$，$i=1,2,\ldots,n$ を求める．u_{ij} の順序統計量を $u_{(1)j}\ge u_{(2)j}\ge\cdots\ge u_{(n)j}$ 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．ここで，r 個すべての PP plot が直線性を示していると見なせる最大の r を決定する．(Tawn, 1988 参照．)

また，次の方法も考えられる．

QQ plot：r を固定し，上位 r 個のデータから形状パラメータの推定値 $\widehat{\xi}_r$ を求める．上位 j 番目のデータ $\{x_{1j}, x_{2j}, \ldots, x_{nj}\}$ $(j = 1, 2, \ldots, r)$ の順序統計量を $x_{(1)j} \geq x_{(2)j} \geq \cdots \geq x_{(n)j}$ とする．数値計算で確率点 $q_{(i)j} = G^{-1}_{\widehat{\xi}_r, j}(1 - i/(n+1))$ を求め，

$$(x_{(i)j},\ q_{(i)j}),\quad i = 1, 2, \ldots, n,$$

をプロットし r 個の QQ plot を作成する．ここで，r 個すべての QQ plot が直線性を示しているとみなせる最大の r を決定する．

分布の同時性をチェックする方法として，2 節の注 5 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 番目のデータ $\{x_{1j}, x_{2j}, \ldots, x_{nj}\}$ $(j = 1, 2, \ldots, r-1)$ に対して，

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

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．ここで $r-1$ 個すべての指数確率紙で直線性を示していると見なせる最大の r を決定する．

従って，PP または QQ と指数確率紙をプロットし得られた r の中の最小のものを採用すればよい．

4 有効性

ここでは，Gumbel モデルの下で上位 $r(\geq 2)$ 個のデータを用いる有効性について議論する．一般の GEV モデルの下での議論は難しいが，以下の議論は $\xi \doteq 0$ の場合にも近似的に成り立つと考えられる．

上位 r 個のデータを用いる場合の T-return level $q(T)$ の推定量の有効性について調べる．

$q(T)$ の推定を行う場合の漸近分散の比で上位 r 個を用いる有効性を示す．推定量 $\widehat{q}_r(T)$ の漸近分散は，付録 A1 の $(A1.4)$ より

$$AV(\widehat{q}_r(T)) = \frac{\sigma^2}{n(rC_r - B_r^2)}\Big[C_r + (g(T))^2 r + 2g(T)B_r\Big]$$

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

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

と表される．これは，T と r の関数である，T =100，1,000，10,000，100,000 と $r = 2, 3, \ldots, 10$ に対する $e(\widehat{q}_r(T), \widehat{q}_1(T))$ の値は次の表の様になる：

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

この表より，例えば 10,000-return level を推定するとき，100 個の極値データは 50 組の上位 3 個のデータとほぼ同じ精度，ほぼ等しい漸近分散，を持つと言うことが出来る．

