平成２６年１２月８日（の修正版）

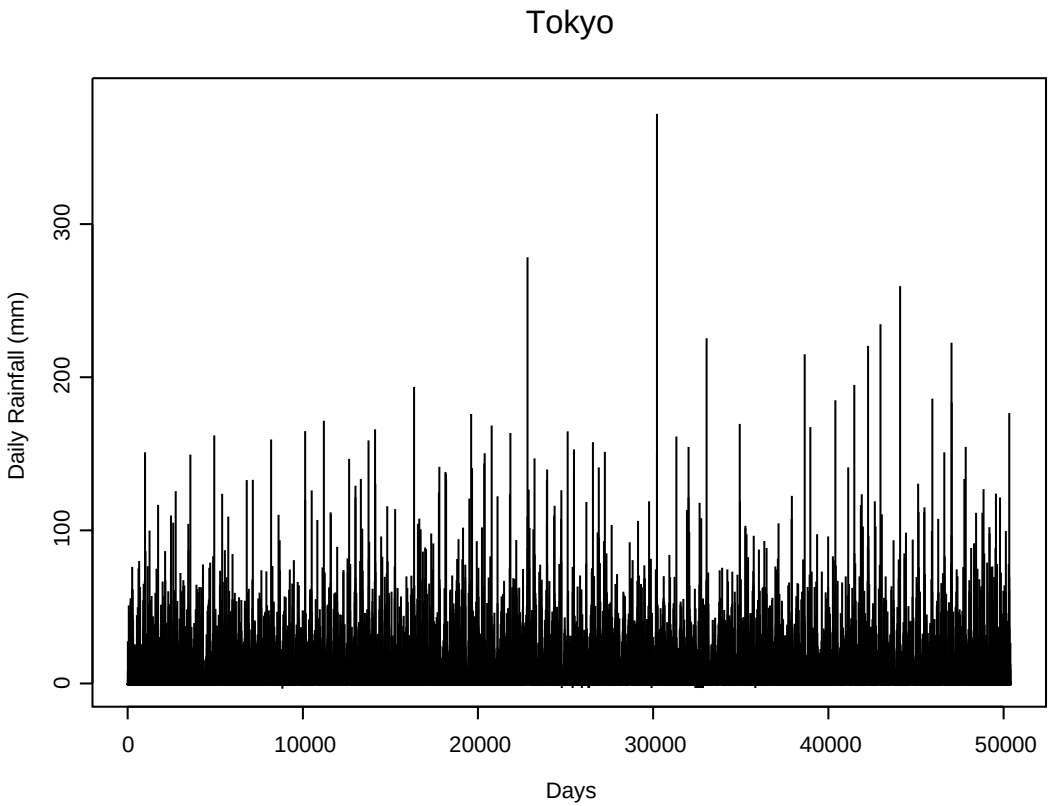
極値統計学

高橋 倫也 (神戸大学・名誉教授)

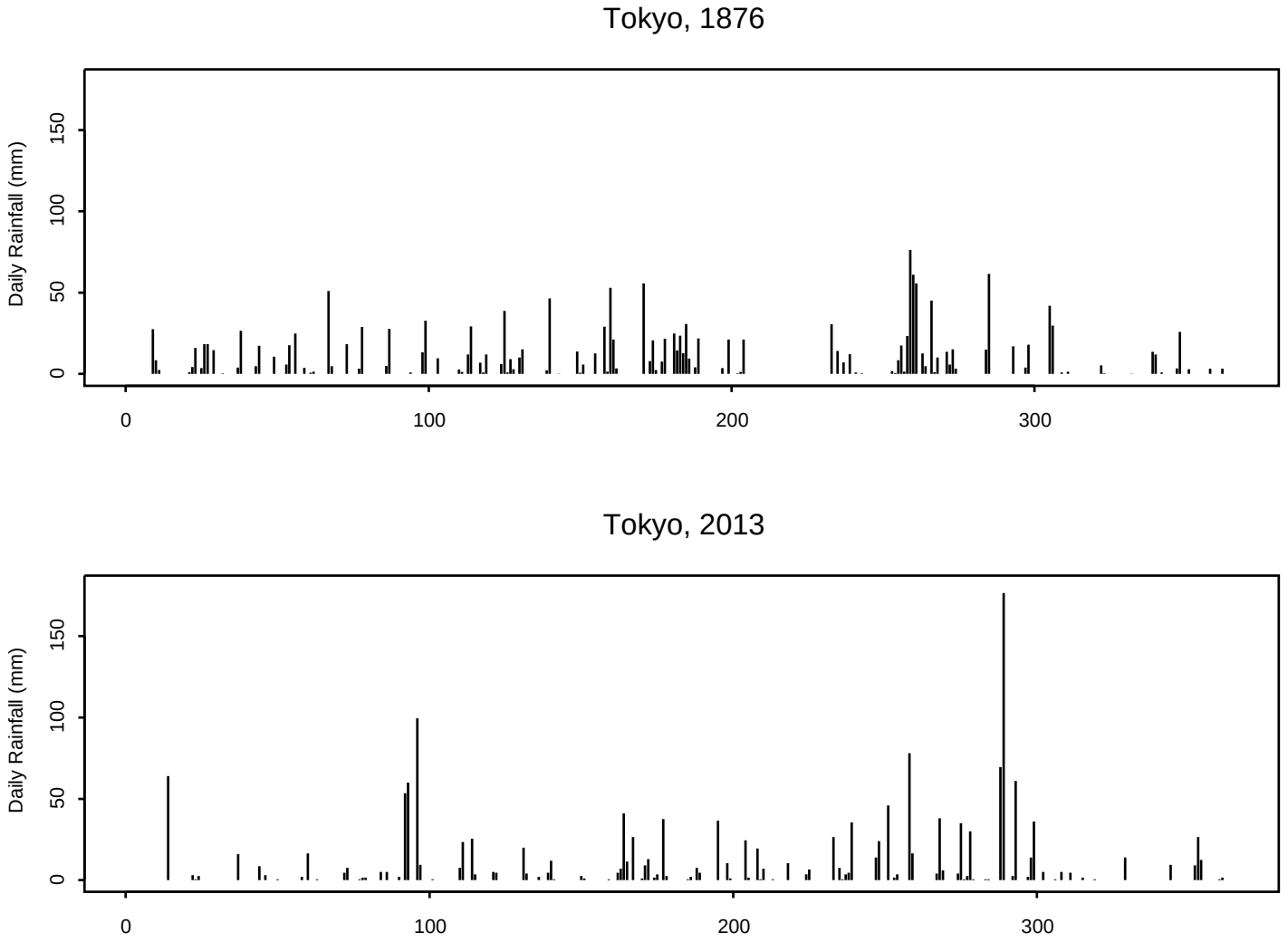
r-taka@maritime.kobe-u.ac.jp

1　はじめに

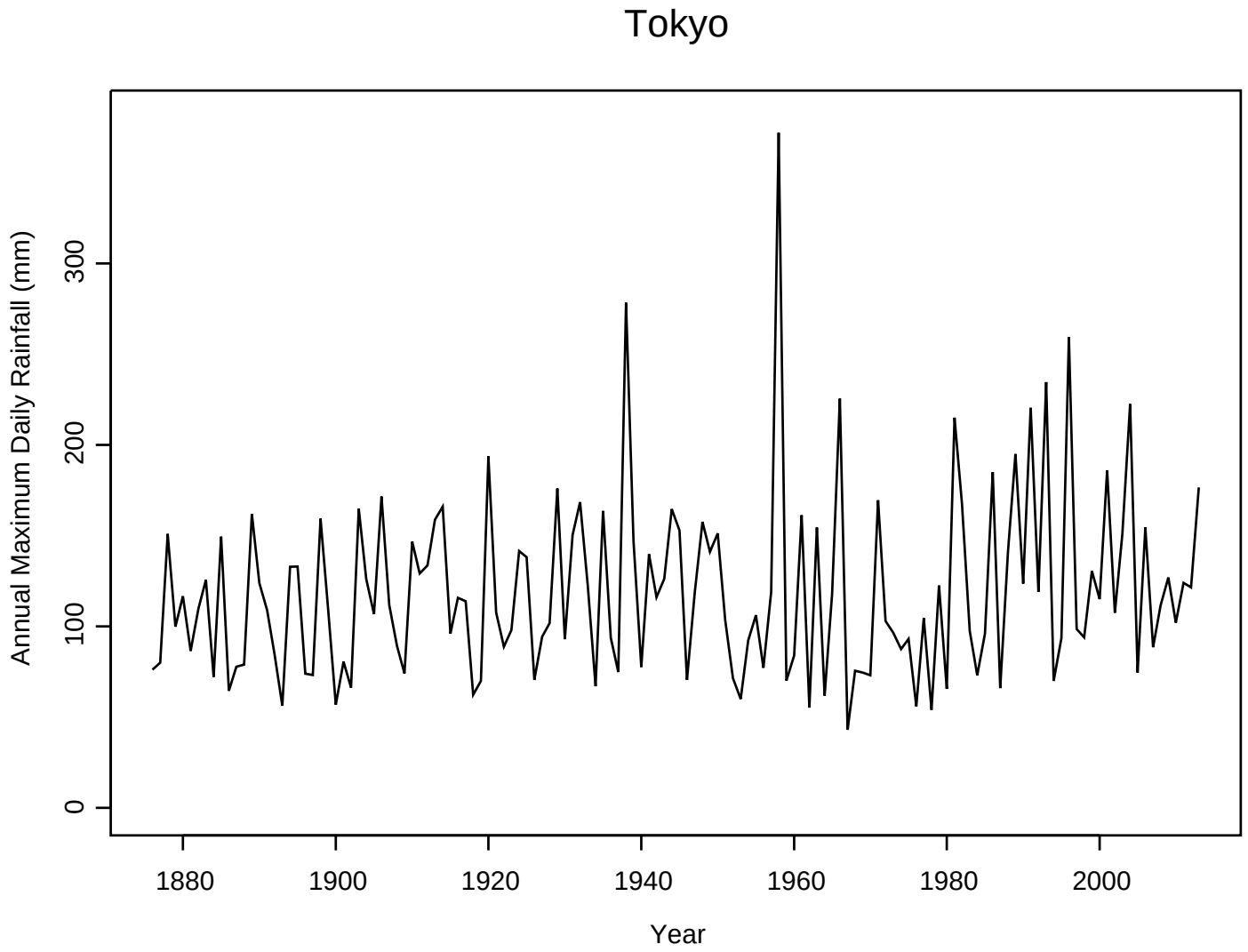
極値統計学で何が出来るか！



東京の日降水量(mm)，1876年1月1日～2013年12月31日.



東京の日降水量(mm)，1876年と2013年.



東京の年最大日降水量(mm)，1876年～2013年.

目標・目的

データから（与えられた空間や時間の中で）
『どの様な大きな値がどれくらいの確率で出現するのか？』
を知りたい.

そのためには
『極値データの確率構造』を明らかにしないといけない.

適切な統計モデルを作成しデータ解析を行う.

分野	数理統計学	信頼性工学	極値統計学
理論	中心極限定理	最弱リンクモデル	極値理論
適合分布	正規分布	ワイブル分布	一般極値分布 一般パレート分布
データ	ランダム	順序統計量	極値データ
目的	平均（分散）	信頼性	再現レベル

極値統計学

大きな値の出現に対して情報を持っている極値データ のみを考える.

データに適合させる分布は，一般極値分布 と 一般パレート分布.

参考文献

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

[2] Katz, R. W., Parlange, M. B. and Naveau, P. (2002). Statistics of extremes in hydrology. *Adv. Water Resour* **25**, 1287–1304.

[3] 高橋倫也，志村隆彰 (2015). 『極値統計学』. 近代科学社（準備中）.

1 変量の場合の極値統計学

A. 極値理論の基礎　　B. 尤度による推測法　　C. 実データ解析

以後の内容

２．極値理論

３．古典的極値データ解析法（GEV モデル，GP モデル）

４．点過程 (PP モデル)

５．東京の日降水量データ（1876年1月1日～2013年12月31日）解析

６．おわりに

レジメの正誤表

2　極値理論

○極値統計学の目的：与えられた観測期間中で大きな値をとる
　　データに関する推測（端の推測）．数理統計学では中心の推測.

