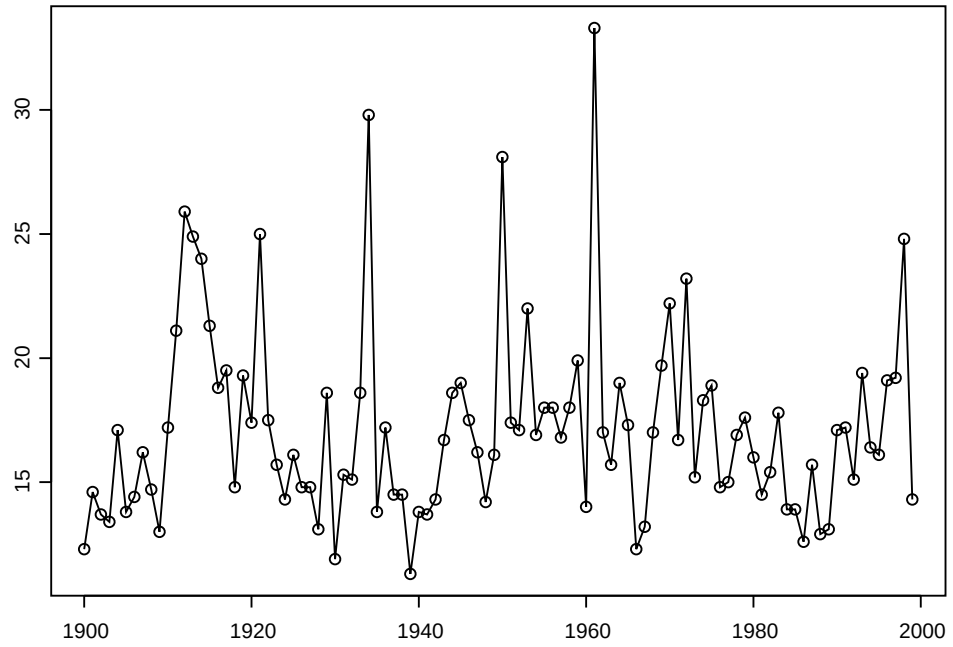
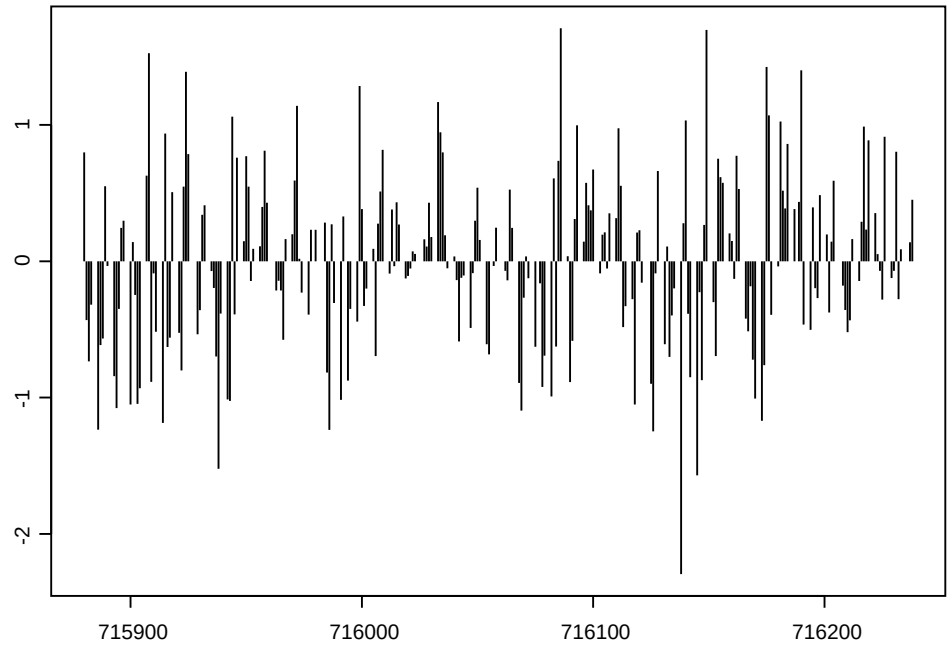
極値統計学

高橋 倫也 (神戸大学・海事科学部)
r-taka@maritime.kobe-u.ac.jp

IBIS2009　　平成２１年１０月１９日

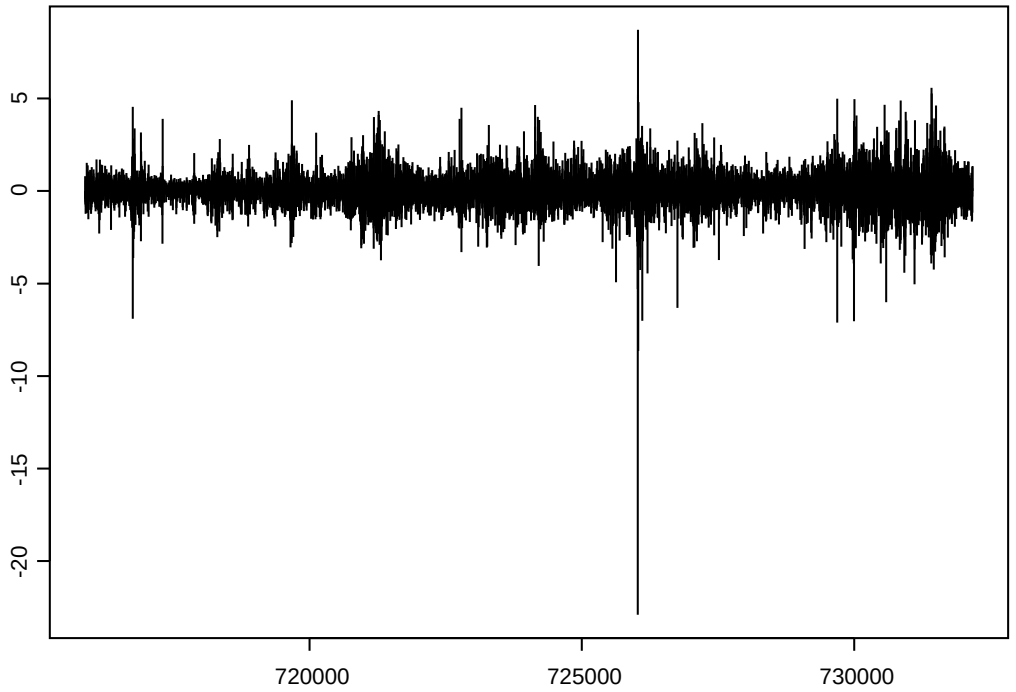


大阪の年最大風速(m/s)，1900年～1999年.



Daily returns of the S&P 500 index. 1960年

Gilli & Këllezi (2006). Computational Econometrics 27(1).



Daily returns of the S&P 500 index.
1960年1月5日～2004年8月16日（11270データ）

目的：リスクを評価し対処するため
　　長期間（または広領域）でどの様な大きな値が発生するか予測したい.

統計学：「期間と確率」を指定　→　「どれくらい大きな値が生起するか」の推測

内容

１．はじめに

統計学で用いる分布とその基礎理論

中心極限定理	**数理統計学**	正規分布
最弱リンクモデル	**信頼性理論**	ワイブル分布
極値理論	**極値統計学**	一般極値分布，　一般パレート分布

○ **極値統計学の目的**

長期間（または広領域等）における最大値（または最小値）に関する推測.

○ 最大値に関する情報を有しているデータ

区分最大値データ（block maxima)

標本を同じ大きさのブロックに分けたときの各ブロックごとの最大値データ.

水準超過データ(threshold exceedances or Peaks Over Threshold)

標本の中で十分大きな水準（閾値）を超えるデータ.

極値理論により，区分最大値データには一般極値分布を，また水準超過データには一般パレート分布を適合させる．通常パラメータは最尤法で推定.

極値統計学で興味がある量

一般極値分布や一般パレート分布のパラメータではなく，それらの分布の上側微小確率点（再現レベルとバリューアットリスクとよばれるもの）．

２．極値理論

○ 分布関数 F を持つ確率変数 X を考える.
　分布 F からの確率標本を

$$X_1, X_2, \ldots, X_n$$

とする.

極値統計量

$$Z_n = \max\left\{X_1, X_2, \ldots, X_n\right\} = \max_{1 \le i \le n} X_i$$

を考える.

○ n を大きくしていくと Z_n は分布 F の上限 x_F へ収束：

$$Z_n \to x_F = \sup\{x : F(x) < 1\} \le \infty, \quad n \to \infty.$$

○ Z_n を基準化：分布 F に依存する数列 $a_n > 0,\ b_n \in \mathbb{R}$ $(n = 1, 2, \ldots)$ と退化していない分布 $G(x)$ を持つ確率変数 Z が存在して,

$$\frac{Z_n - b_n}{a_n} \xrightarrow{\mathcal{L}} Z, \quad n \to \infty.$$

すなわち

$$P\left(\frac{Z_n - b_n}{a_n} \leq x\right) \to P(Z \leq x) = G(x)$$

となる場合を考える.

G：**極値分布**(extreme value distribution), 分布 F は極値分布 G の**吸引領域 (domain of attraction) に属する**（$F \in \mathcal{D}(G)$ ）. $(a_n,\ b_n)$：**基準化係数** .

基本的な問題

(1) 極値分布 $G(x)$ はどのような分布か？

(2) $F \in \mathcal{D}(G)$ となるための必要十分条件？

注意. 以下では最大の場合のみを議論する.

最小 $W_n = \min_{1 \le i \le n} X_i$ に関する結果は，関係式

$$W_n = \min\left\{X_1, X_2, \ldots, X_n\right\} = -\max\left\{-X_1, -X_2, \ldots, -X_n\right\}$$

より，最大の場合からすぐ得られる．例えば，最小の場合の極値分布 W は，G を（対応する）最大の極値分布とすると $W(x) = 1 - G(-x)$ で求まる.