ここでの議論は，「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 実データ解析

ここでは「上位 3 個までの孔食深さ測定」データの解析を行う．データは面積が等しい 30 個の領域内において上位 3 個の孔食深さを測定したものである．

図 5 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 である．

以下，Gumbel と GEV モデルの下で解析した結果を紹介する．

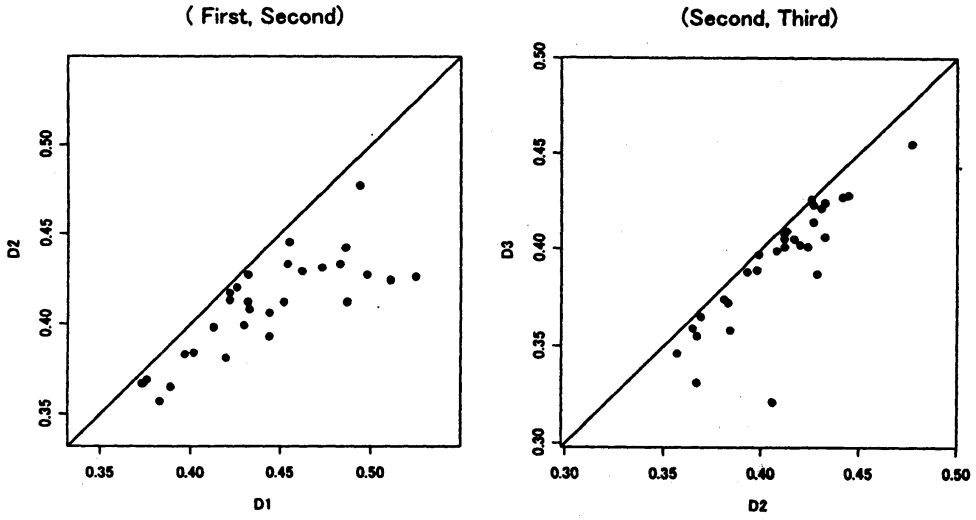


図 5．(a) データの（上位 1 位，上位 2 位）．(b) データの（上位 2 位，上位 3 位）．

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，GEV 分布で形状パラメータ ξ について，検定 $H_0 : \xi = 0$, $H_1 : \xi \neq 0$ を PWM 推定値に基づく方法 (Hosking et al., 1985) 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 であった．

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：

	$r = 1,$	2,	3
$\widehat{\mu}_r$	0.420	0.423	0.431
$\widehat{\sigma}_r$	0.037	0.036	0.045

r の決定を考える．QQ plot を書かせたのが図 6 で，指数確率紙が図 7 である．これらの図より，このデータでは上位 2 個まで使えると考えられる．

従って，T-return level の推定は次の様になる：

$$\widehat{q}_2(T) = 0.423 + 0.036\{ -\log(-\log(1 - 1/T))\}.$$

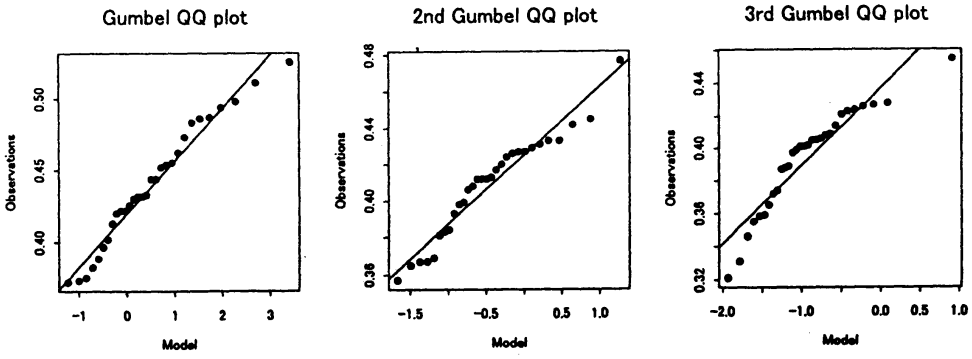


図 6．Gumbel モデルの下での QQ plot.

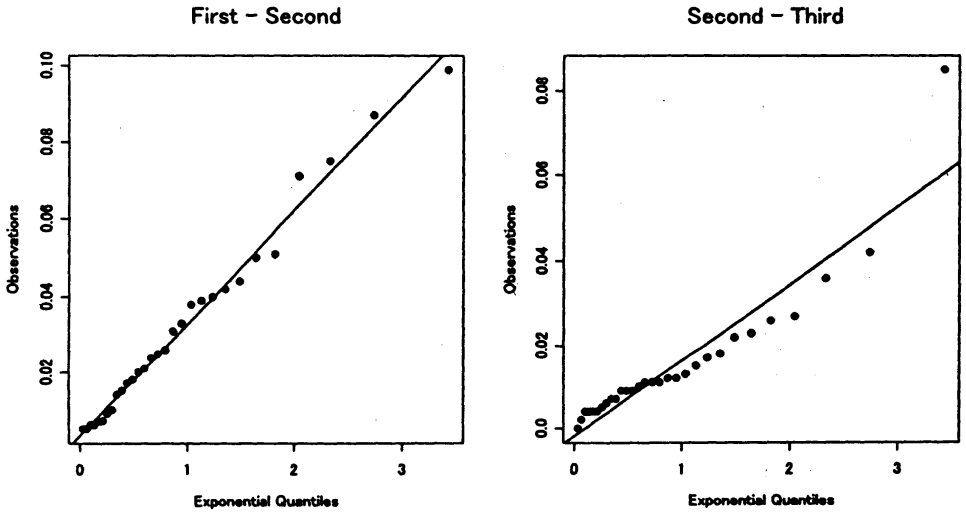


図 7．Gumbel モデルの下での指数確率紙.