○大きな値をとるデータに関して情報を持っている観測値

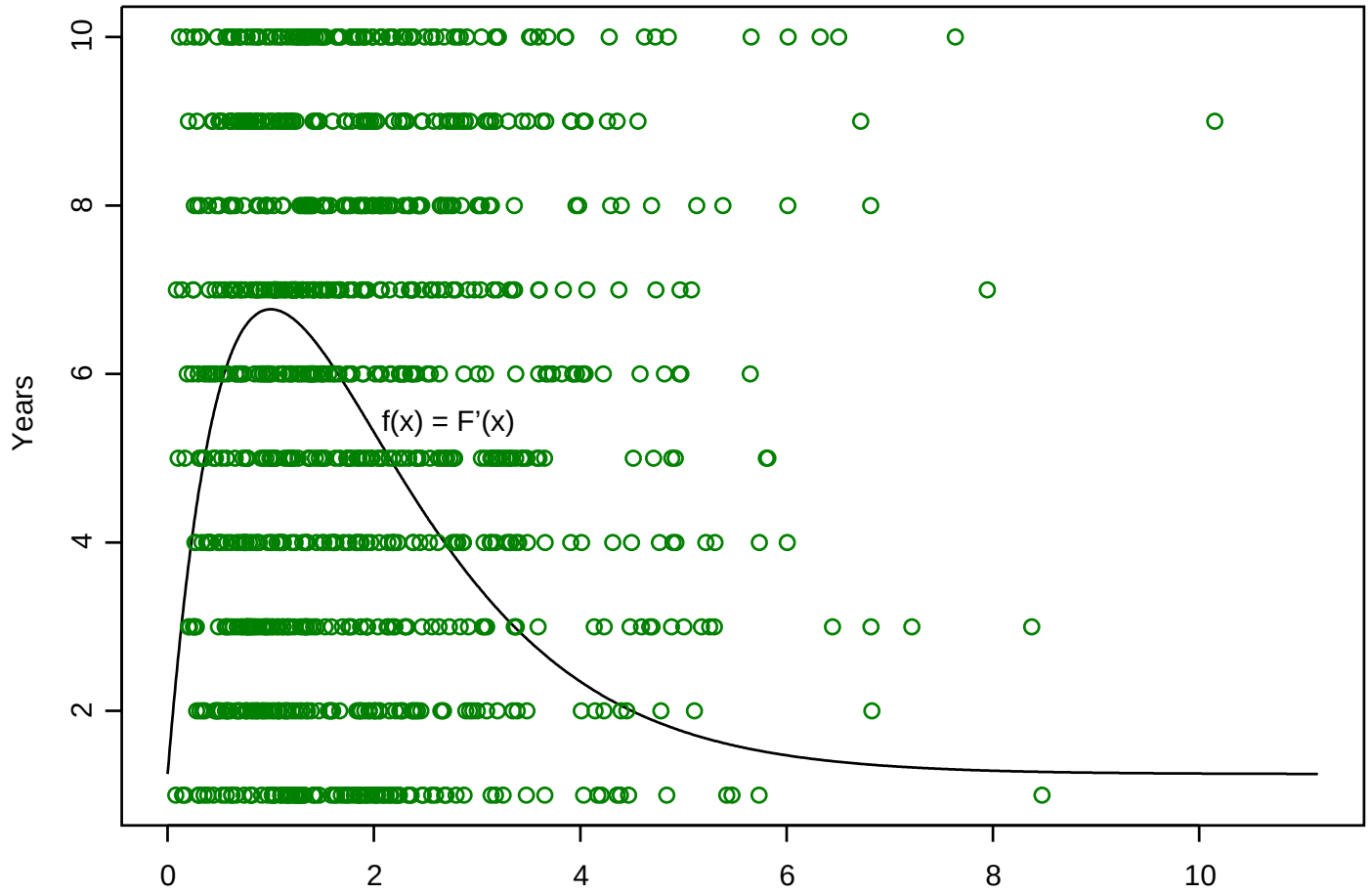
　ブロック最大データ　　Annual Maximum Series，AMS

　閾値超過データ　　Partial Duration Series，PDS

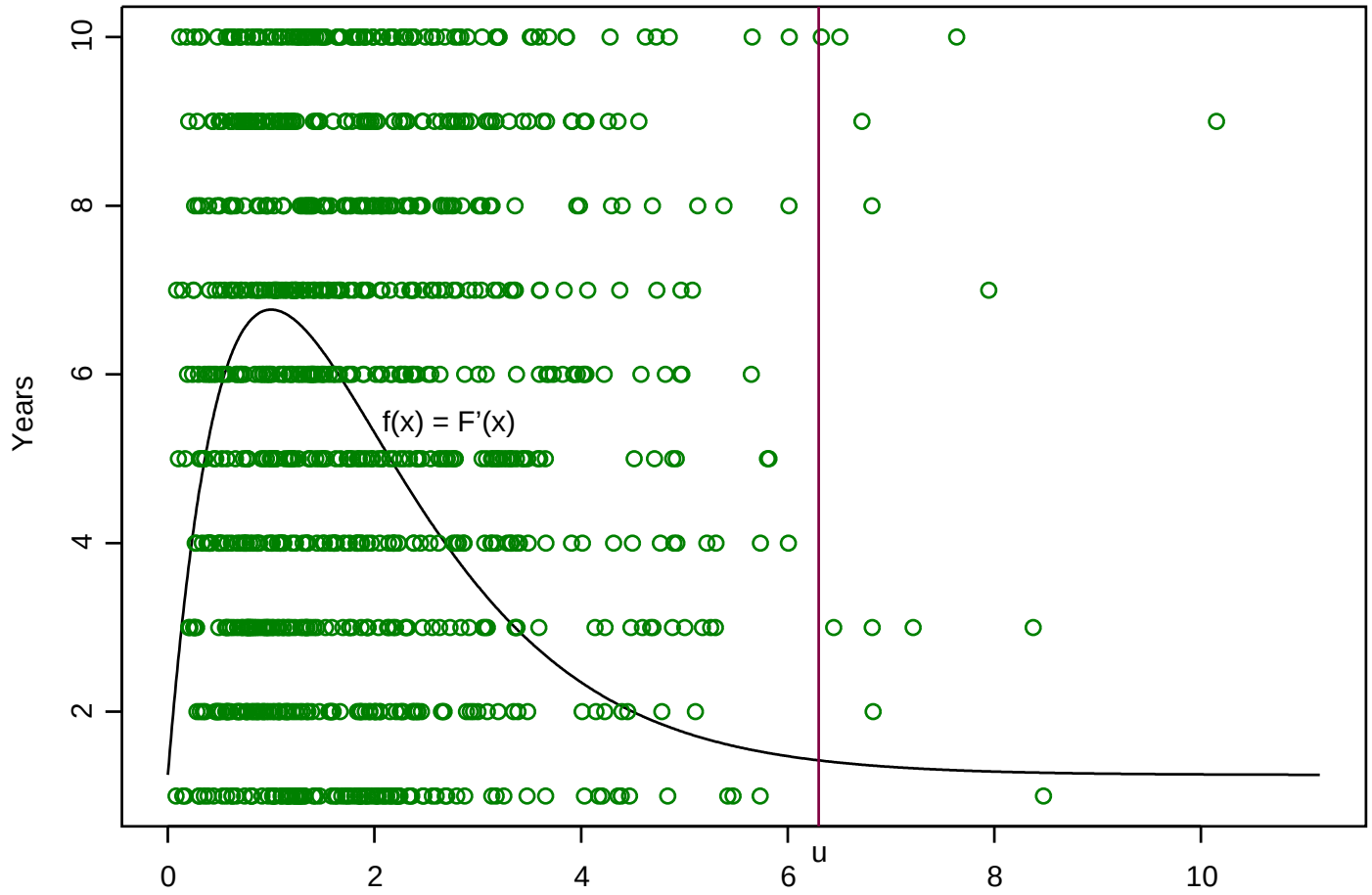
○ 背負い込んだ問題

　ブロック・サイズの決定　　（例えば年単位にする）

　閾値の選択



ブロック最大データ，AMS.　母集団分布の端.



閾値超過データ，PDS.　母集団分布の右裾.

ブロック最大データ と 閾値超過データ

○ 確率モデル　X：確率変数（例えば日降水量）

$F(x) = P(X \leq x)$：母集団分布　　$f(x)$：密度関数

○ブロック最大データ　AMS　n個の観測値の最大

$X_1, X_2, \ldots, X_n$：母集団分布 F からの確率標本

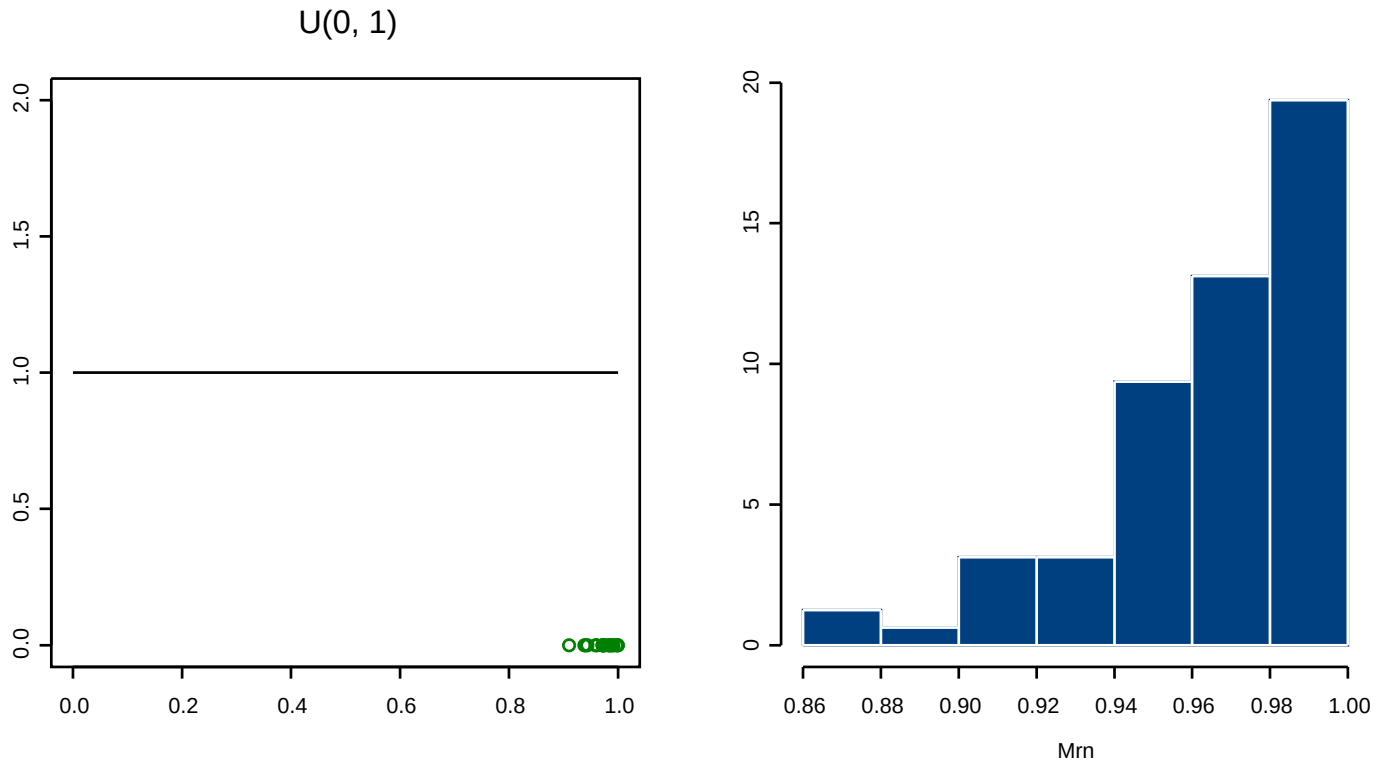
$Z_n = \max\{X_1, X_2, \ldots, X_n\}$：極値統計量

$P(Z_n \leq z) = F^n(z)$ の n が大きいときの分布？

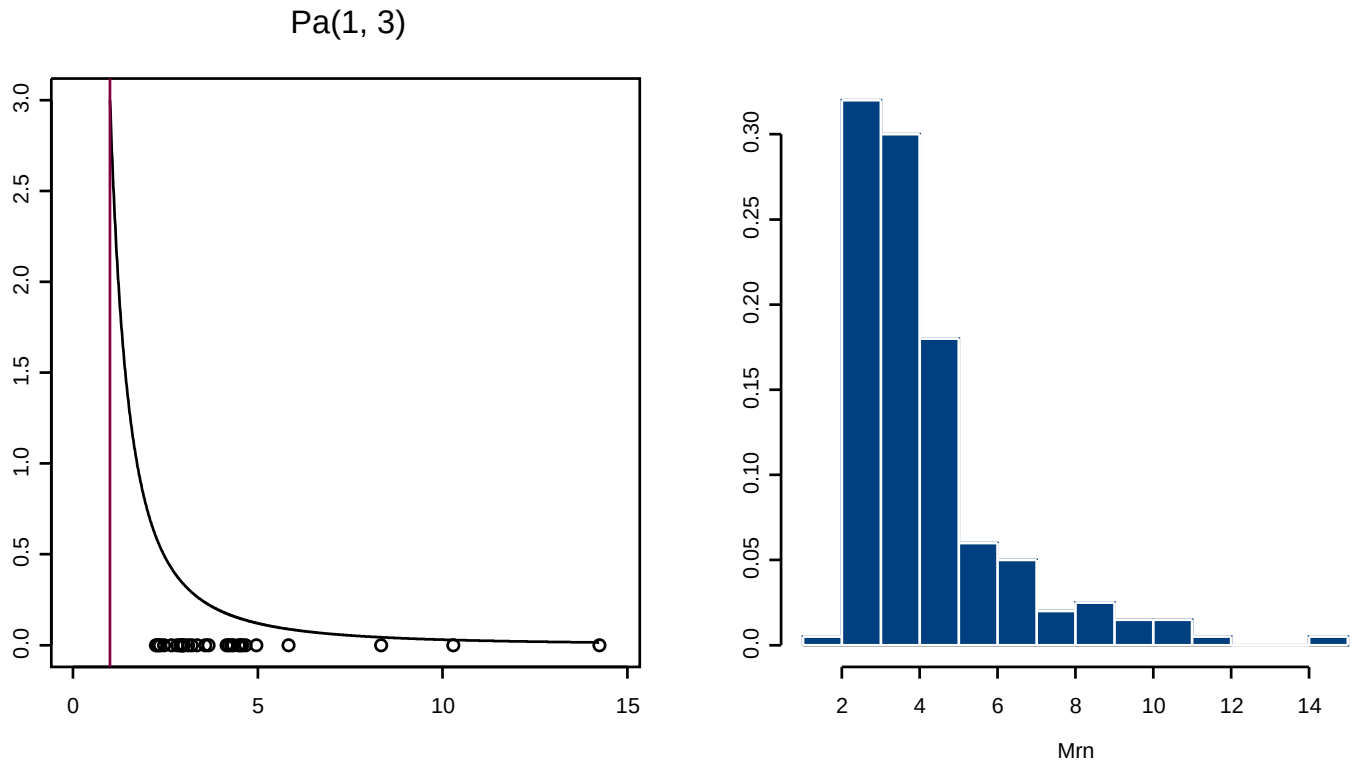
○閾値超過データ　　PDS　u：閾値

$X - u \,|\, X > u$ の u が十分大のときの分布？

$$P(X - u \leq y \,|\, X > u) = \frac{F(u+y) - F(u)}{1 - F(u)}, \quad y > 0.$$



一様分布からの 30個の最大値とヒストグラム．位置と尺度の変換.



パレート分布からの 30 個の最大値とヒストグラム．位置と尺度の変換.

極値統計量の基準化と極値分布

Z_n を基準化：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，

$$\frac{Z_n - b_n}{a_n} \xrightarrow{d} Z \text{ ：分布収束} \quad n \to \infty.$$

すなわち

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

G：極値分布 (extreme value distribution)

分布 F は極値分布 G の値吸引領域に属する：$F \in \mathcal{D}(G)$.

$(a_n,\ b_n)$：基準化定数 .

○標準一般極値分布

$$G_\xi(z) = \begin{cases} \exp[-(1+\xi z)^{-1/\xi}], & \xi \neq 0, \\ \exp[-\exp(-z)], & \xi = 0. \end{cases}$$

○母集団分布 F が適当な条件を満たし，ブロックの大きさ n が十分大

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

$a_n x + b_n = z$ とおくと

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

○ Z_n の分布は位置 b_n，尺度 a_n の一般極値分布 G_ξ で近似できる.

極値統計学の基本仮定：$F \in \mathcal{D}(G_\xi)$.

○統計学の教科書に出てくるほとんどの連続分布 F は $F \in \mathcal{D}(G_\xi)$.
理論的に「極値統計学の基本仮定」は保証される.

○ξ, a_n, b_n は（未知の）母集団分布 F に依存するので未知.

極値データ解析では $a_n = \sigma$, $b_n = \mu$ とおき，ブロック最大データに一般極値分布

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

を適合して $(\mu,\ \sigma,\ \xi)$ を未知パラメータとして推定する.