○ 基準化した極値統計量の分布関数：

$$
\begin{aligned}
P\left(\frac{Z_n - b_n}{a_n} \le x\right) &= P(Z_n \le a_n x + b_n) = P\left(\max_{1\le i\le n} X_i \le a_n x + b_n\right) \\
&= P(X_i \le a_n x + b_n,\ \ i = 1, 2, \ldots, n) \\
&= \prod_{i=1}^{n} P(X_i \le a_n x + b_n) = \prod_{i=1}^{n} F(a_n x + b_n) \\
&= F^n(a_n x + b_n).
\end{aligned}
$$

極値分布　よく知られている分布から３つのタイプの極値分布の導出

具体的な分布の極値統計量の漸近分布（極値分布）
F：分布関数，$f(x)$：密度関数

例１． 標準指数分布 Exp(1)

$$F(x) = 1 - e^{-x}, \quad f(x) = e^{-x}, \ x \geq 0.$$

$$F^n(x + \log n) = \left\{1 - e^{-(x+\log n)}\right\}^n = \left\{1 + \frac{-e^{-x}}{n}\right\}^n$$

$$\to \quad e^{-e^{-x}} = \exp(-\exp(-x)), \ n \to \infty.$$

$$(-\infty < x < \infty).$$

$a_n = 1$，$b_n = \log n$である．

例２． Pareto分布　$(\forall\, \alpha > 0)$

$$F(x) = 1 - \frac{1}{x^{\alpha}}, \quad f(x) = \alpha x^{-\alpha-1}, \ \ x \geq 1.$$

$$\begin{aligned} F^n(n^{1/\alpha}x) &= \left\{1 - \frac{1}{(n^{1/\alpha}x)^{\alpha}}\right\}^n = \left\{1 + \frac{-x^{-\alpha}}{n}\right\}^n \\ &\to \ e^{-x^{-\alpha}} = \exp(-x^{-\alpha}), \ \ n \to \infty. \quad (x > 0). \end{aligned}$$

$a_n = n^{1/\alpha}$, $b_n = 0$である.

例３． ベータ分布　$(\forall\, \alpha > 0)$

$$F(x) = 1 - (1-x)^{\alpha}, \quad f(x) = \alpha(1-x)^{\alpha-1}, \ \ 0 \le x \le 1.$$

$$\begin{aligned} F^n(n^{-1/\alpha}x + 1) &= \left\{1 - (-n^{-1/\alpha}x)^{\alpha}\right\}^n = \left\{1 + \frac{-(-x)^{\alpha}}{n}\right\}^n \\ &\to \ e^{-(-x)^{\alpha}} = \exp(-(-x)^{\alpha}), \ \ n \to \infty. \quad (x \le 0). \end{aligned}$$

$a_n = n^{-1/\alpha}$, $b_n = 1 = x_F$である.

定理１. [Fréchet (1927), Fisher and Tippett (1928), Gnedenko (1943); “ Trinity Theorem”] 極値分布 $G(x)$ は次の３つの型に限る：

$$\text{Gumbel分布：} \quad \Lambda(x) = \exp(-\exp(-x)), \qquad x \in \mathbb{R},$$

$$\text{Fréchet分布：} \quad \Phi_\alpha(x) = \begin{cases} 0, & x \le 0, \\ \exp(-x^{-\alpha}), & x > 0, \end{cases} \qquad \alpha > 0,$$

$$\text{(負の)Weibull分布：} \quad \Psi_\alpha(x) = \begin{cases} \exp(-(-x)^{\alpha}), & x \le 0, \\ 1, & x > 0, \end{cases} \qquad \alpha > 0.$$

吸引領域　　吸引領域に属するための必要十分条件：

定理２． [Gnedenko (1943), de Haan (1970)]

$$F \in \mathcal{D}(\Phi_\alpha) \iff x_F = \infty \ \& \ \lim_{x\to\infty} \frac{1-F(tx)}{1-F(x)} = t^{-\alpha}, \ \forall t > 0.$$

$$F \in \mathcal{D}(\Psi_\alpha) \iff x_F < \infty \ \& \ \lim_{x\uparrow x_F} \frac{1-F(x_F-(x_F-x)t)}{1-F(x)} = t^{\alpha}, \ \forall t > 0.$$

$$F \in \mathcal{D}(\Lambda) \iff \exists\, s(\,\cdot\,) > 0 \ \text{ s.t. } \ (*) \ \lim_{x\uparrow x_F} \frac{1-F(x+ts(x))}{1-F(x)} = e^{-t}.$$

$$(*) \implies \int_x^{x_F} (1-F(y))dy < \infty, \quad x < x_F$$

$$\implies s(x) = \int_x^{x_F} (1-F(y))dy/(1-F(x)) \quad \text{satisfies } (*).$$

○ 3つの極値分布の von Mises-Jenkinson 表現 (1936, 1955)：

$$G_\xi(x) = \exp\Big\{-(1+\xi x)^{-1/\xi}\Big\}, \quad 1+\xi x > 0.$$

これを**一般極値 (generalized extreme value, GEV) 分布**とよぶ.

パラメータ$\xi \in \mathbb{R}$は形状パラメータで, $\xi = 0$の場合は

$$G_0(x) = \lim_{\xi \to 0} G_\xi(x) = \exp\{-\exp(-x)\} = \Lambda(x)$$

とする. また,

$$\Phi_\alpha(x) = G_{1/\alpha}(\alpha(x-1)), \quad \Psi_\alpha(x) = G_{-1/\alpha}(\alpha(x+1)).$$

○ 一般極値分布は**最大値安定（max–stable）性**という性質を持つ：

任意のnに対して$A_n > 0$と$B_n \in \mathbb{R}$が存在して

$$G_\xi^n(A_n x + B_n) = G_\xi(x)$$

となる.

この関数方程式の解が定理１の３つの型の極値分布になる.

一般極値分布を用いると,

$$F^n(a_n x + b_n) \to G_\xi(x).$$

よってnが十分大のとき，$a_n x + b_n = z$とおくと

$$P(Z_n \leq z) = F^n(z) \approx G_\xi\left(\frac{z - b_n}{a_n}\right).$$

問題：G_ξによる近似が十分であるためにはnがどれくらい大きければよいか？

$$
\begin{aligned}
F^n(a_n x + b_n) &= \Big[1 - \Big\{1 - F(a_n x + b_n)\Big\}\Big]^n \\
&= \Big[1 - n\Big\{1 - F(a_n x + b_n)\Big\}\Big/ n\Big]^n \\
&\to \exp\Big[- \lim_{n\to\infty} n\Big\{1 - F(a_n x + b_n)\Big\}\Big]
\end{aligned}
$$

から,

$$
\lim_{n\to\infty} n\{1 - F(a_n x + b_n)\} = -\log G_\xi(x) = (1 + \xi x)^{-1/\xi}.
$$

ここで, $x = 0$ とおくと

$$
\lim_{n\to\infty} n\{1 - F(b_n)\} = 1.
$$

上の２式の比をとると,

$$
\lim_{n\to\infty} \frac{n\{1 - F(a_n x + b_n)\}}{n\{1 - F(b_n)\}} = (1 + \xi x)^{-1/\xi}.
$$

$$P(X > a_n x + b_n \,|\, X > b_n) = \frac{1 - F(a_n x + b_n)}{1 - F(b_n)} \approx (1 + \xi x)^{-1/\xi}.$$

すなわち,

$$P(X - b_n \le a_n x \,|\, X > b_n) \approx 1 - (1 + \xi x)^{-1/\xi}.$$

この右辺の分布

$$H_\xi(x) = 1 + \log G_\xi(x) = 1 - (1 + \xi x)^{-1/\xi}$$

を**一般パレート (Generalized Pareto, GP) 分布**とよぶ.

b_nが十分大のとき　$P(X - b_n \le y \,|\, X > b_n) \approx H_\xi\left(\frac{y}{a_n}\right).$

問題：H_ξによる近似が十分であるためにはb_nがどれくらい大きければよいか？

○ 分布 F の**条件付き超過分布関数** F_u を

$$F_u(y) = P(X - u \le y \,|\, X > u), \quad 0 \le y \le x_F - u$$

とする.

定理３. [Pickands (1975)]

$$F \in \mathcal{D}(G_\xi) \iff \lim_{u \to x_F} F_u(a(u)y) = H_\xi(y), \quad \forall\, y \ge 0,\ F_u(a(u)y) < 1.$$

ただし，$a(\cdot)$ は a_n を連続化した適当な正の関数.

上の議論は n が確率変数の場合にも拡張されている.

G_ξ：最大値の分布を近似　　H_ξ：分布の右裾を近似

吸引領域に属する分布　（形状パラメータξ：**裾指数**）

$\xi < 0$	$\xi = 0$	$\xi > 0$
ベータ分布	正規分布	Pareto分布
一様分布	指数分布	t分布
	ワイブル分布	Cauchy分布
	対数正規分布	

教科書に出てくる　　連続分布 $\in \mathcal{D}(G_\xi)$,　離散分布 $\notin \mathcal{D}(G_\xi)$.

連続ではあるが裾が重すぎて　$F(x) = 1 - 1/\log x,\ x \geq e \implies F \notin \mathcal{D}(G_\xi)$.

3　GEVモデルの推測

一般極値分布

定義１. 次の分布を**一般極値** (generalized extreme value)**分布** といい GEV(μ, σ, ξ) （$-\infty < \mu < \infty$, $\sigma > 0$, $-\infty < \xi < \infty$）で表す.

$$G(z) = \exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\} = G_\xi\left(\frac{z-\mu}{\sigma}\right), \quad 1+\xi(z-\mu)/\sigma > 0.$$

ただし，G_ξ は標準一般極値分布 GEV(0, 1, ξ) の分布関数

$$G_\xi(z) = \exp\Big[-(1+\xi z)^{-1/\xi}\Big], \quad 1+\xi z > 0,$$

である．μ は位置，σ は尺度，ξ は形状パラメータである.