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：

	$r = 1,$	2,	3
$\widehat{\mu}_r$	0.425	0.426	0.432
$\widehat{\sigma}_r$	0.035	0.034	0.036
$\widehat{\xi}_r$	−0.378	−0.156	−0.280

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 分布と見なせる．

r の決定のために PP plot を書かせたのが図 8，9，10 で，指数確率紙が図 11 である．これらから，このモデルの下でも Gumbel モデルの場合と同様に上位 2 個までのデータが使えると言える．

従って，この場合の T-return level の推定は

$$\widehat{q}_2(T) = 0.426 + 0.034\{(-\log(1 - 1/T))^{0.156} - 1\}/(-0.156)$$

とすればよい．また，最大孔食深さの上限値は 0.646 と推定される．

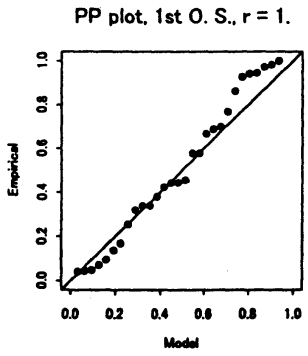
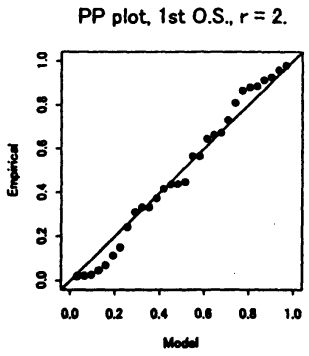


図8．GEV モデルの下での PP plot，$r = 1$.



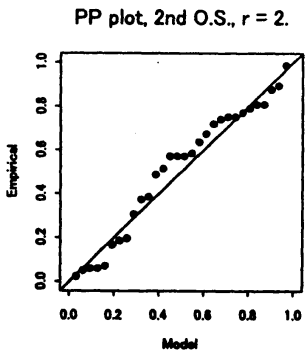
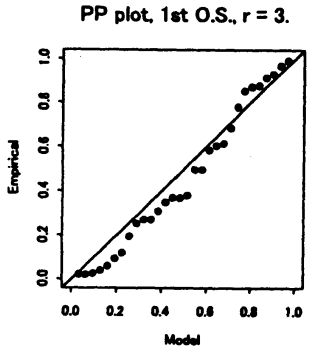
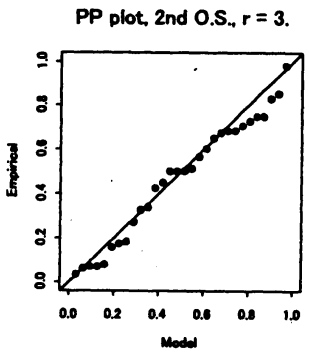


図9．GEV モデルの下での PP plot，$r = 2$.





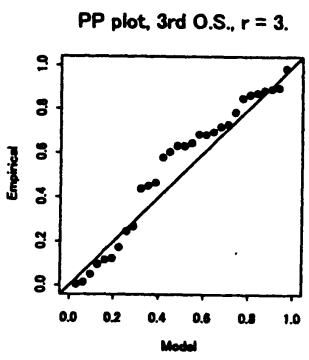
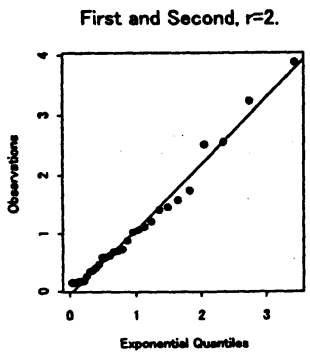
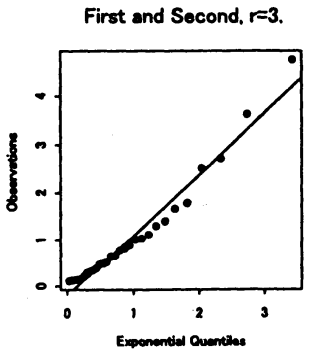


図10．GEV モデルの下での PP plot，$r = 3$.