定義 1．次の分布を一般極値 (generalized extreme value) 分布 といい GEV$(\mu,\ \sigma,\ \xi)$ （$-\infty < \mu < \infty$，$\sigma > 0$，$-\infty < \xi < \infty$）で表す．

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

ただし，G_ξ は標準一般極値分布 GEV$(0,\ 1,\ \xi)$ の分布関数

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

とする．μ は位置，σ は尺度，ξ は形状パラメータ．

この一般極値分布をブロック最大データ AMS に適合して解析を行う．

一般極値分布 GEV$(\mu,\ \sigma,\ \xi)$ $G_\xi((z-\mu)/\sigma)$

$\xi < 0$ のときは Weibull 分布で $z < \mu - \sigma/\xi$,

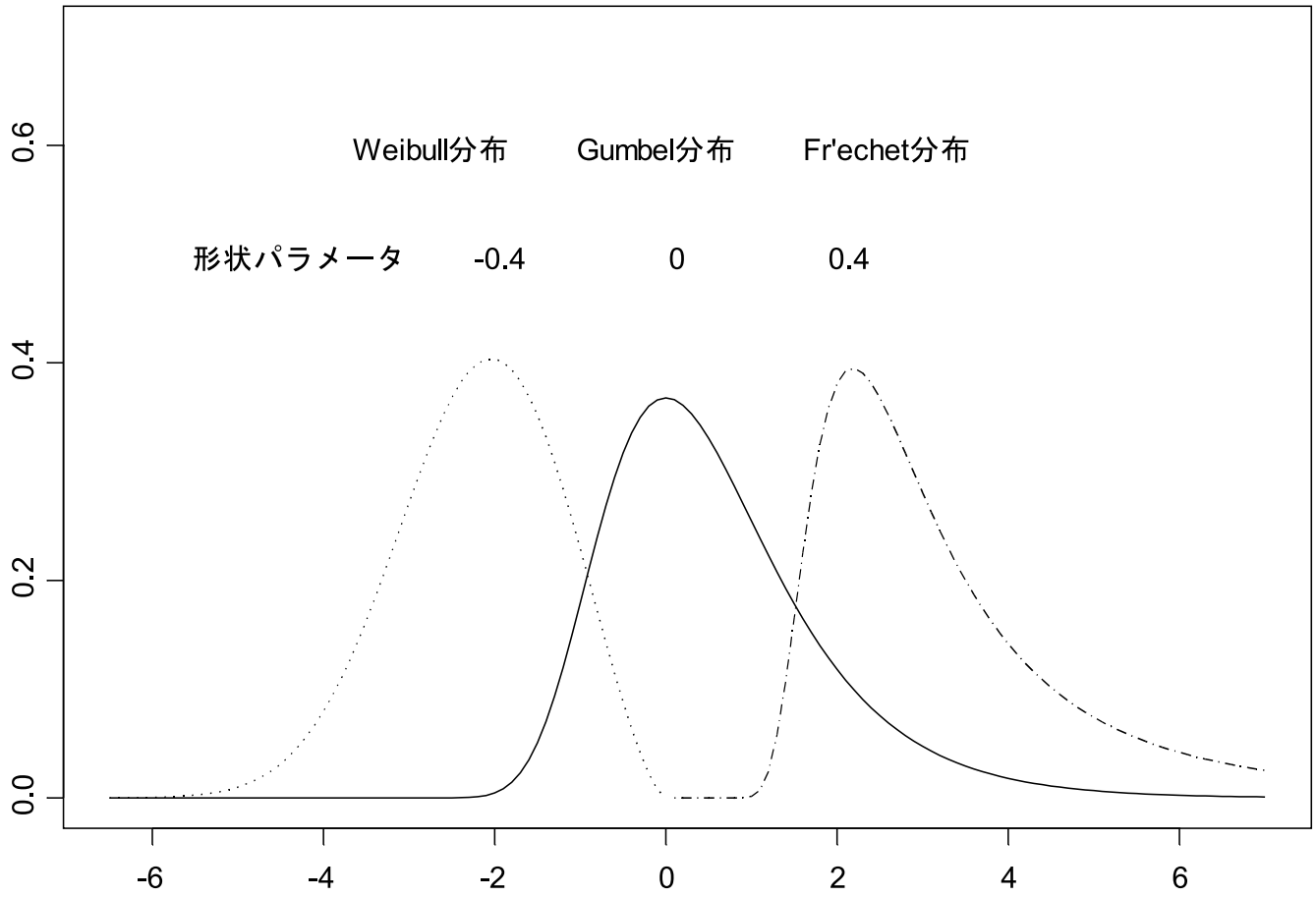
$\xi = 0$ のときは次から Gumbel 分布で $-\infty < z < \infty$,

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

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

一般極値分布 GEV$(0,\ 1,\ \xi)$ $G_\xi(z)$ の密度関数

$$g_\xi(z) = \begin{cases} (1+\xi\, z)^{-1/\xi-1} \exp\big[-(1+\xi\, z)^{-1/\xi}\big], & 1+\xi z > 0, \quad \xi \neq 0, \\ \exp\big[-z-\exp(-z)\big], & z \in \mathbb{R}, \quad \xi = 0. \end{cases}$$



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.

一般パレート分布による近似

○（標準）一般パレート (Generalized Pareto, GP) 分布：

$$H_\xi(x) = \begin{cases} 1 - (1+\xi x)^{-1/\xi}, & \xi \neq 0, \\ 1 - e^{-x}, & \xi = 0. \end{cases}$$

○ $F \in \mathcal{D}(G_\xi)$ のとき，u が十分大きければ

$$P(X - u \leq y \,|\, X > u) \approx H_\xi(y/\sigma_u).$$

ただし，$\sigma_u > 0$ は適当な定数.

○ 同じ形状パラメータ ξ が一般極値分布 G_ξ と一般パレート分布 H_ξ の両方に現れることに注意.

○上の主張の逆も言える.

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，ξ は形状パラメータ．

この一般パレート分布を閾値超過データ PDS に適合して解析を行う．

一般パレート分布 GP$(\sigma,\, \xi)$ $H_\xi(y/\sigma)$

$\xi < 0$ のときはベータ分布で $0 < y < -\sigma/\xi$,

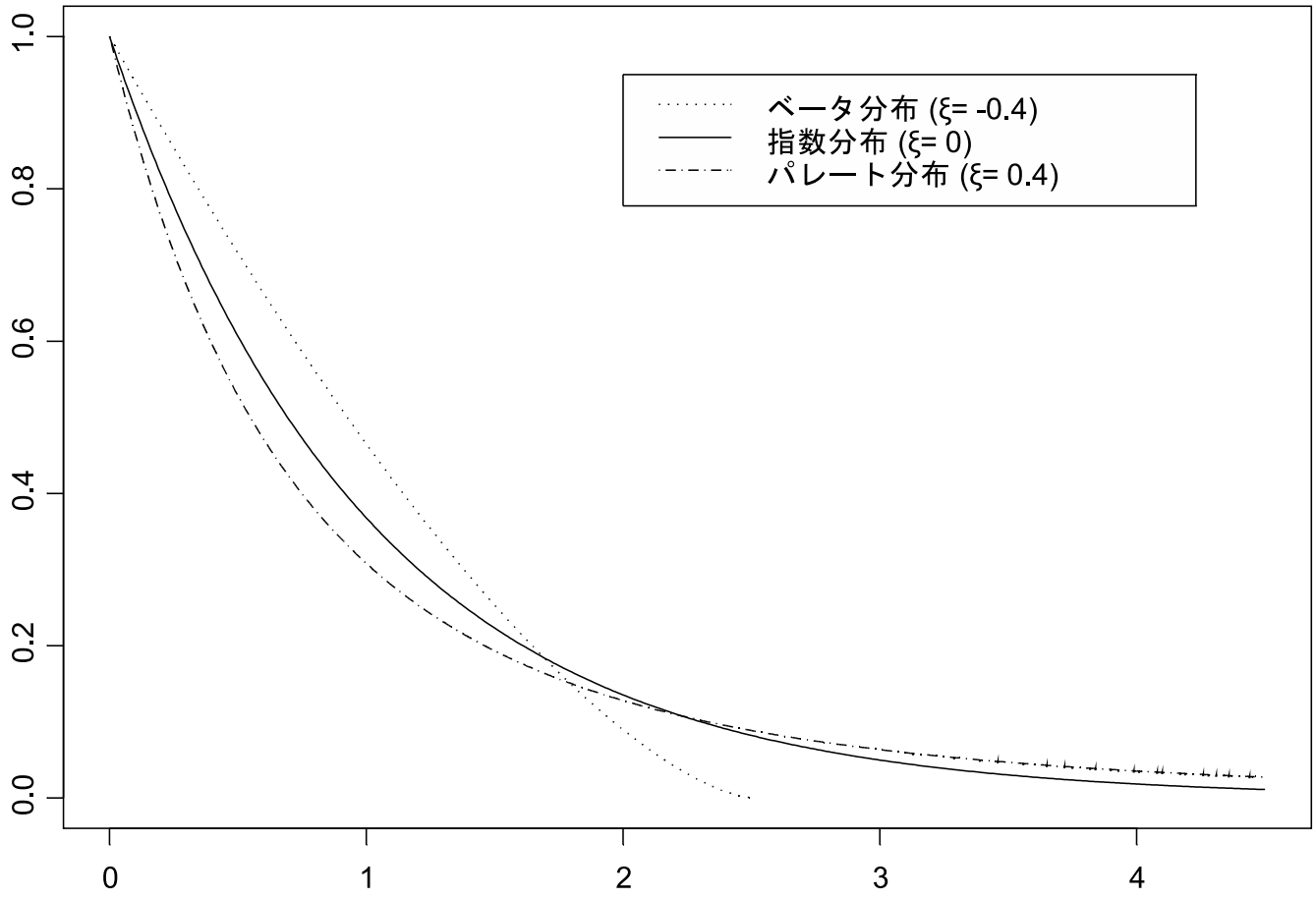
$\xi = 0$ のときは次より指数分布で $0 < y < \infty$,

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

$\xi > 0$ の場合はパレート分布で $0 < y < \infty$.

一般パレート分布 GP$(1,\, \xi)$ $H_\xi(y)$ の密度関数

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



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

3　古典的極値データ解析法

目的は未知の母集団分布の右裾（または左裾）に関する推測.

古典的な２つの極値データ解析法について紹介.
一般極値（GEV)モデル　と　一般パレート(GP)モデル

データの大きさ（サイズ）を n で表す.

一般極値（GEV) モデル

ブロック最大データ $\{z_1, z_2, \ldots, z_n\}$ に一般極値分布 GEV(μ, σ, ξ) を適合.

母集団分布は一般極値分布の吸引領域に属し，データは一般極値分布から得られたと仮定.

「一般極値分布の吸引領域に属する」はデータが得られない分布の上限領域に関する仮定で，それをデータから検証することは出来ない.

「推測による誤差」$=$ 「適合した一般極値分布が近似分布であることによる誤差」 $+$ 「推定による誤差」

データ解析結果の診断が重要.

最尤法

一般極値分布 GEV$(\mu,\, \sigma,\, \xi)$ を適合の場合の対数尤度

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

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

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

対数尤度を最大にする最尤推定値 $(\widehat{\mu},\, \widehat{\sigma},\, \widehat{\xi})$ を数値計算で求める.

最尤推定値は統計ソフト（例えばフリーのソフト R ）で簡単に求まる.

一般極値分布 GEV$(\mu,\, \sigma,\, \xi)$ の 期待情報行列

$I(\boldsymbol{\theta}) = I(\mu,\, \sigma,\, \xi)$ (Prescott and Walden, 1980) :

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

ただし $\boldsymbol{\theta} = (\mu,\, \sigma,\, \xi)$ の順で，$\Gamma(\,\cdot\,)$ はガンマ関数，$\psi(r) = d\log\Gamma(r)/dr$ で$p = (1+\xi)^2\Gamma(1+2\xi)$，$q = \Gamma(2+\xi)\{\psi(1+\xi) + (1+\xi)/\xi\}$，
$\gamma = 0.5772157...$ Euler の定数である．

パラメータ推定は最尤法で

$\{\mathrm{GEV}(\mu,\, \sigma,\, \xi),\ \mu \in \mathbb{R},\ \ \sigma > 0,\, \xi \in \mathbb{R}\}$ は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).

自然現象では $\xi \leq -0.5$ となることは稀：

Hosking *et al.* (1985) ：「年最大洪水ピーク流量資料」では

$$-0.5 < \xi < 0.5.$$

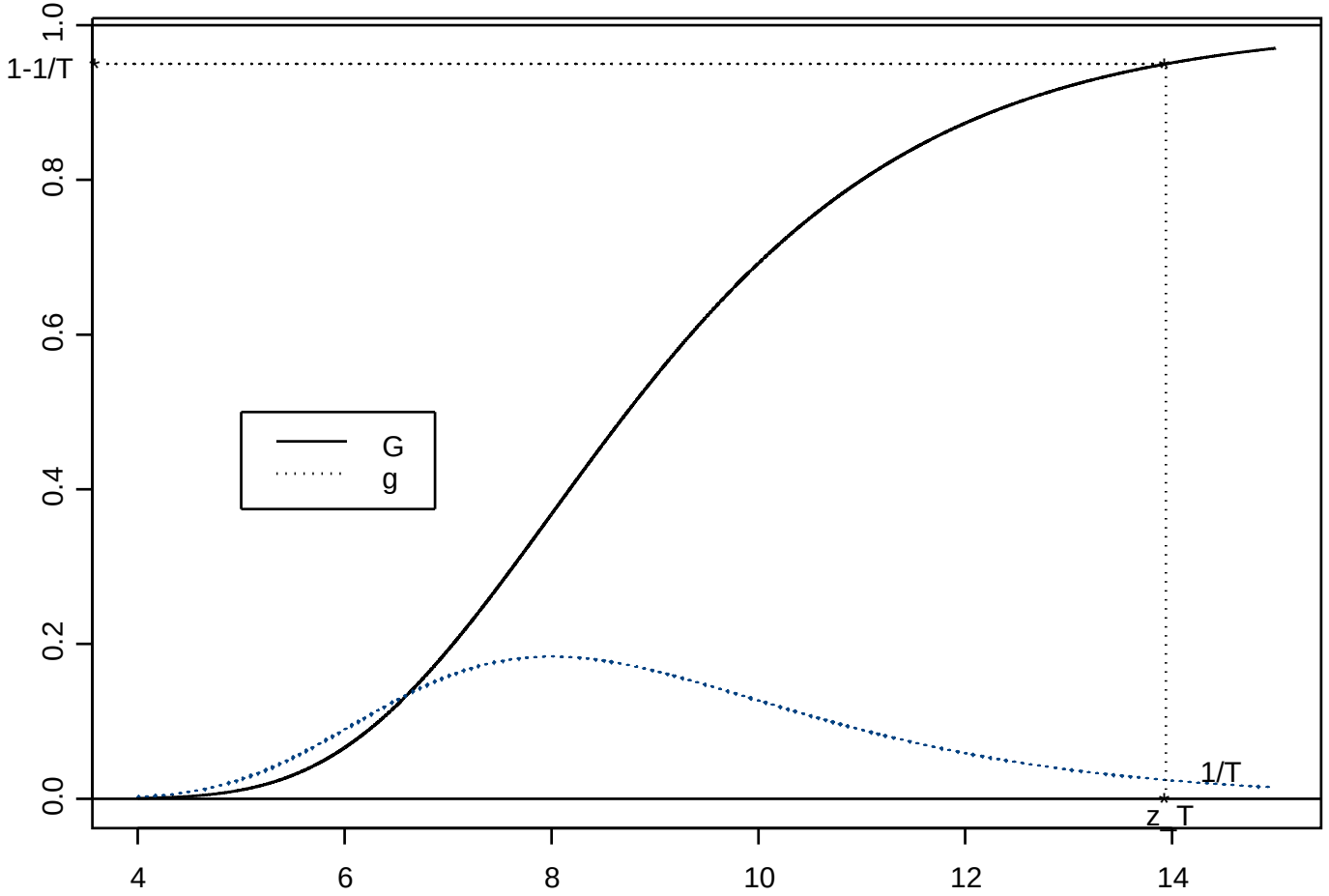
田中 (2010) ：「水文極値頻度解析（日本の日降水量データ）」では

$$-0.4 < \xi < 0.6.$$

最尤推定量 $\widehat{\boldsymbol{\theta}} = (\widehat{\mu},\, \widehat{\sigma},\, \widehat{\xi})^\top$ は，$\xi > -0.5$ のとき

$$\widehat{\boldsymbol{\theta}} \dot{\sim} \mathsf{N}(\boldsymbol{\theta},\, I(\boldsymbol{\theta})^{-1}/n)$$

$\dot{\sim}$ は近似的に従うことを表す.



z_T：再現期間 T 年の再現レベル.　　G：一般極値分布.