一般極値分布 $G(z)$ は,
$\xi < 0$ のときは**Weibull分布**で $z < \mu - \sigma/\xi$,
$\xi = 0$ のときは次から**Gumbel分布**で $-\infty < z < \infty$,

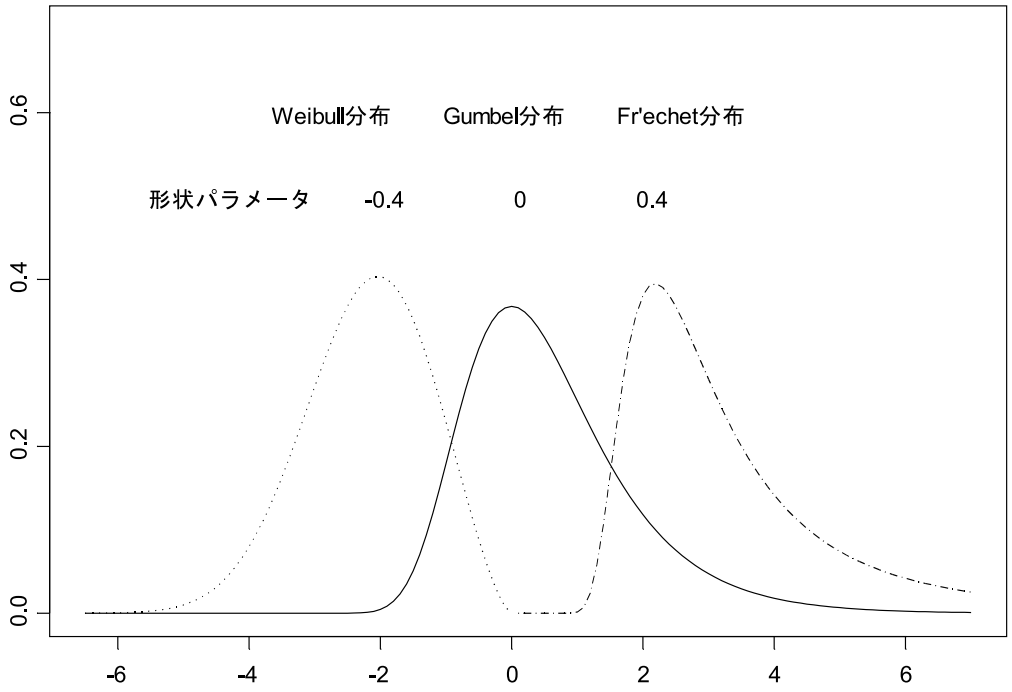
$$G_0((z-\mu)/\sigma) = \lim_{\xi \to 0} G_\xi((z-\mu)/\sigma) = \exp\{-\exp[-(z-\mu)/\sigma]\}$$

$\xi > 0$ の場合は**Fréchet分布**で $z > \mu - \sigma/\xi$.

標準一般極値分布 $G_\xi(z)$ の密度関数は,

$$g_\xi(z) = \begin{cases} (1+\xi z)^{-1/\xi - 1} \exp\left\{ -(1+\xi z)^{-1/\xi} \right\}, & 1+\xi z > 0, \quad \xi \neq 0, \\ \exp\left\{ -z - \exp(-z) \right\}, & -\infty < z < \infty, \quad \xi = 0, \end{cases}$$

となる.



一般極値分布 GEV$(-2.5, 1, -0.4)$（上限 0），GEV$(0,\ 1,\ 0)$,
GEV$(2.5,\ 1,\ 0.4)$（下限 0）の密度関数.

一般極値（GEV) モデル

区分最大値データ $\{z_1, z_2, \ldots, z_n\}$ に一般極値分布 $\mathsf{GEV}(\mu, \sigma, \xi)$,

$$G(z) = \exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\} = G_\xi\left(\frac{z-\mu}{\sigma}\right), \quad 1+\xi(z-\mu)/\sigma > 0,$$

を適合させる.

最尤法　　GEV(μ, σ, ξ)　　**対数尤度**：

$\xi \neq 0$ のとき

$$l(\mu, \sigma, \xi) = -n \log \sigma - (1 + 1/\xi) \sum_{i=1}^{n} \log \left[1 + \xi \left(\frac{z_i - \mu}{\sigma} \right) \right] - \sum_{i=1}^{n} \left[1 + \xi \left(\frac{z_i - \mu}{\sigma} \right) \right]^{-1/\xi},$$

$$1 + \xi(z_i - \mu)/\sigma > 0, \quad i = 1, \ldots, n.$$

$\xi = 0$ のとき

$$l(\mu, \sigma) = -n \log \sigma - \sum_{i=1}^{n} \left(\frac{z_i - \mu}{\sigma} \right) - \sum_{i=1}^{n} \exp \left\{ - \left(\frac{z_i - \mu}{\sigma} \right) \right\}.$$

対数尤度を最大にする最尤推定値 $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$ または $(\hat{\mu}, \hat{\sigma})$ を求める.

GEV(μ, σ, ξ)：**期待情報行列** (Prescott and Walden, 1980)：

$$\frac{n}{\sigma^2\xi^2}\begin{bmatrix} \xi^2 p & \xi\Big\{2\,\Gamma(2+\xi)-p\Big\} & \sigma\xi(p/\xi-q) \\ " & 1-2\,\Gamma(2+\xi)+p & \sigma\left[\dfrac{\Gamma(2+\xi)-1}{\xi}+q-\dfrac{p}{\xi}-1+\gamma\right] \\ " & " & \sigma^2\left[\dfrac{\pi^2}{6}+\left(1-\gamma+\dfrac{1}{\xi}\right)^2-\dfrac{2q}{\xi}+\dfrac{p}{\xi^2}\right] \end{bmatrix}$$

ただし (μ, σ, ξ) の順で，$\Gamma(\,\cdot\,)$ はガンマ関数，$\Psi(r) = d\log\Gamma(r)/dr$ で

$$p = (1+\xi)^2\Gamma(1+2\xi), \quad q = \Gamma(2+\xi)\{\Psi(1+\xi)+(1+\xi)/\xi\},$$

$\gamma = 0.5772157...$ Euler の定数.

○ 分布族 $\{\mathrm{GEV}(\mu, \sigma, \xi),\ -\infty < \mu < \infty,\ \sigma > 0,\ -\infty < \xi < \infty\}$：正則条件を満たしていない.

○ しかし，$\xi > -0.5$ の場合は最尤推定量は一致推定量で漸近正規性を持ち漸近有効推定量になる（Smith, 1985）.

○ 応用上 $\xi \leq -0.5$ となることは稀で，特に自然現象では $-0.5 < \xi < 0.5$ となることが多いと言われている.

○ パラメータの推定は最尤法で行えばよい.

○ 各最尤推定値の標準誤差の推定値は観測情報行列から求めることが出来る．これから，(μ, σ, ξ) の各パラメータの信頼区間が求まる.

極値統計学では，次の極値分布の上側 (微少) 確率点の推定が目的の場合が多い.

一般極値分布 GEV(μ, σ, ξ) の $1 - 1/T$ 確率点 R_T，すなわち

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

は次のように表される：

$$R_T = \begin{cases} \mu + \sigma\Big[\Big\{-\log(1-1/T)\Big\}^{-\xi} - 1\Big]\Big/\xi, & \xi \neq 0, \\ \mu + \sigma\Big[-\log\Big\{-\log(1-1/T)\Big\}\Big], & \xi = 0. \end{cases}$$

この確率点 R_T は**再現期間** (return period) T の **再現レベル** (return level) という．例えば年最大値データを扱うとき，再現期間 $T = 200$ 年の再現レベル R_{200} は 200 年に平均 1 度**超えられる**様な（大きな）値と解釈できる.

一般に $n \ll T$ で，これはデータの存在しない領域の推測（外挿）になる.

○ 再現レベル R_T の最尤推定値は

$$\widehat{R}_T = \begin{cases} \widehat{\mu} + \widehat{\sigma}\left\{ y_T^{-\widehat{\xi}} - 1 \right\} / \widehat{\xi}, & \widehat{\xi} \neq 0, \\ \widehat{\mu} + \widehat{\sigma}\left\{ -\log y_T \right\}, & \widehat{\xi} = 0, \end{cases}$$

となる，ただし $(\widehat{\mu}, \widehat{\sigma}, \widehat{\xi})$ は最尤推定値で $y_T = -\log(1 - 1/T)$.

デルタ法より，分散は $V(\widehat{R}_T) \approx \nabla R_T^{\mathbf{T}} \, V \, \nabla R_T$ で求まる.
ここで V は $(\widehat{\mu}, \widehat{\sigma}, \widehat{\xi})$ の分散共分散行列で

$$\begin{aligned} \nabla R_T^{\mathbf{T}} &= \left[\frac{\partial R_T}{\partial \mu}, \frac{\partial R_T}{\partial \sigma}, \frac{\partial R_T}{\partial \xi} \right] \\ &= \left[1, \, (y_T^{-\xi} - 1)/\xi, \, \sigma y_T^{-\xi} (-\log y_T)/\xi - \sigma (y_T^{-\xi} - 1)/\xi^2 \right] \end{aligned}$$

を $(\widehat{\mu}, \widehat{\sigma}, \widehat{\xi})$ で求める.

○ $\widehat{\xi} < 0$の場合は 分布の上限 $R_* = \mu - \sigma/\xi$ の推定

$$\widehat{R}_* = \widehat{\mu} - \widehat{\sigma}/\widehat{\xi}$$

が重要．この場合は

$$\nabla R_*^{\mathbf{T}} = \left[1,\ -1/\xi,\ \sigma/\xi^2\right]$$

となる．

○ 一般に，ξ や R_T の信頼区間はプロファイル尤度から求めた方が精度が良い.

○ **形状パラメータ** ξ のプロファイル尤度は，$\xi = \xi_0$ として $l(\mu, \sigma, \xi_0)$ を μ と σ に関して最大化して求めればよい.

○ ξ の 95%**信頼区間**は近似的に

$$\begin{aligned} & \left\{\xi : \ 2\Big\{l(\hat{\mu}, \hat{\sigma}, \hat{\xi}) - \max_{\mu, \sigma} l(\mu, \sigma, \xi)\Big\} \leq \chi_1^2(0.05) = 3.841\right\} \\ = \ & \left\{\xi : \ \max_{\mu, \sigma} l(\mu, \sigma, \xi) \geq l(\hat{\mu}, \hat{\sigma}, \hat{\xi}) - 1.921\right\} \end{aligned}$$

となる.