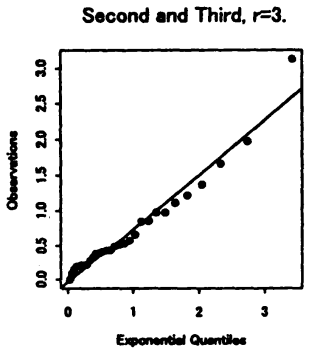


図11．GEV モデルの下での指数確率紙，$r = 2$，$r = 3$.

付録

A1. Gumbel モデルの場合の Fisher 情報量, Smith (1986)

サイズ n の上位 r 個の同時分布の Fisher 情報行列は

$$I_F(\boldsymbol{\theta}) = \frac{n}{\sigma^2}\begin{pmatrix} r & -B_r \\ -B_r & C_r \end{pmatrix} \tag{A1.1}$$

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

$$B_r = r\psi(r+1), \quad C_r = r\Big\{\psi^2(r+1) + \psi'(r+1) + 1\Big\}. \tag{A1.2}$$

逆行列は

$$\frac{\sigma^2}{n(rC_r - B_r^2)}\begin{pmatrix} C_r & B_r \\ B_r & r \end{pmatrix} \tag{A1.3}$$

と表される. これが, 最尤推定量 $\hat{\boldsymbol{\theta}}_r = (\hat{\mu}_r, \hat{\sigma}_r)$ の漸近分散行列である.

ここで, T-return level $q(T)$ の推定は,

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

とすればよかった. 推定量 $\hat{q}_r(T)$ の漸近分散は

$$\begin{aligned} AV(\hat{q}_r(T)) &= AV(\hat{\mu}_r) + (g(T))^2 AV(\hat{\sigma}_r) + 2g(T)\mathrm{ACov}(\hat{\mu}_r, \hat{\sigma}_r) \\ &= \frac{\sigma^2}{n(rC_r - B_r^2)}\Big[C_r + (g(T))^2 r + 2g(T)B_r\Big] \end{aligned} \tag{A1.4}$$

となる.

A2. GEV モデルの場合の Fisher 情報量, Tawn (1988)

GEV モデルの場合の Fisher 情報行列はかなり複雑になる.

サイズ n の上位 r 個の同時分布の Fisher 情報行列を

$$I_F(\boldsymbol{\theta}) = n\begin{pmatrix} E_{\boldsymbol{\theta}}\left\{-\frac{\partial^2 l}{\partial\mu^2}\right\} & E_{\boldsymbol{\theta}}\left\{-\frac{\partial^2 l}{\partial\mu\partial\sigma}\right\} & E_{\boldsymbol{\theta}}\left\{-\frac{\partial^2 l}{\partial\mu\partial\xi}\right\} \\ E_{\boldsymbol{\theta}}\left\{-\frac{\partial^2 l}{\partial\mu\partial\sigma}\right\} & E_{\boldsymbol{\theta}}\left\{-\frac{\partial^2 l}{\partial\sigma^2}\right\} & E_{\boldsymbol{\theta}}\left\{-\frac{\partial^2 l}{\partial\sigma\partial\xi}\right\} \\ E_{\boldsymbol{\theta}}\left\{-\frac{\partial^2 l}{\partial\mu\partial\xi}\right\} & E_{\boldsymbol{\theta}}\left\{-\frac{\partial^2 l}{\partial\sigma\partial\xi}\right\} & E_{\boldsymbol{\theta}}\left\{-\frac{\partial^2 l}{\partial\xi^2}\right\} \end{pmatrix} \tag{A2.1}$$

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

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

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

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

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

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

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

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

参考文献

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

Nagaraja, H. N. (1982). Record values and extreme value distributions. *J. Appl. Probab.*, **19**, 233–239.

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. *J. Electrochem. Soc.*, **139**, 2621–2627.

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

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

Weissman, I. (1978). Estimation of parameters and large quantiles based on the k largest observations. *J. Amer. Statist. Assoc.* **73**, 812–815.