再現レベル

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

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

は

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

z_T は再現期間 (return period) T の再現レベル (return level)

例えば年最大値データを扱うとき，再現期間 $T = 100$ 年の再現レベル z_{100} は 100 年に平均 1 度現れる様な（大きな）値.

一般に $n \ll T$ の場合を考える，これはデータの存在しない領域の推測（外挿）.

再現レベル z_T の最尤推定値は，(μ, σ, ξ) の最尤推定値を用いて

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

デルタ法より標準誤差を求めることが出来る．

プロファイル信頼区間

形状パラメータ ξ の 95% の近似信頼区間：

$$\big\{\xi : 2\big\{l(\widehat{\mu}, \widehat{\sigma}, \widehat{\xi}) - \max_{\mu, \sigma} l(\mu, \sigma, \xi)\big\} \leq \chi^2_1(0.05)\big\}$$

非定常のモデル

次を考える：$i = 1, 2, \ldots, n$

$$\mu(t_i) = \alpha_0 + \alpha_1 t_i + \alpha_2 t_i^2, \quad \sigma(t_i) = \exp(\beta_0 + \beta_1 t_i), \quad \xi(t_i) = \gamma_0 + \gamma_1 t_i.$$

t_i は z_i の観測時点で $(\alpha_0, \alpha_1, \alpha_2, \beta_0, \beta_1, \gamma_0, \gamma_1)$ はパラメータ.

モデルを M_{ijk} $(i = 0, 1, 2,\ \ j = 0, 1,\ \ k = 0, 1)$ で表す．M_{ijk} では $\mu(t)$，$\log \sigma(t)$，$\xi(t)$ はそれぞれ i，j，k 次の多項式．例えば，M_{110} は

$$\mu(t_i) = \alpha_0 + \alpha_1 t_i, \quad \sigma(t_i) = \exp(\beta_0 + \beta_1 t_i), \quad \xi(t_i) = \xi = \gamma_0$$

のモデルになる．$\sigma(t) = \exp(\beta_0 + \beta_1 t)$ は，$\sigma(t) > 0$ を保証するため.

モデル（$3 \times 2 \times 2 = 12$ 個）の中で統計的に最適なものをAICで選択.

一般パレート(GP)モデル

閾値超過データ $\{y_1, y_2, \ldots, y_n\}$ に一般パレート分布 GP(σ, ξ) を適合.
閾値超過データは一般パレート分布からのものと仮定.

最尤法
一般パレート分布 GP(σ, ξ) の対数尤度

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

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

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

最尤推定量の性質

一般パレート分布 GP(σ, ξ) の期待情報量行列

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

$\xi > -1/2$ ならば情報行列は有限で，n が十分大のとき，最尤推定量は漸近的に平均 $(\sigma, \xi)^\top$，分散共分散行列が

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

の２変量正規分布に従い漸近有効推定量となる（Smith, 1985）.

閾値の選択

応用上データに一般パレート分布を適合させ解析するには,
閾値 (threshold) の選択が必要.

閾値の選択には一般パレート分布の性質を利用する.

一般パレート分布の性質　$Y \sim \mathrm{GP}(\sigma, \xi)$

○ $\xi < 1$ で平均は存在

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

ただし $\omega = \sup\{y \,|\, H_\xi(y/\sigma) < 1\}$.

○ $u > 0$ のときの条件付き確率変数 $Y - u \,|\, Y > u$ の分布

$$P(Y - u > y \,|\, Y > u) = \frac{1 - H_\xi((y+u)/\sigma)}{1 - H_\xi(u/\sigma)} = \frac{\left(1 + \xi(y+u)/\sigma\right)^{-1/\xi}}{\left(1 + \xi u/\sigma\right)^{-1/\xi}}$$

$$= \left(1 + \xi \frac{y}{\sigma + \xi u}\right)^{-1/\xi}$$

同じ形状パラメータ ξ の一般パレート分布 $\mathrm{GP}(\sigma + \xi u,\, \xi)$ に従う.

○ $Y - u \,|\, Y > u \sim \mathrm{GP}(\sigma_u,\, \xi_u)$, $\sigma_u = \sigma + \xi_u\, u, \quad \xi_u = \xi$.
これから

$$\sigma = \sigma_u - \xi_u\, u \text{（修正尺度）}, \quad \xi = \xi_u \text{：一定.}$$

○ $e(u)$：Y の 平均超過 (mean excess) 関数

$$e(u) = E(Y - u \,|\, Y > u)$$

$Y - u \,|\, Y > u \ \sim \mathrm{GP}(\sigma + \xi u,\, \xi)$ より

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

u の一次関数．特に，指数分布（$\xi = 0$）の場合は $e(u)$ 定数.

○ $\widehat{e}_n(u)$：標本平均超過関数

$$\widehat{e}_n(u) = \frac{1}{N_u} \sum_{i=1}^{n} (X_i - u)_+, \quad N_u：u より大のデータ数$$

ただし，$X_1, X_2, \ldots, X_n$ は生のデータで，$(a)_+ = \max(a,\, 0)$.

閾値の選択法

1) 修正尺度と形状パラメータのプロット　値 u を動かして，各 u を超過したデータに一般パレート分布 GP(σ_u, ξ_u) を適合し形状と尺度パラメータの最尤推定値 $(\widehat{\sigma}_u, \widehat{\xi}_u)$ を求める．修正尺度の推定値 $\widehat{\sigma} = \widehat{\sigma}_u - \widehat{\xi}_u u$ と $\widehat{\xi}_u$ を u に対してプロットした図で，その値より右側では２つの推定値が一定になっていると見なせる最小の値を閾値とする．

2) 標本平均超過関数プロット　値 u を動かして，各u に対して標本平均超過関数を描いた図で，それより右側で関数が直線に近いと見なせる最小の値を閾値とする．

バリューアットリスク，m 観測再現レベル

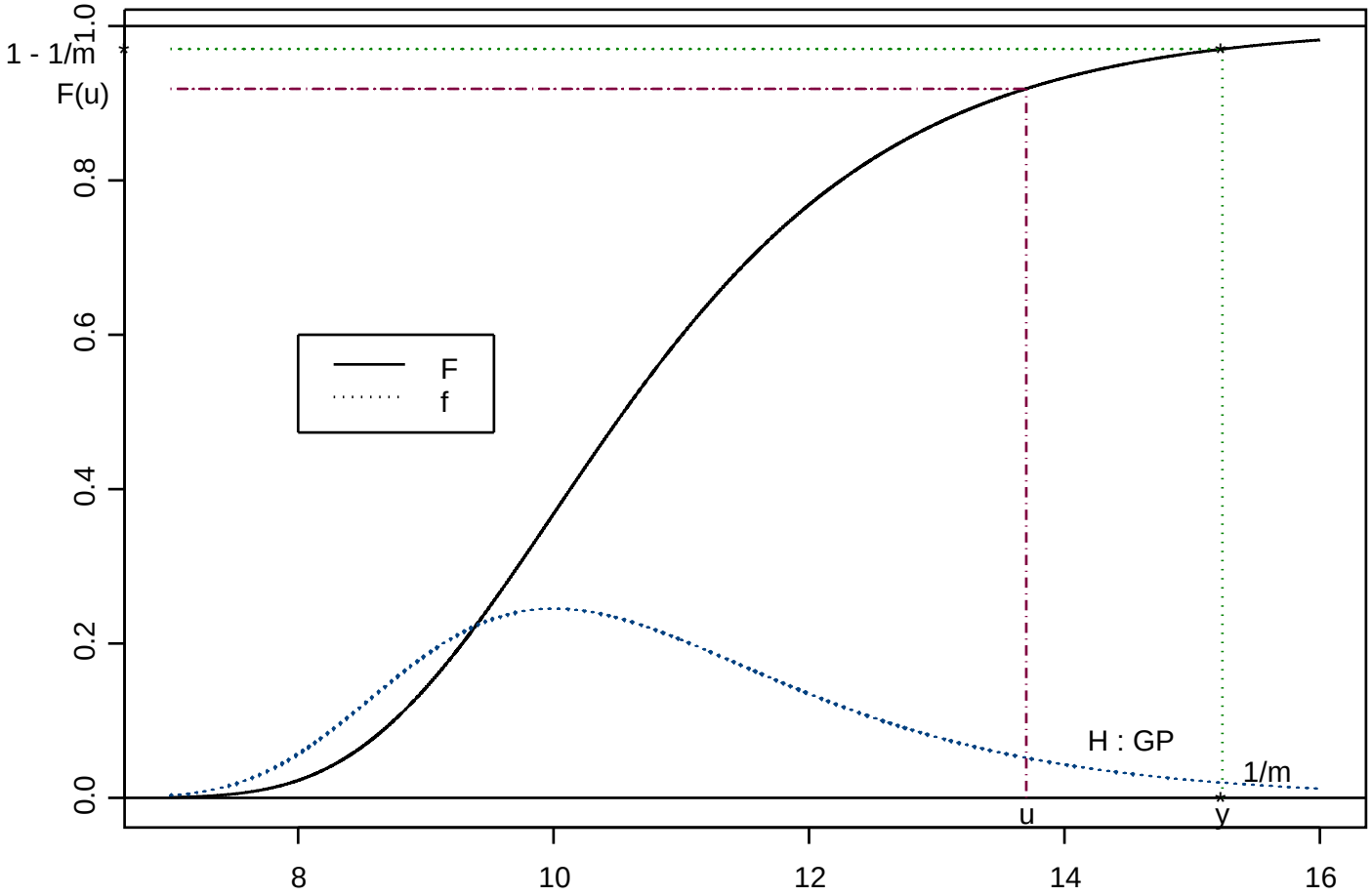
極値統計学では，母集団分布 F の上側微小確率点の推定が目的の場合が多い．

分布 F で

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

となる確率点 $y_p = \mathsf{VaR}_p$ は最近ファイナンスの分野でバリューアットリスク (Value–at–Risk) とよばれる．$p = 1/m$ として m 観測再現レベルとよぶこともある．すなわち，m 回の観測で平均一度 $y_{1/m}$ 以上の値が観測される．

以下，おおきさ n の生の観測データが与えられているとする．このデータから閾値 u を決定し y_p を推定する．



$y_{1/m}$：m 観測再現レベル.　　F：母集団分布.

母集団分布 $F(x) = P(X \leq x)$ を次の様に分解：$x > u$

$$
\begin{aligned}
P(X \leq x) &= P(X \leq u) + P(u < X \leq x) \\
&= P(X \leq u) + \frac{P(u < X \leq x)}{P(X > u)}\, P(X > u) \\
&= P(X \leq u) + P(X - u \leq x - u \,|\, X > u)\, P(X > u)
\end{aligned}
$$

十分大きい u に対して $P(X - u \leq x - u \,|\, X > u)$ を GP 分布 H_ξ で置き換え

$$
F(x) = F(u) + H_\xi \left(\frac{x - u}{\sigma} \right) [1 - F(u)]
$$

と仮定．ここで $\zeta_u = 1 - F(u)$ とおくと，$F(y_p) = 1 - p$ より

$$
y_p = u + \frac{\sigma}{\xi} \left\{ \left(\frac{\zeta_u}{p} \right)^{\xi} - 1 \right\}
$$

となる．

閾値uを選択し，閾値を超過するデータ（その個数を N_u とする）で分布 GP(σ, ξ) のパラメータの最尤推定値$(\widehat{\sigma}, \widehat{\xi})$を求める．また，$\zeta_u$ は N_u/n で推定する．