○ **再現レベル** R_T のプロファイル尤度を求めるために,

$$\mu = R_T - \sigma\left[y_T^{-\xi} - 1\right]/\xi$$

から，パラメータ $(\mu,\, \sigma,\, \xi)$ を $(R_T,\, \sigma,\, \xi)$ へ換える．このとき，対数尤度は

$$l(R_T,\, \sigma,\, \xi) = -n\log\sigma - (1+1/\xi)\sum_{i=1}^{n}\log\left[y_T^{-\xi} + \xi\left(\frac{z_i - R_T}{\sigma}\right)\right] - \sum_{i=1}^{n}\left[y_T^{-\xi} + \xi\left(\frac{z_i - R_T}{\sigma}\right)\right]^{-1/\xi}$$

と表される．これから，R_T の 95%信頼区間は近似的に

$$\left\{R_T :\ 2\left\{l(\hat{R}_T,\, \hat{\sigma},\, \hat{\xi}) - \max_{\sigma,\,\xi} l(R_T,\, \sigma,\, \xi)\right\} \le \chi_1^2(0.05)\right\}$$

$$= \left\{R_T :\ \max_{\sigma,\,\xi} l(R_T,\, \sigma,\, \xi) \ge l(\hat{R}_T,\, \hat{\sigma},\, \hat{\xi}) - \chi_1^2(0.05)/2\right\}$$

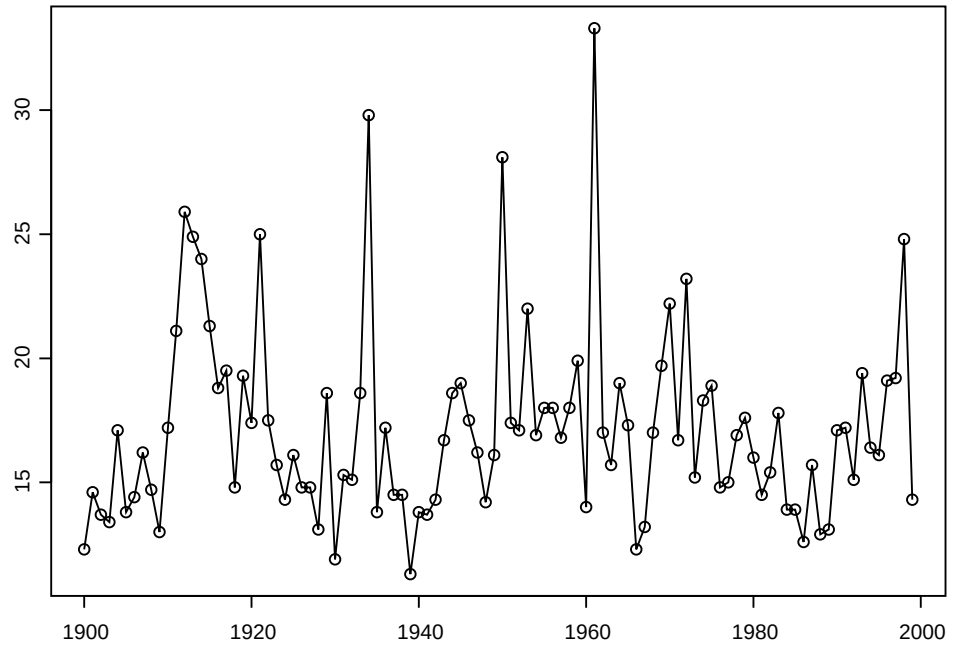
となる．

データ解析例

大阪の1900年から1999年までの百年間の年最大風速(m/s)データ（大阪管区気象台編 (1982)，石原他 (2002)）をGEVモデルで解析する．

気象データに関しては観測基準や観測機器が変わっていることがあるが，ここではこれらの影響を無視する．

気象データ等の品質に関しては木下 (2004) を参照．



大阪の年最大風速(m/s)，1900年～1999年.
データの最大値は33.3(m/s).

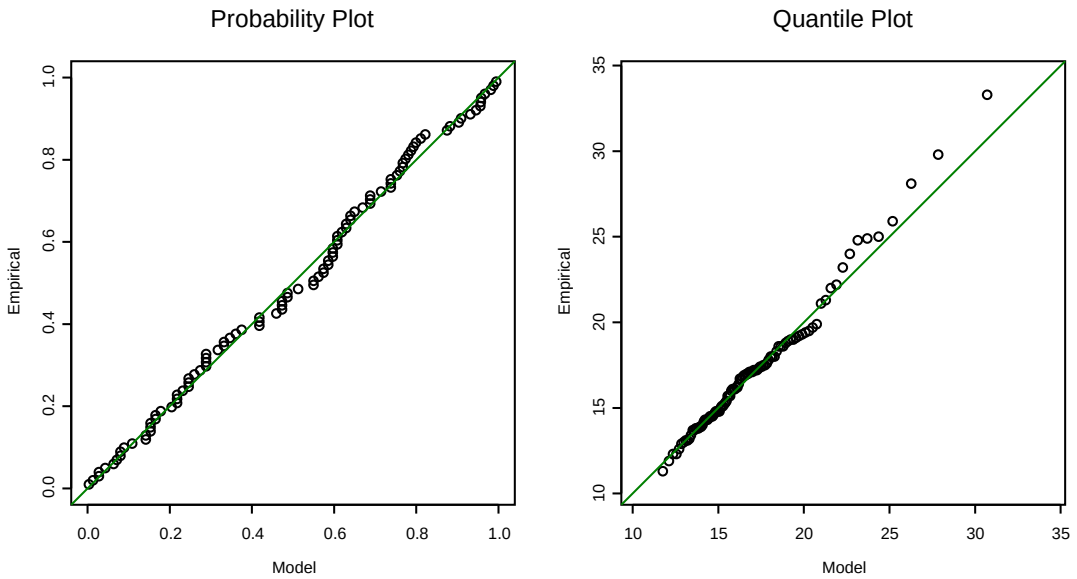
データに一般極値分布 GEV(μ, σ, ξ) を適合させる．

最尤推定値は $(\hat{\mu}, \hat{\sigma}, \hat{\xi}) = (15.349, 2.550, 0.111)$，最大対数尤度は -257.78 で，分散共分散行列は

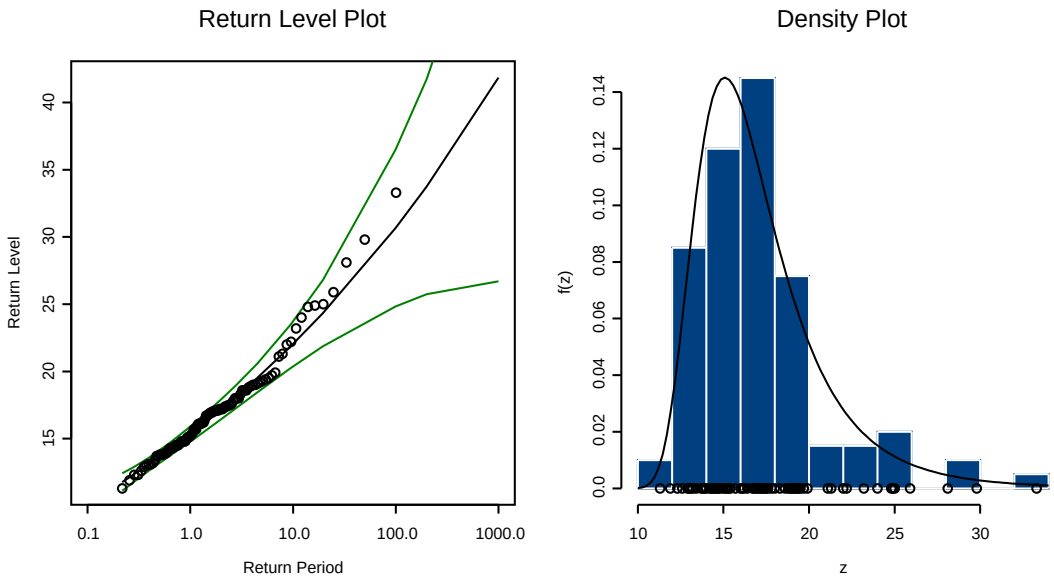
$$V = \begin{bmatrix} 0.08235 & 0.03024 & -0.00686 \\ 0.03024 & 0.04703 & -0.00231 \\ -0.00686 & -0.00231 & 0.00567 \end{bmatrix}$$

となった．この行列の対角成分が $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$ の分散である．

データは一般極値分布に従うとして解析を行っている．再現レベルの推定はデータの存在しない領域への外挿になりデータによる検証は行えない．しかし，データへの一般極値分布の適合の診断は PP plot や QQ plot で出来る．この場合の適合は次の図で見る様に良好である．



大阪の年最大風速データ解析の診断.



大阪の年最大風速データ解析の診断.

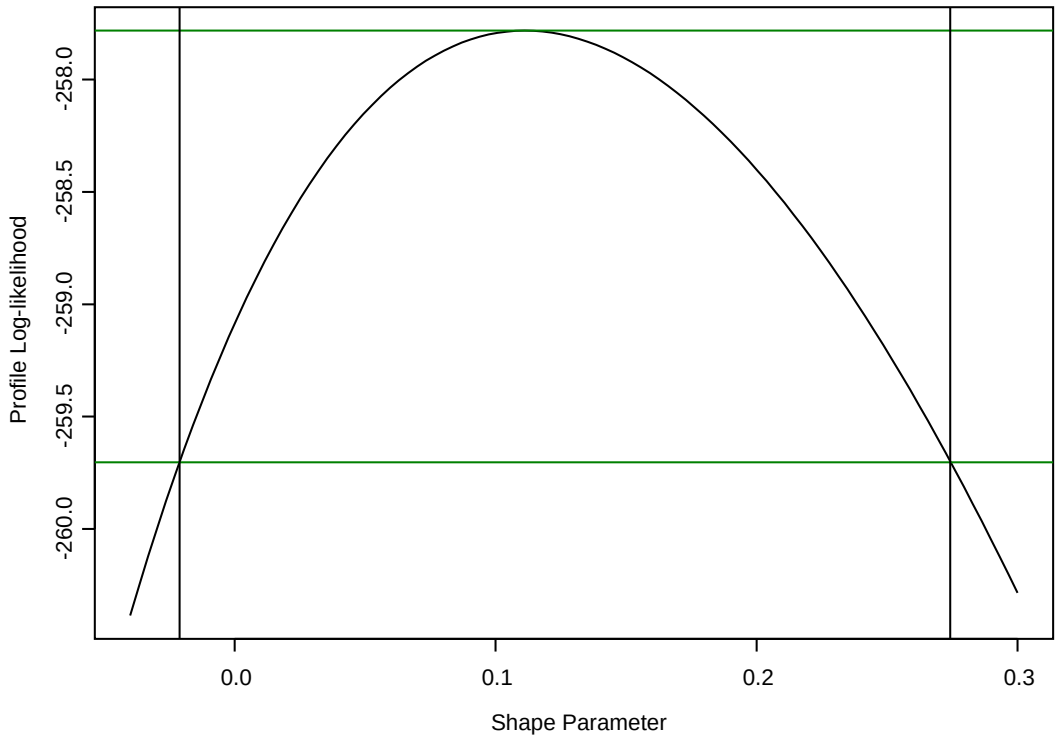
大阪の年最大風速では $\widehat{\xi} = 0.111 > 0$ で Fréchet 分布が適合する.

ξ の 95%信頼区間は，最尤推定量の漸近正規性より

$$\xi \in 0.111 \pm 1.96\sqrt{0.00567} = [-0.037,\ 0.259].$$

プロファイル尤度に基づく 95%信頼区間は次の図より $[-0.021,\ 0.274]$.

プロファイル尤度に基づく信頼区間は右にずれている.
一般にプロファイル尤度に基づく信頼区間の方が精度が良い.



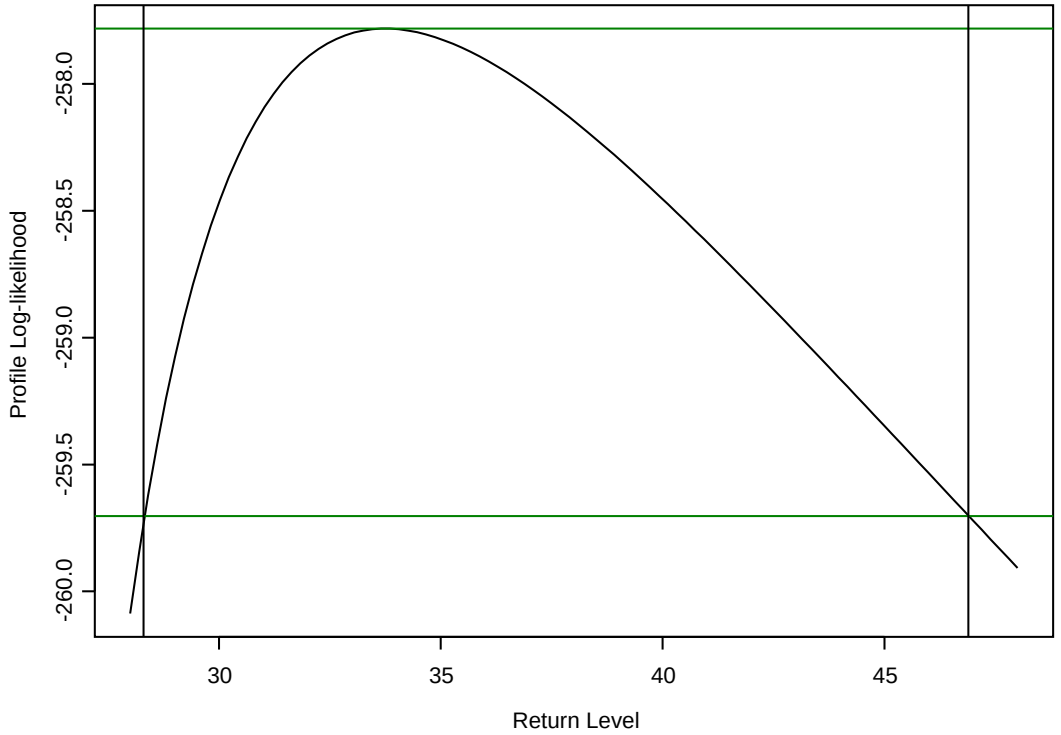
大阪の年最大風速の形状パラメータξ のプロファイル尤度.
$\widehat{\xi} = 0.111 > 0$, 95%信頼区間は$[-0.021, 0.274]$.

再現期間200年の再現レベルR_{200}の最尤推定値は33.74で，デルタ法による95%信頼区間は

$$R_{200} \in \hat{R}_{200} \pm 1.96\sqrt{\nabla R_{200}^{\mathbf{T}} V \nabla T_{200}} = [25.75,\ 41.73].$$

プロファイル尤度に基づく95%信頼区間は次の図より [28.29, 46.90].

プロファイル尤度に基づく信頼区間は右にずれている.
一般にプロファイル尤度に基づく信頼区間の方が精度が良い.



大阪の年最大風速の200年再現レベルのプロファイル尤度.
$\widehat{R}_{200} = 33.74$, 95%信頼区間は [28.29, 46.90].

4　GP モデルの推測

定義２． 次の分布を**一般パレート (generalized Pareto) 分布**といい GP(σ, ξ) （$\sigma > 0$, $-\infty < \xi < \infty$）で表す.

$$H(y) = 1 - \left(1 + \xi \frac{y}{\sigma}\right)^{-1/\xi} = H_\xi\left(\frac{y}{\sigma}\right), \quad 1 + \xi y/\sigma > 0.$$

ただし，H_ξ は標準一般パレート分布 GP$(1, \xi)$ の分布関数

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

とする．σ は尺度パラメータ, ξ は形状パラメータである.

一般パレート分布 $H(y)$ は,
$\xi < 0$ のときは**ベータ分布**で $0 < y < -\sigma/\xi$,
$\xi = 0$ のときは次より**指数分布**で $0 < y < \infty$,

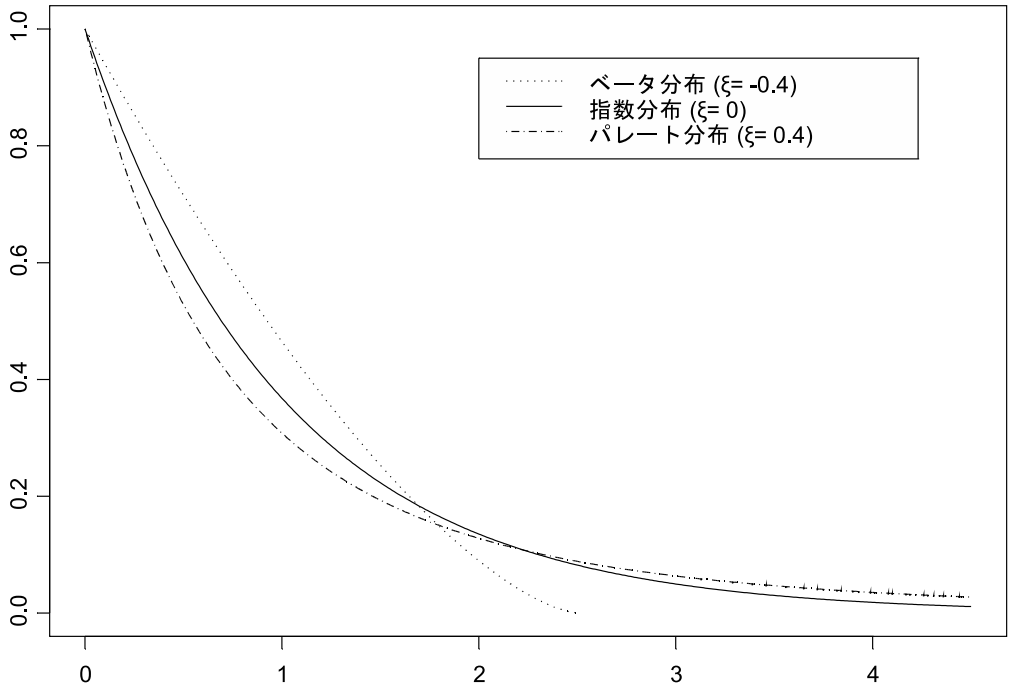
$$H_0(y/\sigma) = \lim_{\xi \to 0} H_\xi(y/\sigma) = 1 - e^{-y/\sigma}$$

$\xi > 0$ の場合は**Pareto分布**で $0 < y < \infty$.

標準一般パレート分布 $H_\xi(y)$ の密度関数は,

$$h_\xi(y) = \begin{cases} (1+\xi y)^{-1/\xi - 1}, & 1 + \xi y > 0, \quad \xi \neq 0, \\ \exp(-y), & 0 < y < \infty, \quad \xi = 0, \end{cases}$$

となる.



一般パレート分布 GP$(1, \xi)$, $\xi = -0.4,\ 0,\ 0.4$ の密度関数.

一般パレート(GP)モデル

水準超過データ $\{y_1, y_2, \ldots, y_n\}$ に一般パレート分布 $\mathrm{GP}(\sigma, \xi)$

$$H(y) = 1 - \left(1 + \xi \frac{y}{\sigma}\right)^{-1/\xi} = H_\xi\left(\frac{y}{\sigma}\right), \quad 1 + \xi y/\sigma > 0,$$

を適合させる.

最尤法　GP(σ, ξ)　　**対数尤度**：

$\xi \neq 0$ のとき

$$l(\sigma, \xi) = -n \log \sigma - (1 + 1/\xi) \sum_{i=1}^{n} \log(1 + \xi y_i/\sigma),$$

$$1 + \xi y_i/\sigma > 0, \quad i = 1, 2, \ldots, n.$$

$\xi = 0$ のとき

$$l(\sigma) = -n \log \sigma - \frac{1}{\sigma} \sum_{i=1}^{n} y_i.$$

対数尤度を最大にする最尤推定値 $(\hat{\sigma}, \hat{\xi})$ または $\hat{\sigma}$ を求める.

一般パレート分布 GP(σ, ξ) の Fisher 情報量行列は

$$\frac{n}{(1+\xi)(1+2\xi)}\begin{bmatrix} (1+\xi)/\sigma^2 & 1/\sigma \\ 1/\sigma & 2 \end{bmatrix}$$

となる．したがって$\xi > -1/2$ならば情報行列は有限で，最尤推定量は漸近的に分散共分散行列が

$$V_n = \frac{1}{n}\begin{bmatrix} 2\sigma^2(1+\xi) & -\sigma(1+\xi) \\ -\sigma(1+\xi) & (1+\xi)^2 \end{bmatrix}$$

の 2 変量正規分布に従う（Smith, 1985）.