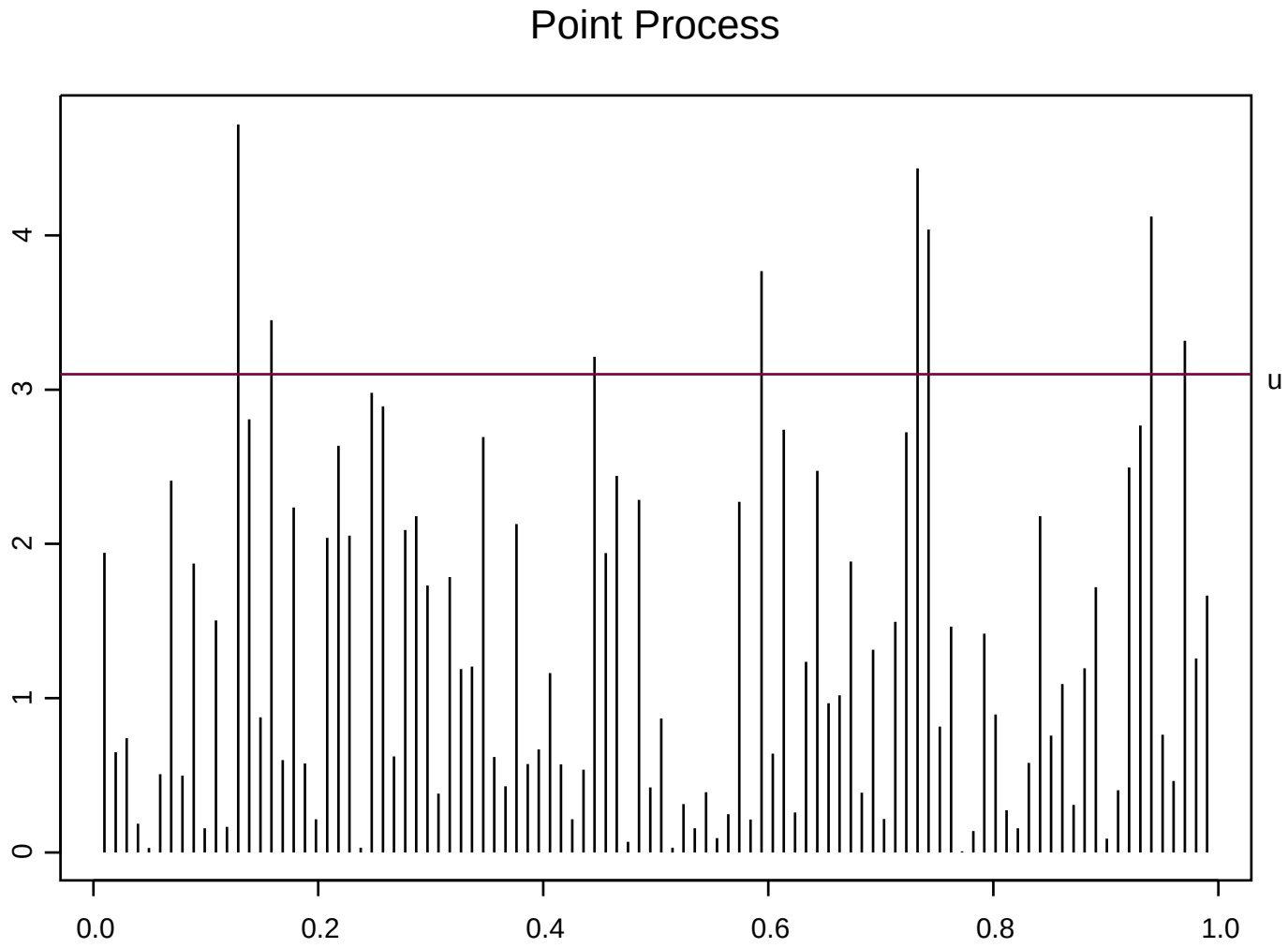
これらを代入して確率点 y_p の最尤推定値

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

を得る．推定の標準誤差はデルタ法から求まる．

プロファイル尤度を用いた ξ の 95% 近似信頼区間

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



点過程 PP.

4　点過程モデル

柔軟な極値データ解析が可能.

独立で同一分布 F に従う確率変数列 $X_1, X_2, \ldots$ を考える．$F \in \mathcal{D}(G_\xi)$ と仮定すると，定数列 $a_n > 0$，$b_n \in \mathbb{R}$ が存在して

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

が成立．ここで，$1 - F(a_n z + b_n)$ は基準化した確率変数 $(X_i - b_n)/a_n$ が閾値 z を超える確率．よって，$n[1 - F(a_n z + b_n)]$ は基準化した n 個の確率変数 $(X_1 - b_n)/a_n, \ldots, (X_n - b_n)/a_n$ が閾値 z を超える平均個数になる．もし n が十分大であれば，ポアソンの小数の法則から，閾値 z を超える標本数はポアソン分布で近似できる.

定理４．互いに独立に同一分布 F に従う確率変数列を $X_1, X_2, \ldots$ とし，$Z_n = \max_{1 \le i \le n} X_i$ に対して $a_n > 0$，$b_n \in \mathbb{R}$ が存在して，

$$P\{(Z_n - b_n)/a_n \le z\} \to G_\xi(z) = \exp[-(1+\xi z)^{-1/\xi}], \quad n \to \infty$$

とする．また，α，ω をそれぞれ分布 F の下限，上限とする．このとき点過程列

$$\left\{ \left(\frac{i}{n+1}, \frac{X_i - b_n}{a_n} \right) : \ i = 1, \ldots, n \right\}$$

は $n \to \infty$ のとき，任意の $z > \alpha$ に対して，領域 $[0,\, 1] \times (z,\, \omega)$ でポアソン過程に収束し，$A = [t_1,\, t_2] \times (z,\, \omega)$ $([t_1,\, t_2] \subset [0,\, 1])$ の平均強度は $\Lambda(A) = (t_2 - t_1)(1+\xi z)^{-1/\xi}$ で与えられる．

データ解析では，基準化定数 $(a_n,\, b_n)$ は未知，これを $(\sigma,\, \mu)$ と置き，次の点過程

$$\mathbb{P}_n = \left\{ \left(\frac{i}{n+1},\, X_i \right) :\ \ X_i > u,\ i = 1, \ldots, n \right\}$$

を考える．このとき

$$P(X_i > u) = P\left(\frac{X_i - b_n}{a_n} > \frac{u - b_n}{a_n} \right) = P\left(\frac{X_i - b_n}{a_n} > \frac{u - \mu}{\sigma} \right)$$

で，$x = (u - b_n)/a_n$ とおくと

$$u = a_n x + b_n \to \omega_F = \sup\{x \,|\, F(x) < 1\}, \quad n \to \infty$$

となる．ポアソン過程の近似を保証するためには閾値 u は十分大きく取る．

極値統計学の基本仮定 の下で

点過程 $\mathbb{P}_n$ は u が十分大のとき，$A = [t_1,\, t_2] \times (u,\, \omega)$ の平均強度が

$$\Lambda(A) = (t_2 - t_1)\left[1 + \xi\left(\frac{u-\mu}{\sigma}\right)\right]^{-1/\xi}$$

で与えられるポアソン過程 $\mathbb{P}$，PP$(\mu,\, \sigma,\, \xi)$，で近似できる．閾値 u は，漸近理論が使えるように選択する．
選択した閾値 u を超過するデータを考える．n_y 年間のデータで領域 $A = [0,\, 1] \times (u,\, \omega)$ に入っている点を

$$\left\{(t_1,\, x_1), \ldots, (t_{N(A)},\, x_{N(A)})\right\}$$

とする．領域 A 内では $\mathbb{P}_n \approx \mathbb{P}$ である．

近似的な尤度：観測年数を n_y として

$$\Lambda(A) = n_y \left[1 + \xi \left(\frac{u-\mu}{\sigma}\right)\right]^{-1/\xi}$$

とおけば，(μ, σ, ξ) は年最大分布 (GEV) のパラメータに相当する．このとき尤度は

$$L_A(\mu, \sigma, \xi; x_1, \ldots, x_{N(A)}) = \exp\{-\Lambda(A)\} \prod_{i=1}^{N(A)} \lambda(t_i, x_i)$$

$$\propto \exp\left\{-n_y \left[1 + \xi \left(\frac{u-\mu}{\sigma}\right)\right]^{-1/\xi}\right\} \prod_{i=1}^{N(A)} \frac{1}{\sigma} \left[1 + \xi \left(\frac{x_i-\mu}{\sigma}\right)\right]^{-1/\xi-1}$$

と表される．ただし，$\lambda(t, x) = [1 + \xi(x-\mu)/\sigma]^{-1/\xi-1}/\sigma$.

この尤度を最大化して最尤推定値 $(\widehat{\mu}, \widehat{\sigma}, \widehat{\xi})$ を求める．

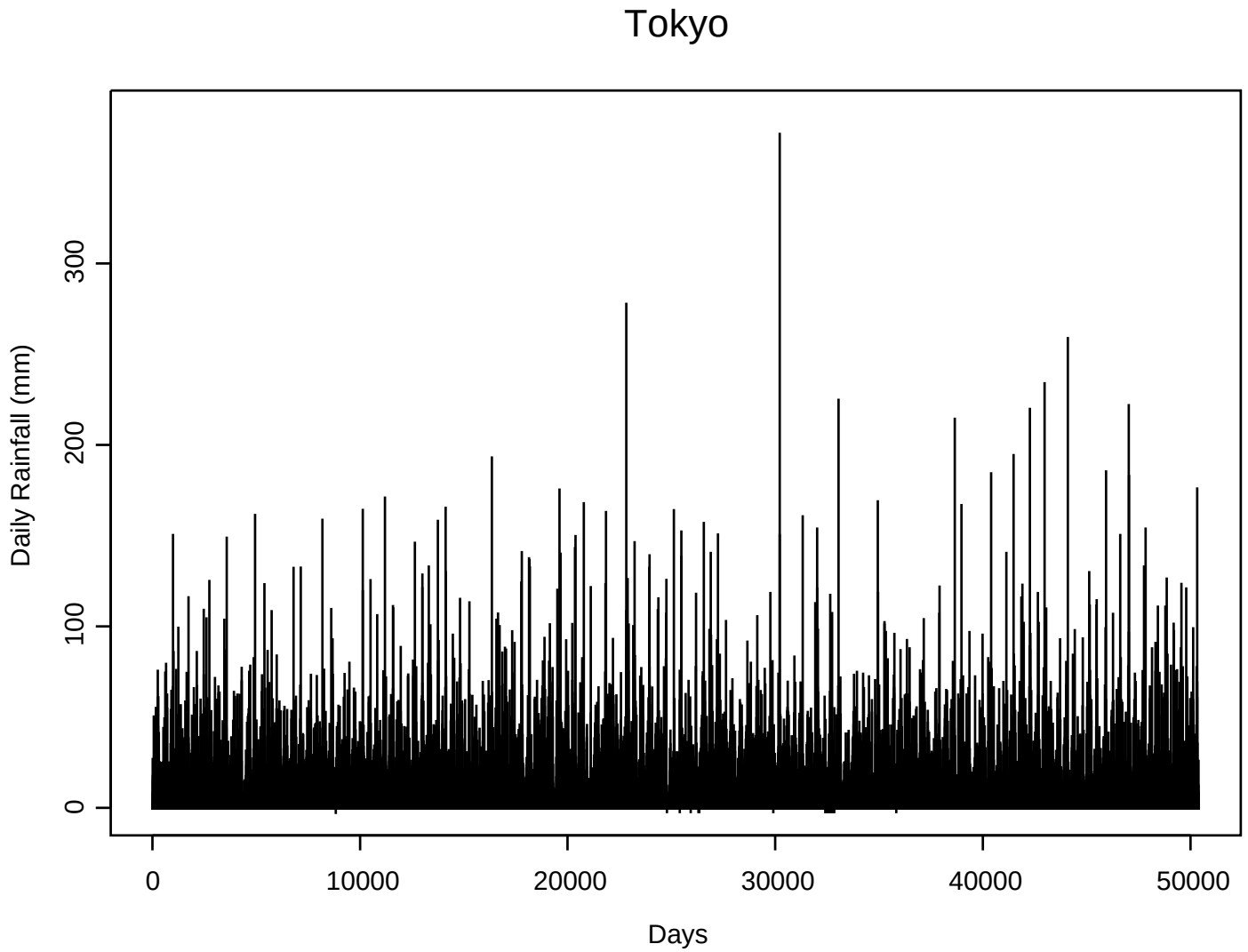
5　東京の日降水量データの極値解析

東京の1876年1月1日から2013年12月31日までの138年間の日降水量 (mm) データの極値解析.

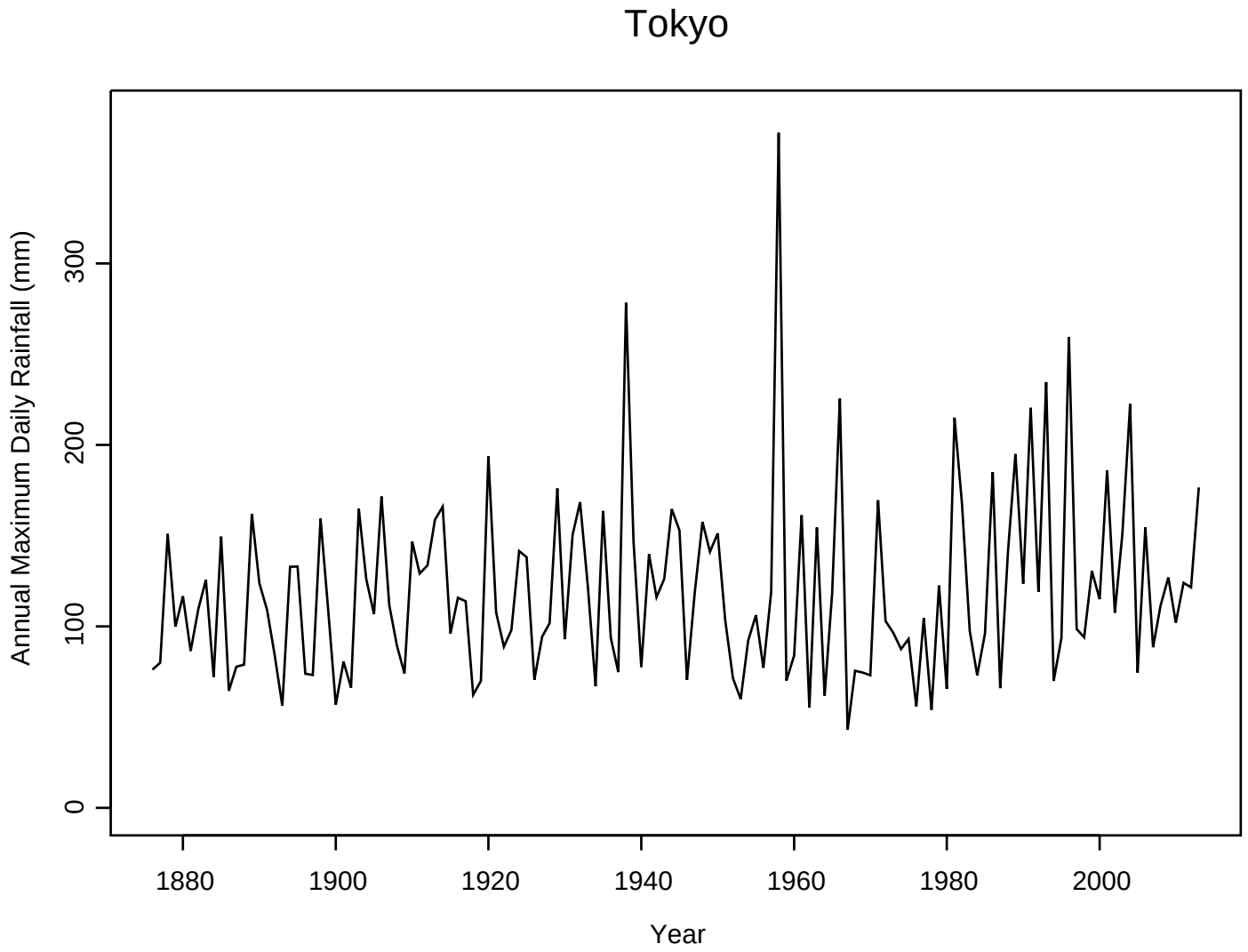
データの中には33個の欠測値等があるが，それらはすべて 0 (mm) として処理．欠測日の前後の日の測定値等から，それらは日降水量の最大に関して影響がないと判断.