最尤推定値の分散共分散行列は観測情報行列から求めた方が精度が良い．これを用いて信頼区間を構成することができるが，プロファイル尤度を用いて求めたほうが精度が良い.

一般パレート分布の性質　（閾値の決定で用いる）

○ $Y \sim \mathrm{GP}(\sigma, \xi)$　　$\xi < 1$ で平均は存在

$$E(Y) = \int_0^{y_+} (1 - H(y))dy = \int_0^{y_+} \left(1 + \xi \frac{y}{\sigma}\right)^{-1/\xi} dy = \frac{\sigma}{1-\xi}.$$

ただし，$y_+ = \sup\{y : H(y) < 1\}$.

○ $Y - v | Y > v \sim \mathrm{GP}(\sigma + \xi v, \xi)$　$(v > 0)$
　次から明らか,

$$\begin{aligned} P(Y - v > y | Y > v) &= \frac{1 - H(y+v)}{1 - H(v)} = \frac{\left(1 + \xi(y+v)/\sigma\right)^{-1/\xi}}{\left(1 + \xi v/\sigma\right)^{-1/\xi}} \\ &= \left(1 + \xi \frac{y}{\sigma + \xi v}\right)^{-1/\xi}. \end{aligned}$$

○ $e(v) = E(Y - v \,|\, Y > v)$: Y の **平均超過** (mean excess) **関数**

○ $\tilde{e}(v)$: Y の **メジアン超過** (median excess) **関数**

$$P(Y - v \le \tilde{e}(v) \,|\, Y > v) = 1/2$$

○ $Y - v \,|\, Y > v \sim \mathsf{GP}(\sigma + \xi v,\, \xi)$ より

$$e(v) = \frac{\sigma + \xi v}{1 - \xi} = \frac{\sigma}{1 - \xi} + \frac{\xi}{1 - \xi}\, v,$$

$$\tilde{e}(v) = \begin{cases} \sigma(2^{\xi} - 1)/\xi + (2^{\xi} - 1)v, & \xi \neq 0, \\ \sigma \log 2, & \xi = 0. \end{cases}$$

閾値の決定

閾値 u を小さくとるとデータ数は多くなり，分散は小さくなるが一般パレート分布への適合は悪くなり偏りは大きくなる．

閾値u を大きくとるとデータ数は少なくなり，一般パレート分布への適合は良くなり偏りは小さくなるが分散は大きくなる．

平均超過プロット (mean excess plot)

データで u を超えるものを $x_{[1]}, x_{[2]}, \ldots, x_{[N_u]}$ とし，x_{max} でデータの最大値を表す．値 $u < x_{max}$ に対して標本平均超過，すなわち

$$\left\{ \left(u,\ \frac{1}{N_u} \sum_{i=1}^{N_u} (x_{[i]} - u) \right) :\ \ u < x_{max} \right\}$$

をプロットする．この図で，ある u 以上ではプロットが直線と見なせるとき，この u で最小のものを閾値と決める．

$$Y \sim \mathsf{GP}(\sigma, \xi) \implies E(Y - v \,|\, Y > v) = \frac{\sigma}{1-\xi} + \frac{\xi}{1-\xi} v.$$

パラメータ ξ と σ^* の安定性

値 $u < x_{max}$ に対して，データ $\{x_{[i]} - u\}_{i=1}^{N_u}$ に $\mathrm{GP}(\sigma_u, \xi)$ を適合させ最尤法でパラメータ σ_u と ξ を推定する．値 u を変化させて $(u, \hat{\sigma}^*)$ と $(u, \hat{\xi})$ をプロットする．ただし $\sigma^* = \sigma_u - \xi u$．このとき，ある u 以上では推定値 $\hat{\sigma}^*$ と $\hat{\xi}$ が共に一定であると見なせるとき，この u の最小のものを閾値と決める．

○ $Y - v \,|\, Y > v \sim \mathrm{GP}(\sigma + \xi v, \xi) \quad (v > 0), \quad \sigma_v = \sigma + \xi v.$

バリューアットリスク

極値統計学では次の確率点の推定が目的の場合が多い.

母集団分布 F で

$$F(\mathrm{VaR}_p) = 1 - p$$

となる確率点は**バリューアットリスク** (Value–at–Risk) とよばれる．これを推定するために母集団分布 $F(x)$ を次の様に分解する：

$$F(x) = (1 - F(u))F_u(y) + F(u), \quad y = x - u.$$

適当な条件の下で u が十分大きいとき，F_u は GP 分布 H_ξ で近似できる：

$$F(x) \approx (1 - F(u))H_\xi\left(\frac{x-u}{\sigma}\right) + F(u), \quad x \geq u.$$

上の式で$=$とし，$\zeta_u = 1 - F(u)$ とおくと

$$\mathsf{VaR}_p = u + \frac{\sigma}{\xi}\left\{\left(\frac{\zeta_u}{p}\right)^{\xi} - 1\right\},$$

と表される.

閾値uを決定し，閾値を超過するデータでパラメータの最尤推定値$(\hat{\sigma}, \hat{\xi})$を求める．また，$\zeta_u$は$N_u/n$で推定する．ここで，$n$は標本サイズで$N_u$は閾値$u$を超過したデータ数.

これらを代入して最尤推定値

$$\widehat{\mathsf{VaR}}_p = u + \frac{\hat{\sigma}}{\hat{\xi}}\left\{\left(\frac{\hat{\zeta}_u}{p}\right)^{\hat{\xi}} - 1\right\}$$

を得る.

$\widehat{\mathsf{VaR}}_p$ の標準誤差はデルタ法で求める.

$(\hat{\zeta}_u,\, \hat{\sigma},\, \hat{\xi})$ の分散共分散行列は近似的に

$$V^* = \begin{bmatrix} \hat{\zeta}_u(1-\hat{\zeta}_u)/n & \mathbf{0}^\mathsf{T} \\ \mathbf{0} & V_{N_u} \end{bmatrix}$$

となる，ここで $\mathbf{0}^\mathsf{T} = (0,\, 0)$.

VaR_p の最尤推定量の分散はデルタ法から

$$V(\widehat{\mathsf{VaR}}_p) \approx \nabla \mathsf{VaR}_p^\mathsf{T}\; V^*\; \nabla \mathsf{VaR}_p,$$

ただし

$$\nabla \mathsf{VaR}_p^\mathsf{T} = \left[\frac{\partial \mathsf{VaR}_p}{\partial \zeta_u},\, \frac{\partial \mathsf{VaR}_p}{\partial \sigma},\, \frac{\partial \mathsf{VaR}_p}{\partial \xi} \right]$$

で，$(\hat{\zeta}_u,\, \hat{\sigma},\, \hat{\xi})$ で評価する.

プロファイル尤度を用いた信頼区間もGEVモデルの場合と同様に求まる．
例えば，ξの 95%信頼区間は近似的に

$$\{\xi : \max_{\sigma} l(\sigma, \xi) \geq l(\hat{\sigma}, \hat{\xi}) - 1.921\}$$

となる．

一方，VaR_pの信頼区間は

$$\sigma = \frac{(\mathrm{VaR}_p - u)\xi}{(\zeta_u/p)^{\xi} - 1}$$

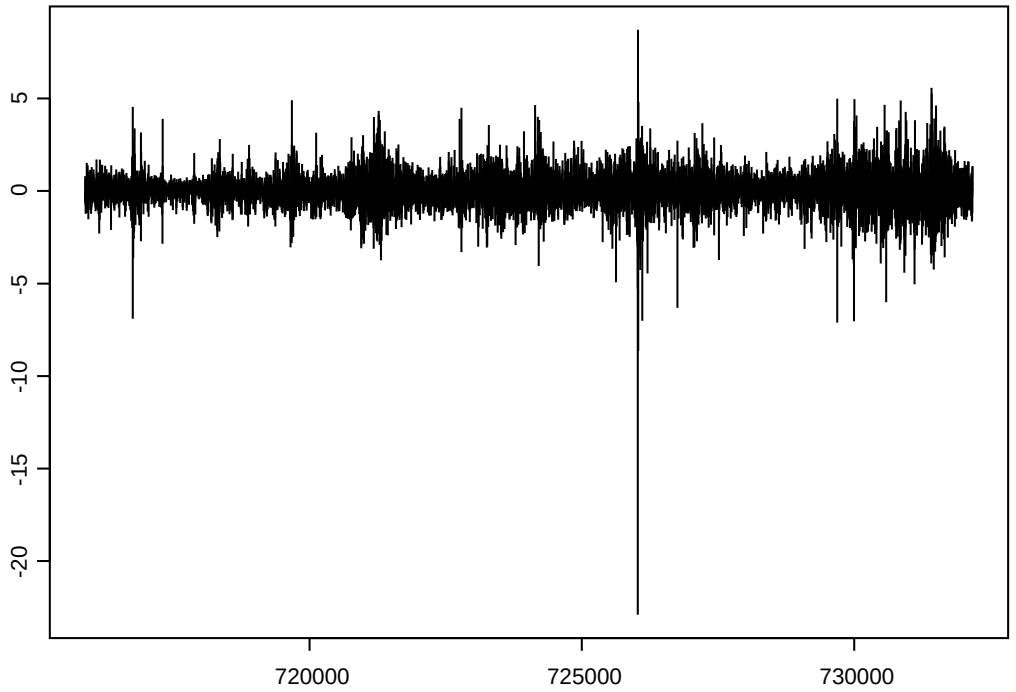
として，対数尤度 $l(\mathrm{VaR}_p, \xi)$を考える．VaR_pを固定して$\max_{\xi} l(\mathrm{VaR}_p, \xi)$を求めればよい．ここで応用上$\zeta_u$のバラツキは小さいので無視している．

データ解析例

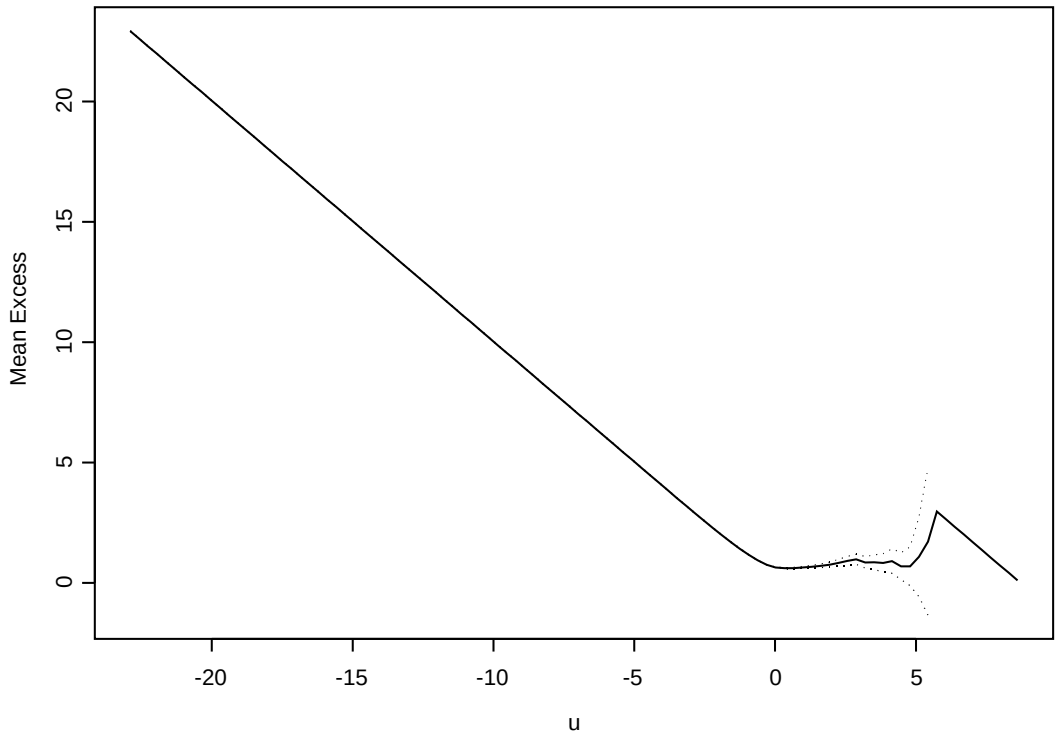
○ M. Gilli & E. Këllezi
An Application of Extreme Value Theory for Measuring Financial Risk.

S&P 500 のデータ　　**右裾**についての推測

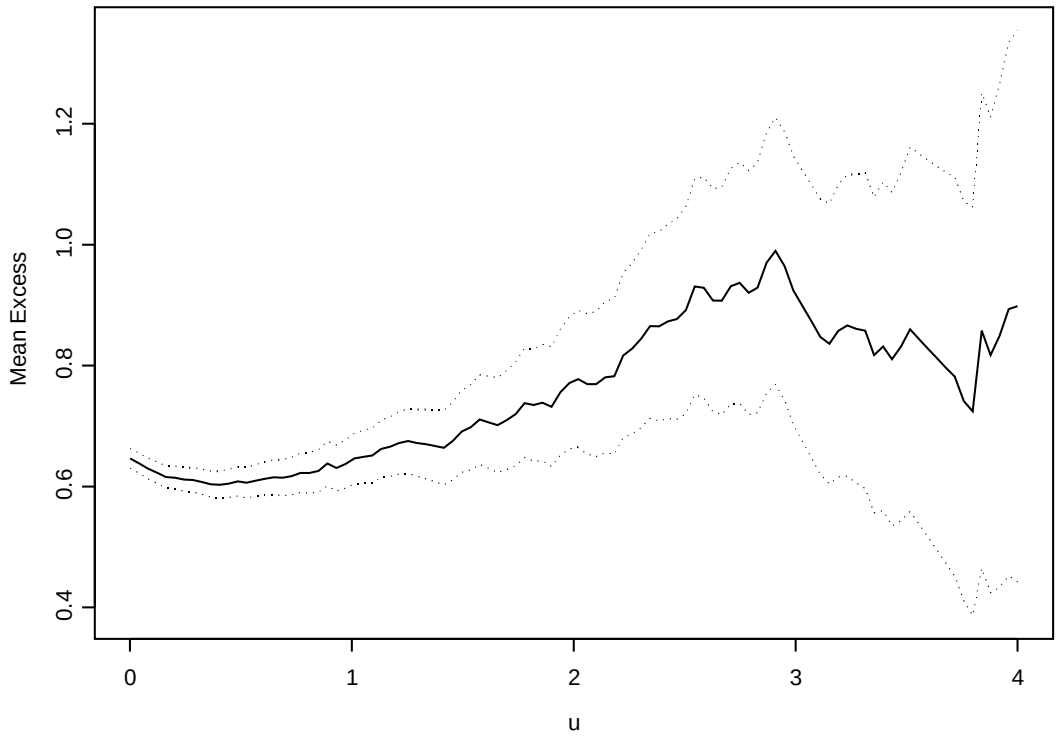
閾値の決定のためにデータから「標本平均超過関数」と「修正尺度と形状パラーメータの推定値」を閾値を変化させてプロット．標本平均超過関数の図では，データが少ない右端では変動が激しい．そこでデータの多い箇所を拡大．同様に「修正尺度と形状パラーメータの推定値」の図もデータの多い箇所について作成.



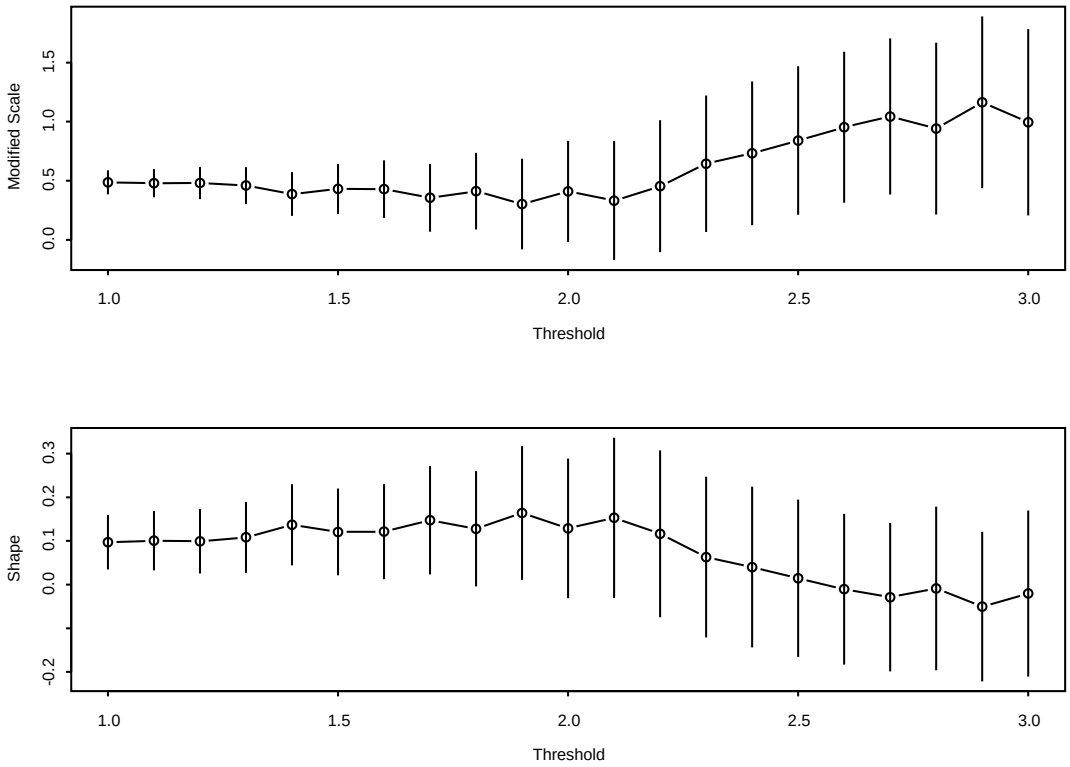
Daily returns of the S&P 500 index.
1960年1月5日～2004年8月16日（11270データ）



データの標本平均超過関数.



データの個数の多い箇所を拡大した標本平均超過関数.



修正尺度σ^*（上）と形状パラメータξ（下） の最尤推定値.

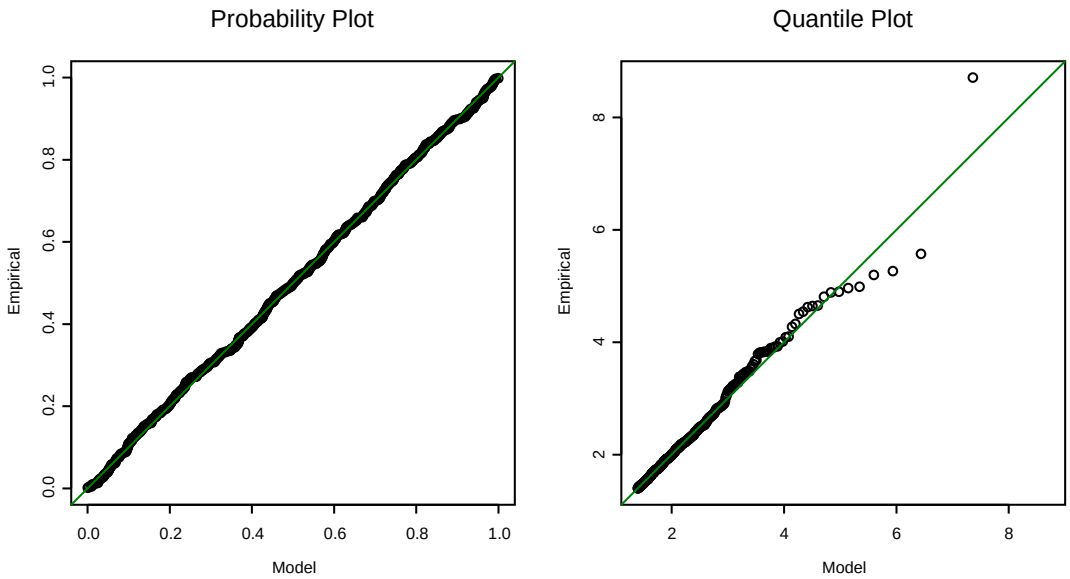
閾値を $u = 1.4$ とする．このとき閾値以上のデータは614個.