以下，GEV，GP，PPの３モデルによる解析結果を紹介する.

138年 $=$ 50404日，　降雨の観測日数 $=$ 18343日　(36.4%)，
33日/138年 $=$ 0.0006546



東京の日降水量(mm)，1876年1月1日～2013年12月31日.



東京の年最大日降水量(mm)，1876年～2013年.

一般極値 GEV モデルによる解析

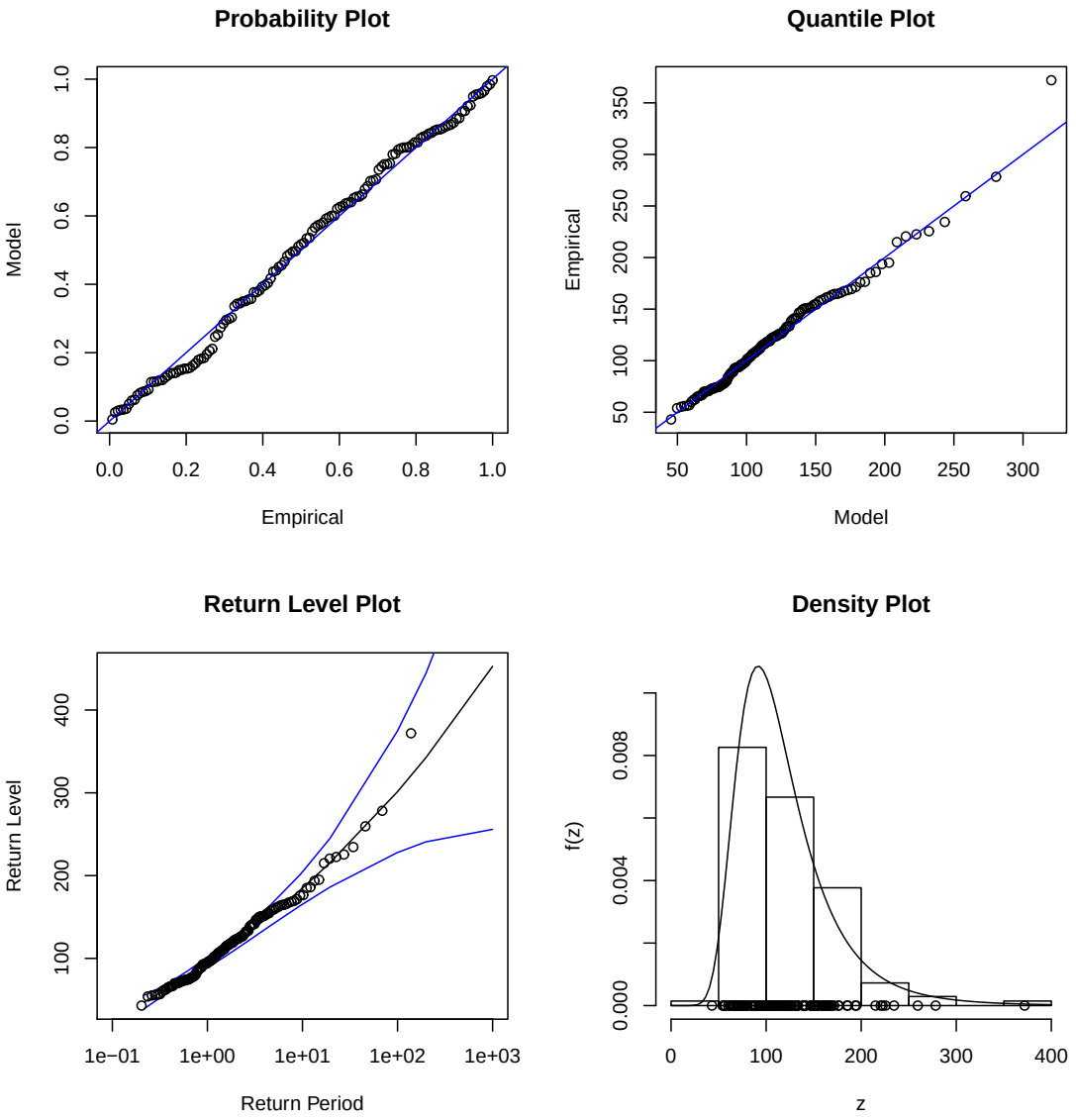
年最大日降水量データの最小値は 43.0 で 最大値は 371.9.

年最大日降水量データに一般極値分布 GEV(μ, σ, ξ) を適合して解析.

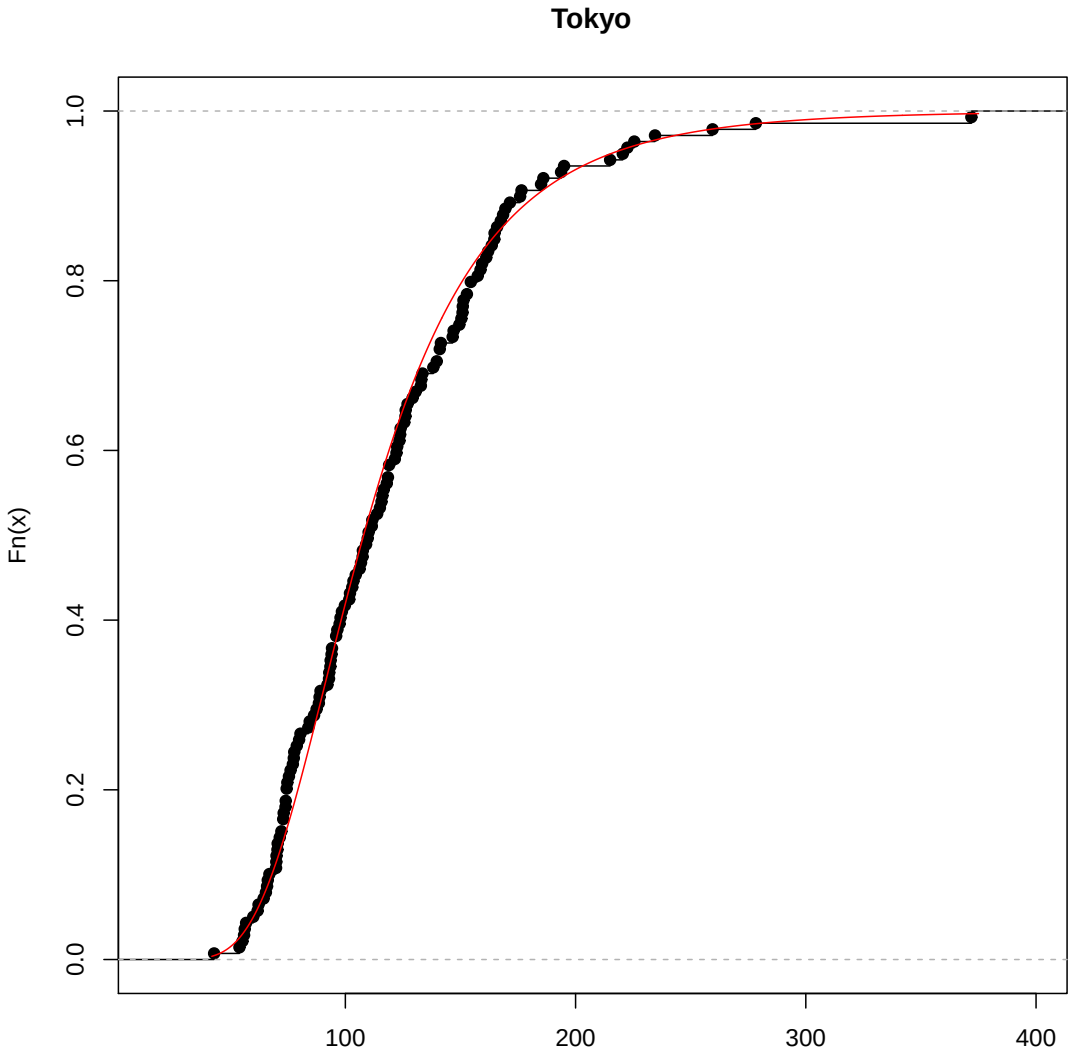
最大対数尤度は -713.8478，最尤推定値（標準誤差）

$$\widehat{\mu} = 95.17\,(3.35), \quad \widehat{\sigma} = 34.08\,(2.58), \quad \widehat{\xi} = 0.114\,(0.075).$$

形状パラメータ ξ の推定値 0.114 は正で，最大値の分布は Fréchet 分布と推定され，非常に大きな値が観測される可能性がある.


Probability Plot
Model
Empirical
0.0
0.2
0.4
0.6
0.8
1.0
Quantile Plot
Empirical
Model
50
100
150
200
250
300
350
Return Level Plot
Return Level
Return Period
100
200
300
400
1e−01
1e+00
1e+01
1e+02
1e+03
Density Plot
f(z)
z
0.000
0.004
0.008
0
100
200
300
400

経験分布関数 G_n と推定分布関数 $\widehat{G}$（赤）.



経験分布関数　$G_n(z) = \dfrac{i}{n+1}, \quad z_{(i)} \leq z < z_{(i+1)}.$

GEVモデルによる解析診断

$z_{(1)} \leq z_{(2)} \leq \cdots \leq z_{(n)}$：ブロック最大データを大きさの順に並べたもの

確率プロット (Probability Plot)

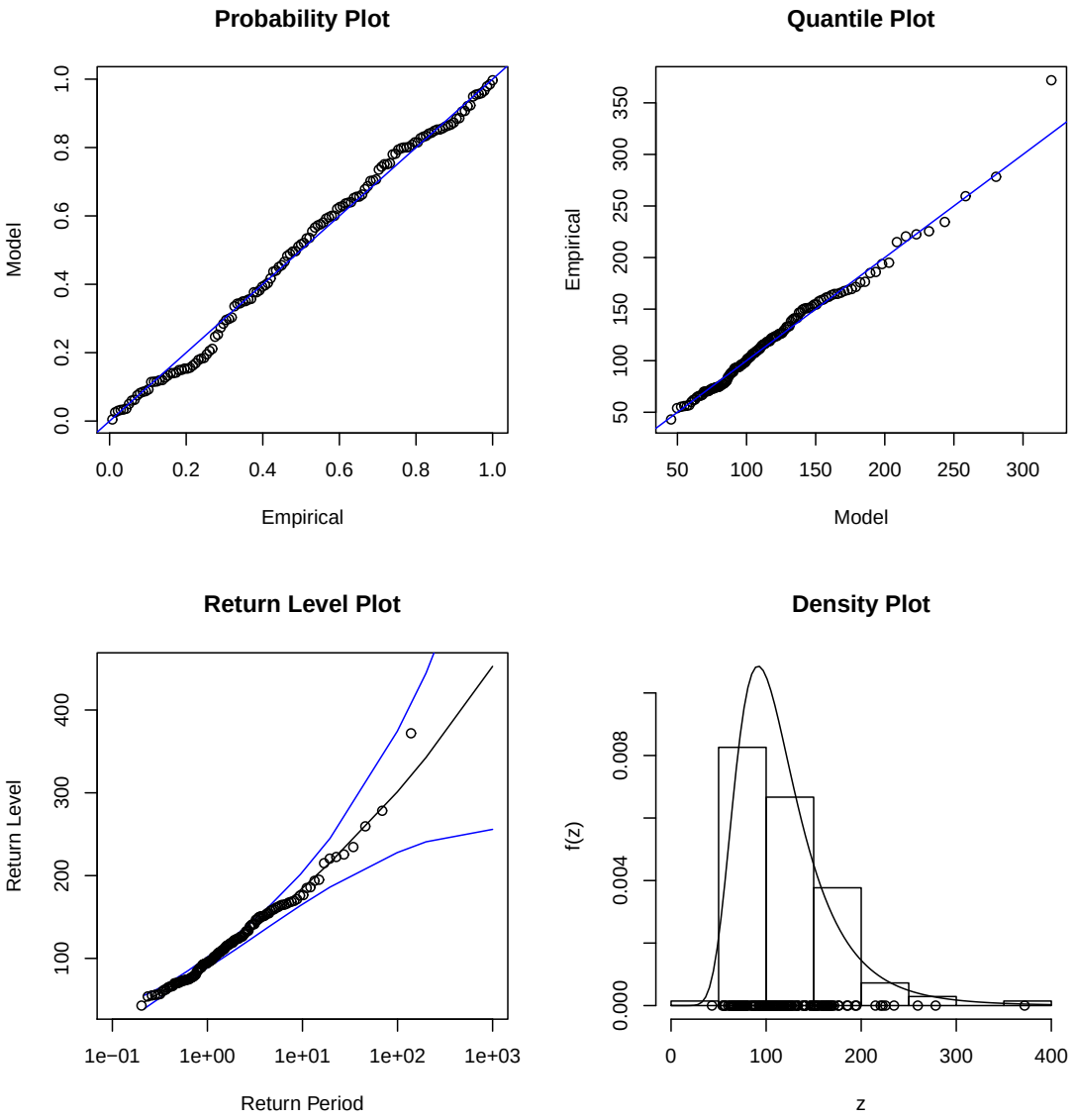
$$\left\{ \left(\frac{i}{n+1},\ \widehat{G}(z_{(i)}) \right) :\ \ i = 1, 2, \ldots, n \right\}, \quad \widehat{G}(z_{(i)}) = \exp\left\{ - \left[1 + \widehat{\xi} \left(\frac{z_{(i)} - \widehat{\mu}}{\widehat{\sigma}} \right) \right]^{-1/\widehat{\xi}} \right\}.$$

確率点プロット (Quantile Plot)

$$\left\{ \left(\widehat{G}^{-1} \left(\frac{i}{n+1} \right),\ z_{(i)} \right) :\ \ i = 1, 2, \ldots, n \right\},$$

$$\widehat{G}^{-1} \left(\frac{i}{n+1} \right) = \widehat{\mu} + \widehat{\sigma} \left[\left\{ - \log \left(\frac{i}{n+1} \right) \right\}^{-\widehat{\xi}} - 1 \right] / \widehat{\xi}.$$

経験分布関数 (Empirical) と推定分布関数 (Model) の点 $z_{(i)}$ でのズレを見る.