最尤推定値と標準誤差は

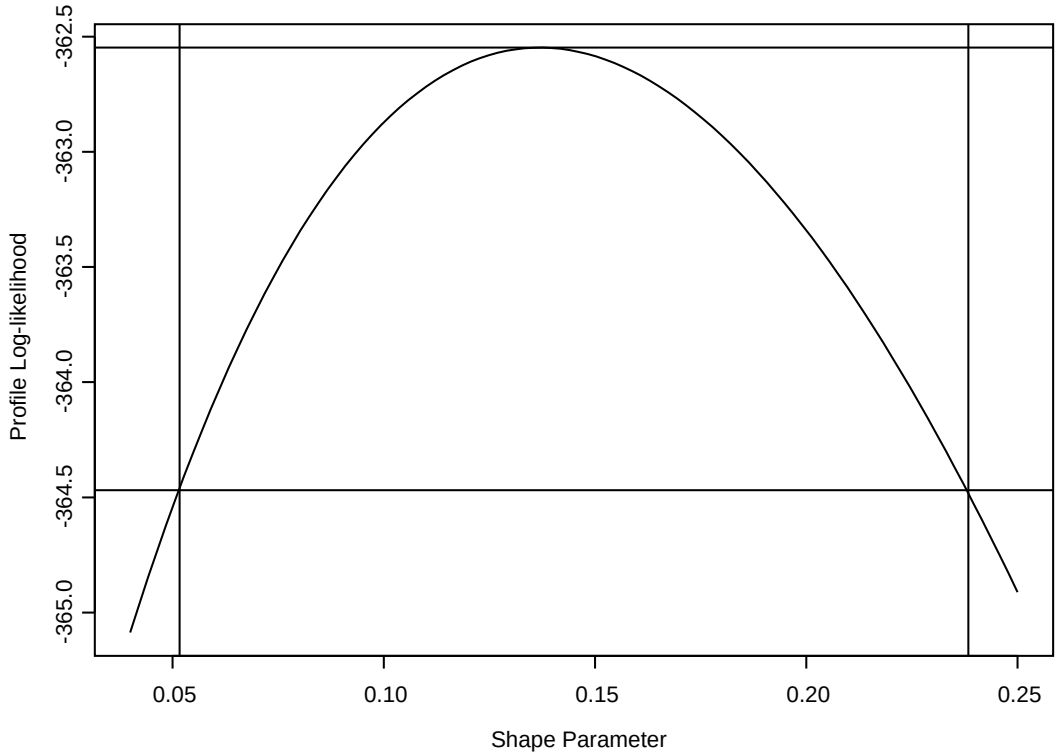
$$\hat{\sigma} = 0.579\ (0.0359), \quad \hat{\xi} = 0.137\ (0.0474)$$

で，対応する尤度は -362.548 となった.

データ解析の診断は概ね良好.

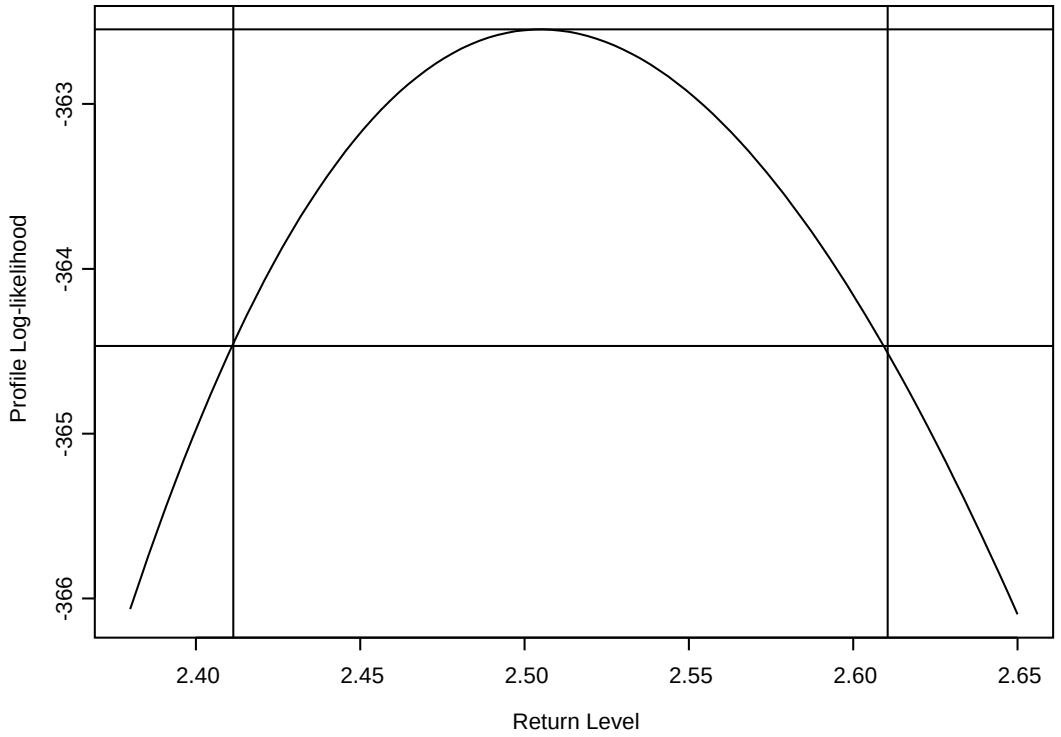


S&P 500 データ解析の診断.



形状パラメータξ の 95%信頼区間は$[0.052, 0.238]$.

($\hat{\xi} = 0.137$. 漸近正規性による信頼区間は$[0.044, 0.230]$.)



$\mathrm{VaR}_{0.01}$ の 95%信頼区間は [2.411, 2.610].

ソフトウェア　極値統計学のデータ解析用ソフトの総合報告

Stephenson, A. and Gilleland, E. (2006).
Software for the analysis of extreme events: The current state and future directions. *Extremes* **8**, 87–109.

雑誌

極値の専門雑誌 Extremes が1998年に創刊された.

和雑誌「統計数理」(2004) Vol. 52 , No. 1 は極値理論の特集号である.
(`http://www.ism.ac.jp/editsec/toukei.html`からダウンロード可能.)

参考書

- Beirlant, J., Goegebeur, Y., Segers, J. and Teugels, J. (2004). *Statistics of Extremes, Theory and Applications*. Wiley.

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

- Embrechts, P., Klüppelberg, C. and Mikosch, T. (2001). *Modelling Extremal Events for Insurance and Finance*, 3rd ed. Springer.

- Falk, M., Hüsler, J. and Reiss, R.-D. (2004). *Laws of Small Numbers: Extremes and Rare Events*, 2nd ed. Birkhäuser.

- de Haan, L. and Ferreira, A. (2006). *Extreme Value Theory: An Introduction*. Springer.

- Kotz, S. and Nadarajah, S. (2000). *Extreme Value Distributions: Theory and Applications*. Imperial College Press.

- Leadbetter, M. R., Lindgren, G. and Rootzén, H. (1983). *Extremes and Related Properties of Random Sequences and Processes*. Springer-Verlag, New York.

- McNeil, J., Frey, R. and Embrechts, P. (2005). *Quantitative Risk Management*. Princeton University Press.

- Reiss, R.-D. and Thomas, M. (2007). *Statistical Analysis of Extreme Values, with Applications to Insurance, Finance, Hydrology and Other Fields*, 3rd ed. Birkhäuser.

- Resnick, S. I. (1987). *Extreme Values, Regular Variation and Point Processes*. Springer-Verlag, Berlin.

- Resnick, S. I. (2007). *Hervy-Tail Phenomena.* Springer.

文献

木下武雄 (2004). 自然災害研究のための利用可能データ. 統計数理, 第52巻, 第1号, 5–24.

渋谷政昭 (2008). 極値理論とその応用. 国友直人編「保険と金融の統計学II」.

渋谷政昭, 高橋倫也 (2008). 極値理論, 信頼性, リスク管理. 21世紀の統計科学II. 89–124.

志村隆彰 (2008). 吸引領域と離散分布.「極値理論の工学への応用(5)」. 統計数理研究所共同研究リポート212, 25–32.

高橋倫也，渋谷政昭 (2004)．上位 r 個の観測値に基づく確率点の推定．統計数理，第52巻，第1号，93–116.

Fisher, R. A. and Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Proc. Cambridge Philos. Soc.* **24**, 180–190.

Fréchet, M. (1927). Sur la loi de probabilité de l'écart maximum. *Ann. Soc. Math. Polon.* **6**, 93–116.

Gnedenko, B. (1943). Sur la distibution limite du terme maximum d'une serie aleatoire. *Ann, Math.* **44**, 423–453. Translated and

reprinted in: *Breakthroughs in Statistics*, Vol.I, 1992, eds. S. Kotz and N. L. Johnson, Springer-verlag, pp. 195–225.

Gumbel, E. J. (1958). *Statistics of Extremes*. Columbia University Press.

de Haan, L. (1970). *On Regular Variation and Its Application to the Weak Convergence of Sample Extremes*. Math. Centre Tracts, Vol.**32**, Amsterdam.

de Haan, L. (1976). Sample extremes: an elementary introduction. *Statist. Neerlandica* **30**, 161–172.

Jenkinson, A. F. (1955). The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *Quart. J. R. Met. Soc.* **81**, 158–171.

Mises, R. von (1936). La distribution de la plus grande de *n* valeurs. *Rev. Math. Union Interbalcanique* **1**, 141–160. Reproduced in *Selected Papers of Richard von Mises*, Amer. Math. Soc. **2** (1964), pp. 271–294.

Pickands, J. (1975). Statistical inference using extreme order statistics. *Ann. Statist.* **3**, 119–131.

Prescott, P. and Walden, A. T. (1980). Maximum likelihood estimation of the parameters of the generalized extreme value distribution. *Biometrika* **67**, 723–724.

Smith, R. L. (1985). Maximum likelihood estimation in a class of nonregular cases. *Biometrika* **72**, 67–90.

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

Smith, R. L. (1987). Estimating tails of probability distributions. *Ann. Statist.* **15**, 1174–1207.

Smith, R. L. (1989). Extreme value analysis of environmental time series: an example based on ozone data (with discussion). *Statistical Science* **4**, 367–393.

Smith, R. L. (1990). Extreme value theory. In *Handbook of Applicable Mathematics* VII, Ed. W. Ledermann, Chichester, Wiley. pp. 437–472.

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