Probability Plot
Model
Empirical
0.0
0.2
0.4
0.6
0.8
1.0
Quantile Plot
Empirical
Model
50
100
150
200
250
300
350
Return Level Plot
Return Level
Return Period
100
200
300
400
1e−01
1e+00
1e+01
1e+02
1e+03
Density Plot
f(z)
z
0.000
0.004
0.008
0
100
200
300
400

再現レベルプロット (Return Level Plot)：プロット

$$\left\{\left(-1\Big/\log\left(\frac{i}{n+1}\right),\ \ z_{(i)}\right):\ \ i=1,2,\ldots,n\right\}$$

に，一般極値分布の T 再現レベルの推定値

$$\left\{\left(-1/\log(1-1/T),\ \ \widehat{\mu}+\widehat{\sigma}\left[\left\{-1/\log(1-1/T)\right\}^{\widehat{\xi}}-1\right]/\widehat{\xi}\right):\quad 0.1<T<1000\right\}$$

と，その 95% 信頼区間を描き加えたものである．

この図では x 対数軸にするので極値確率紙プロットに相当する．

ブロック・サイズの決め方

ブロック最大データに一般極値分布を適合できるのは，極値確率紙で，プロットが直線に近い，上に凸，そして下に凸の形状の場合である．

このとき，それぞれの形状のブロック最大データの適合候補分布は，Gumbel，Weibull そして Fréchet分布となる．

極値確率紙でプロットが上記以外の複雑な形状になる場合は，ブロック・サイズを増やす等の処置が必要になる．

極値（グンベル）確率紙へのプロット：

$$\left\{\left(-\log\left[-\log\left(\frac{i}{n+1}\right)\right],\ z_{(i)}\right):\ i=1,2,\ldots,n\right\}$$

モデル選択

定常モデルを入れて12個のモデルの中で，最適なものをAICで選択.

AIC最小のモデルは M_{010} で

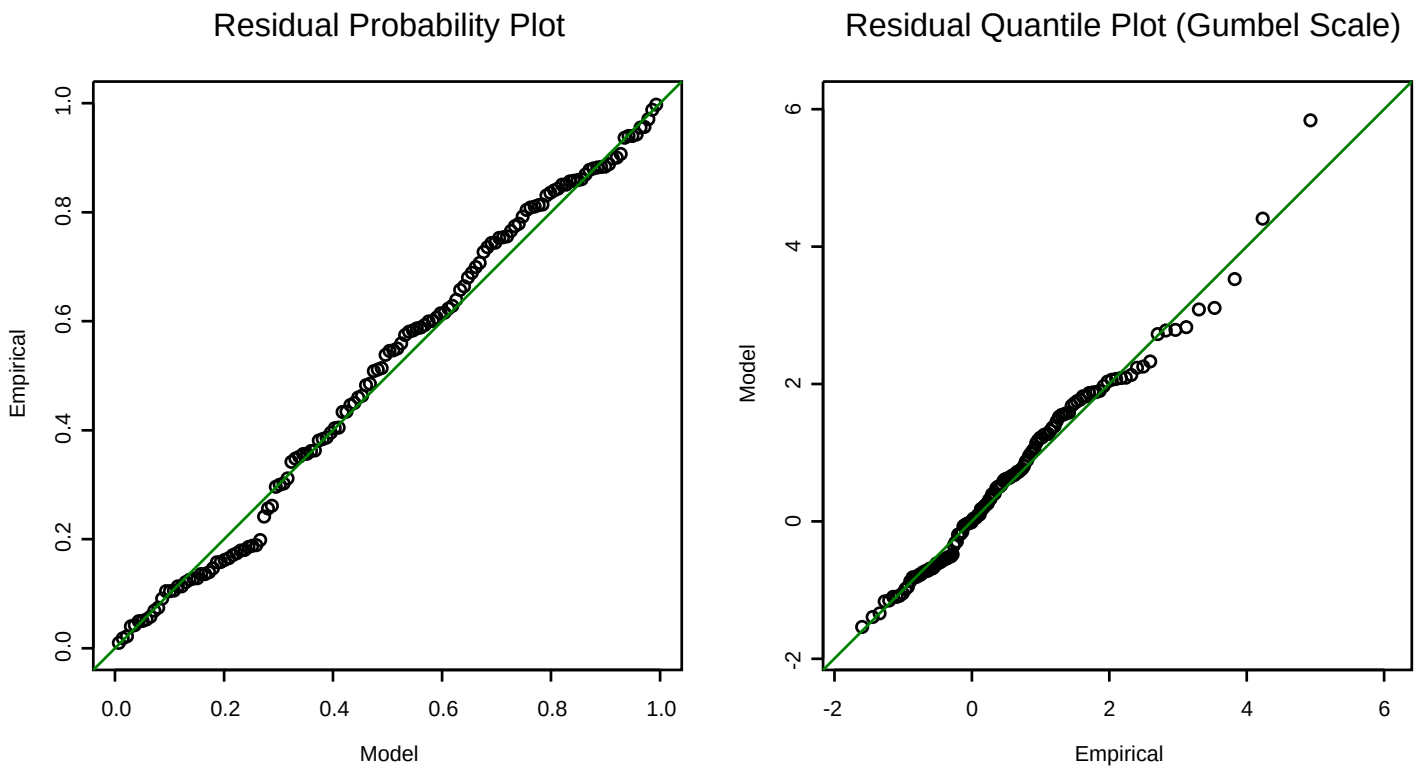
$$\widehat{\mu}(y) = 94.658, \quad \widehat{\sigma}(y) = \exp(3.527 + 0.192y^*),$$

$$\widehat{\xi}(y) = 0.097, \quad y^* = (y - 1945)/69, \quad y = 1876, \ldots, 2013$$

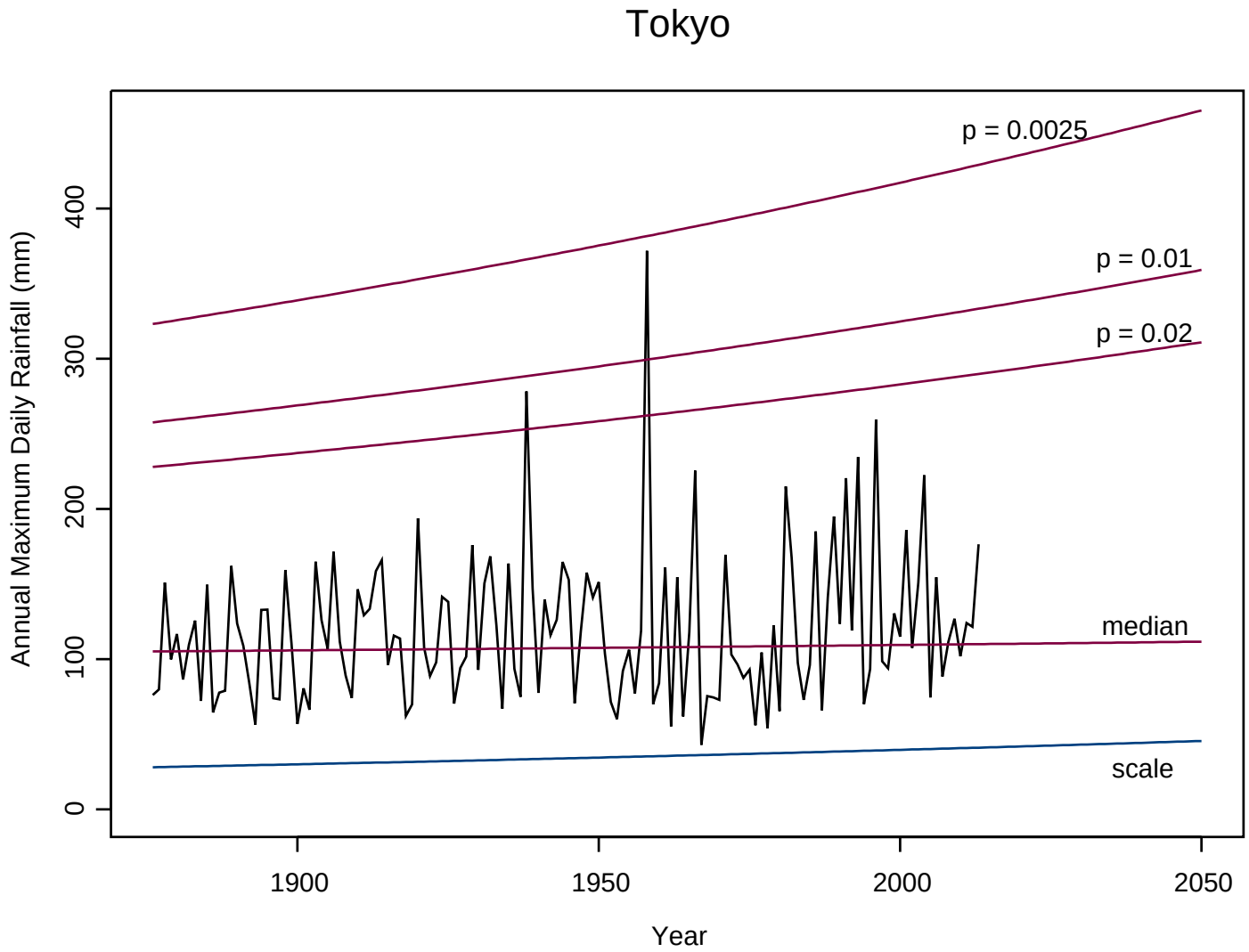
となった.

年最大データの従う分布として，位置と形状パラメータは一定であるが，尺度パラメータが年とともに増加するFréchet分布が選ばれた.

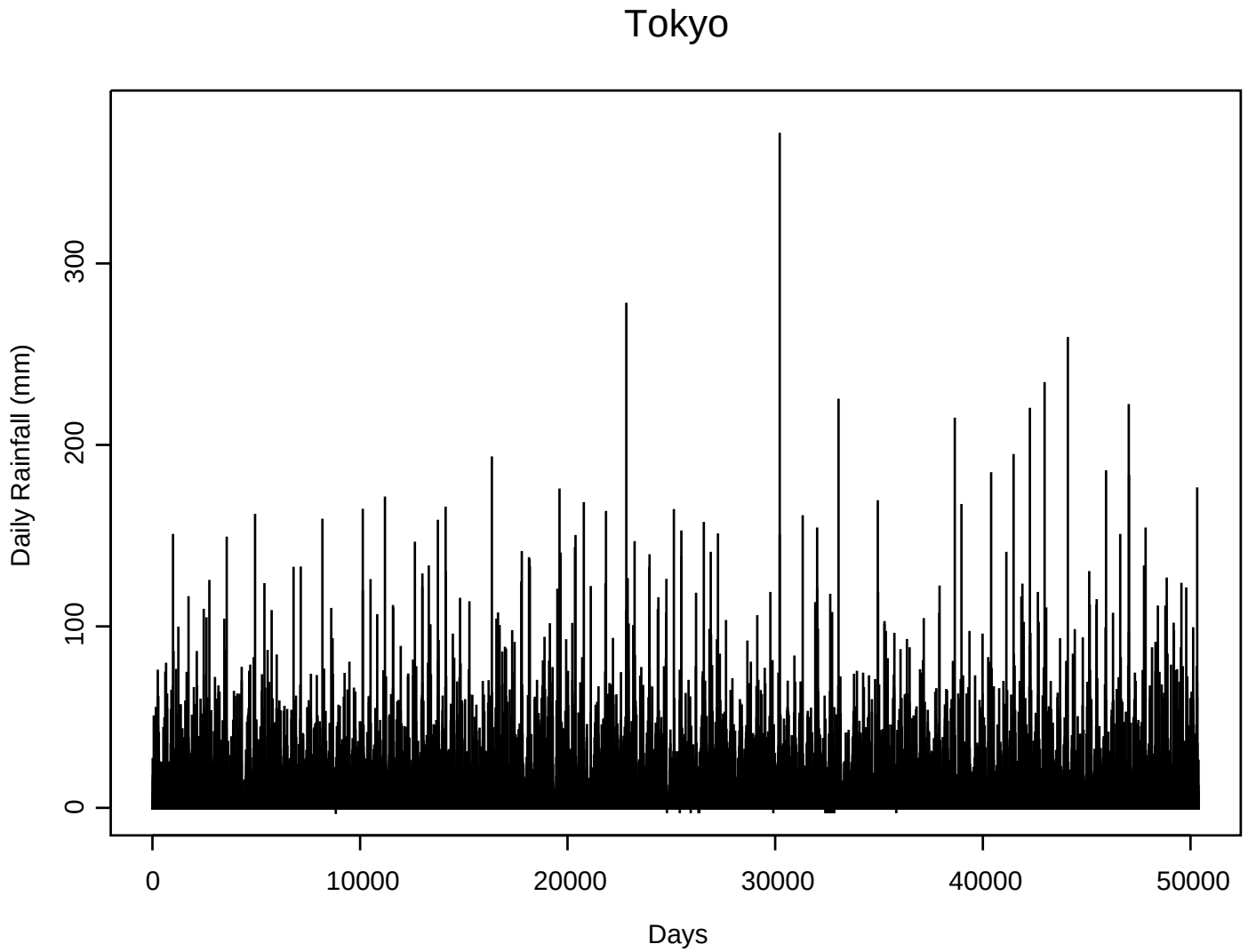
形状パラメータが正で尺度パラメータが増加すると今後，今までに経験したことの無いような大雨が降る可能性がある.



モデル M_{010} での解析診断.



東京の年最大日降水量，各年の上側 p 確率.

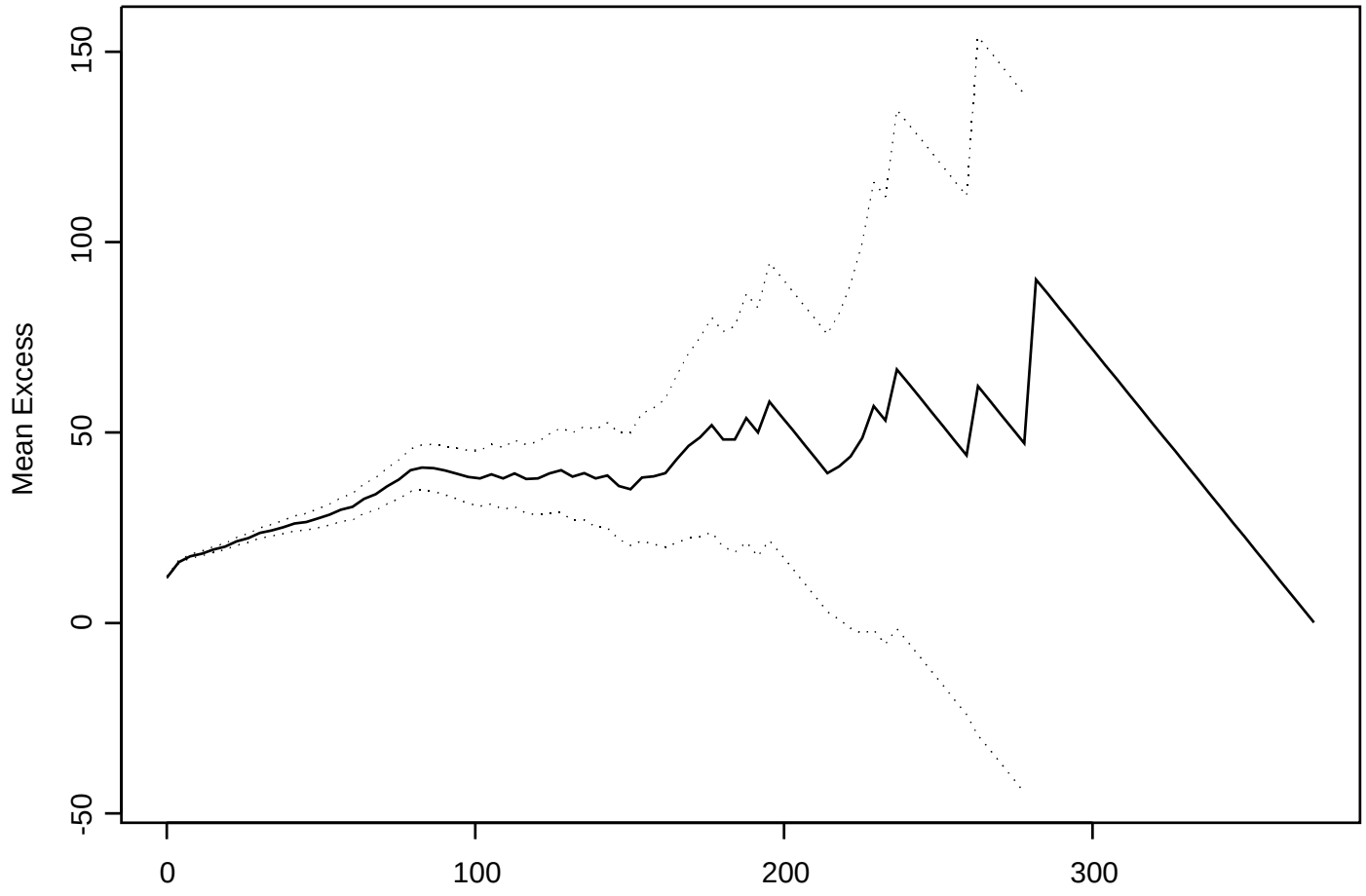


東京の日降水量(mm)，1876年1月1日～2013年12月31日.

一般パレート (GP) モデルによるデータ解析

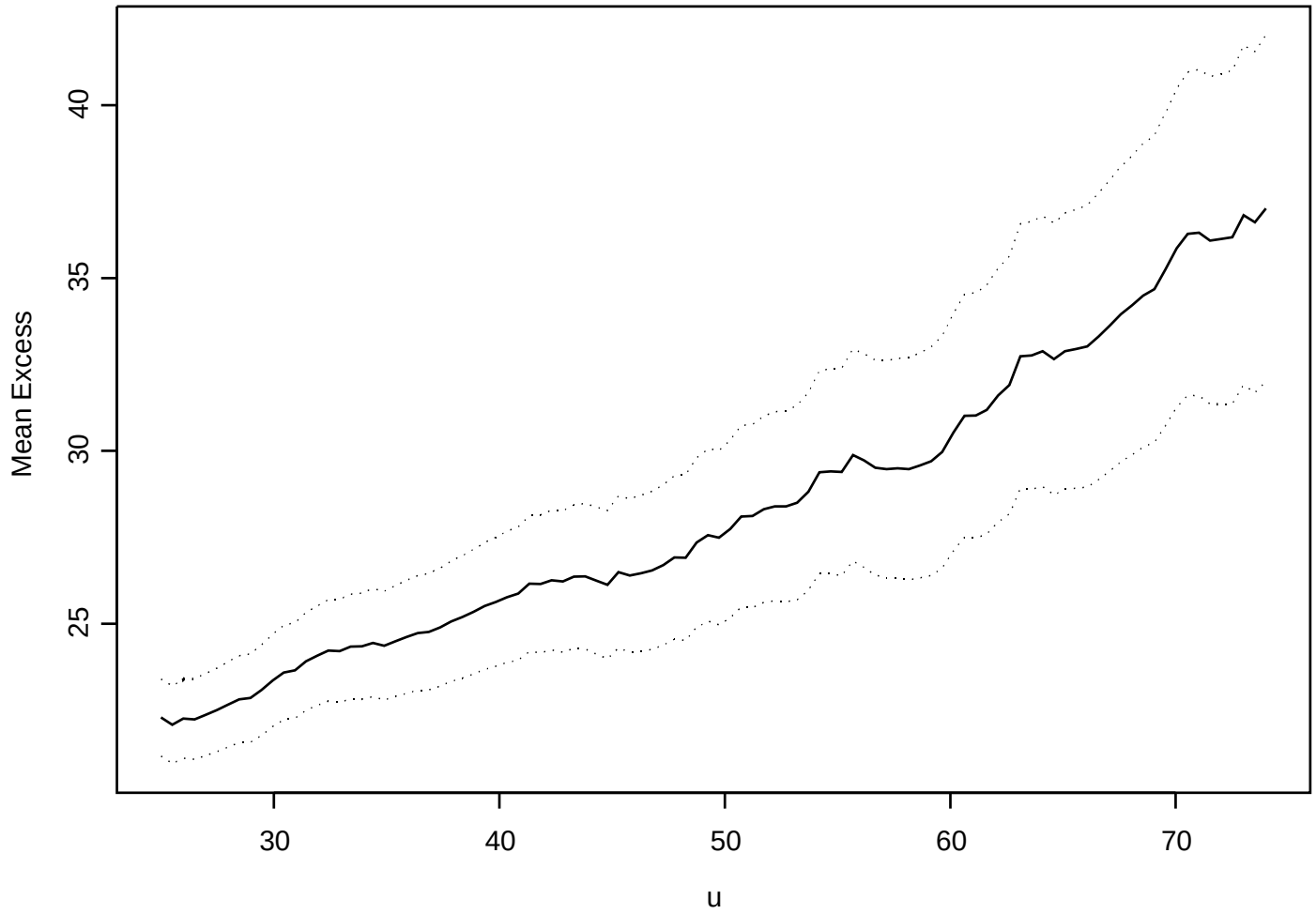
東京の日降水量データを GP モデルで解析する.

まず，データは定常として閾値を決める.

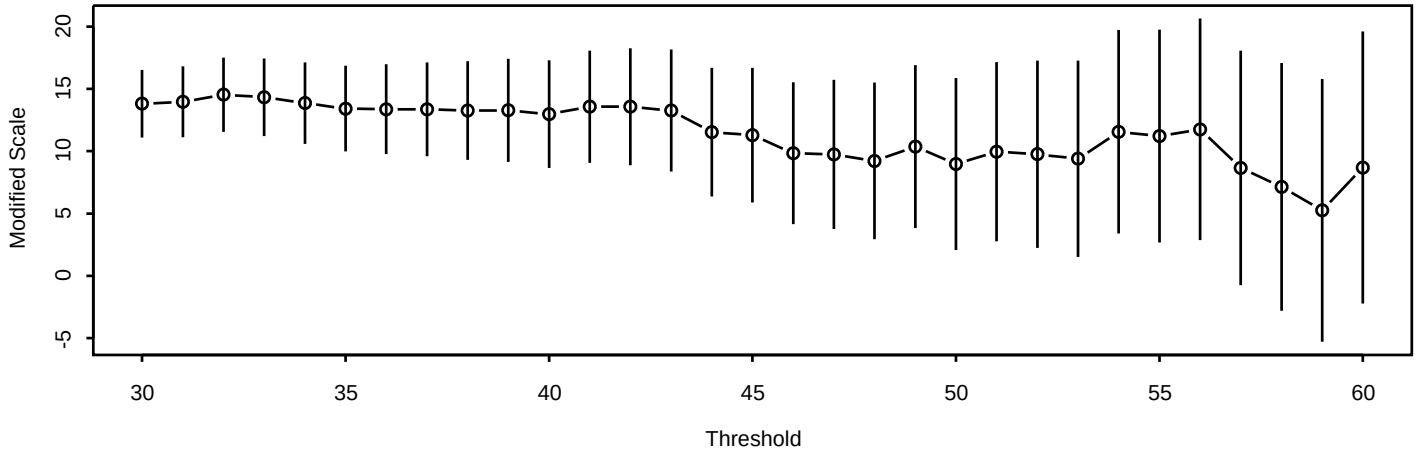


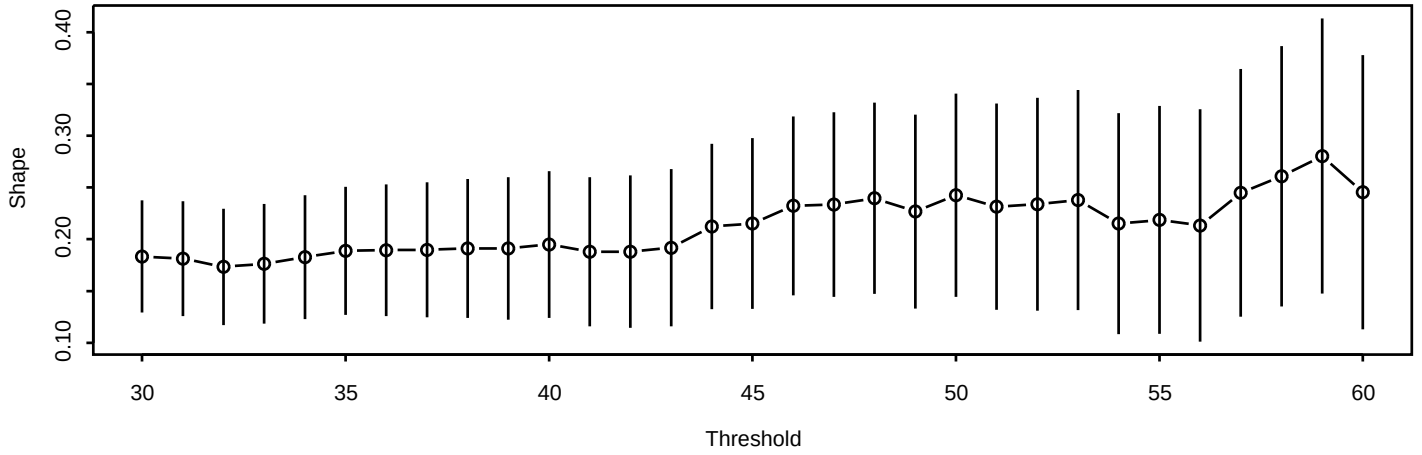
標本平均超過プロット.

データ数，200より大8個，100より大122個，50より大703個.



標本平均超過プロット.





修正尺度と形状パラメータの推定値プロット.

解析結果

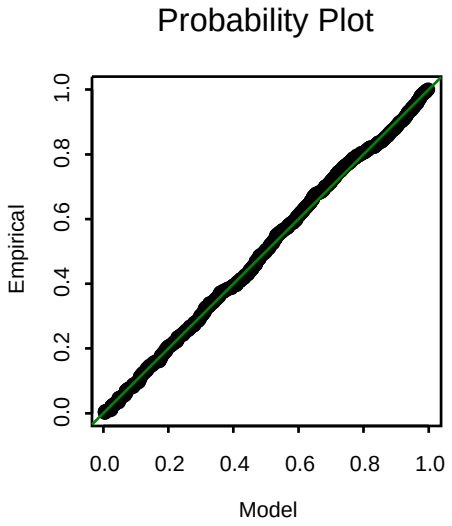
図から閾値 $u = 46$ を選択．この閾値を用いて，超過するデータに一般パレート分布 GP(σ, ξ) を適合して解析．

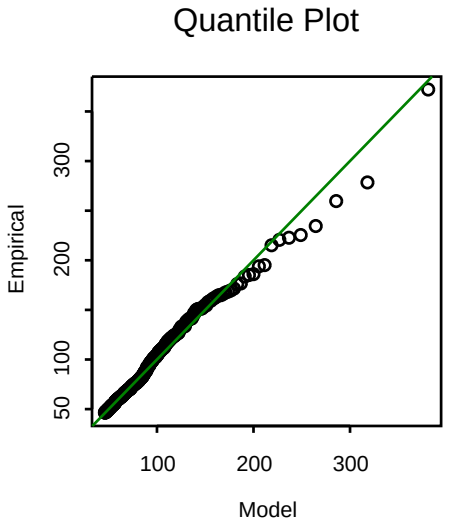
最大対数尤度は -3603.12，最尤推定値（標準誤差）は

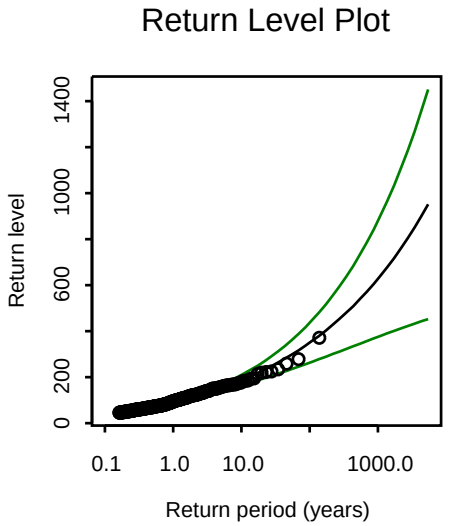
$$\widehat{\sigma} = 20.52\,(1.14), \quad \widehat{\xi} = 0.232\,(0.044).$$

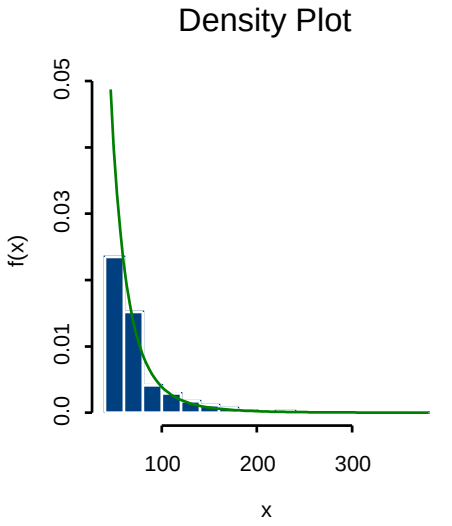
形状パラメータ ξ の最尤推定値 0.232 は正で，GEV の場合と比べてかなり大きい．

十分大きいデータの分布は Pareto 分布と推定され，非常に大きな値が観測される可能性がある．


Probability Plot
Empirical
Model
0.0
0.2
0.4
0.6
0.8
1.0


Quantile Plot
Empirical
Model
50
100
200
300


Return Level Plot
Return level
Return period (years)
0
200
600
1000
1400
0.1
1.0
10.0
1000.0


Density Plot
f(x)
x
0.0
0.01
0.03
0.05
100
200
300

GP 解析の診断.

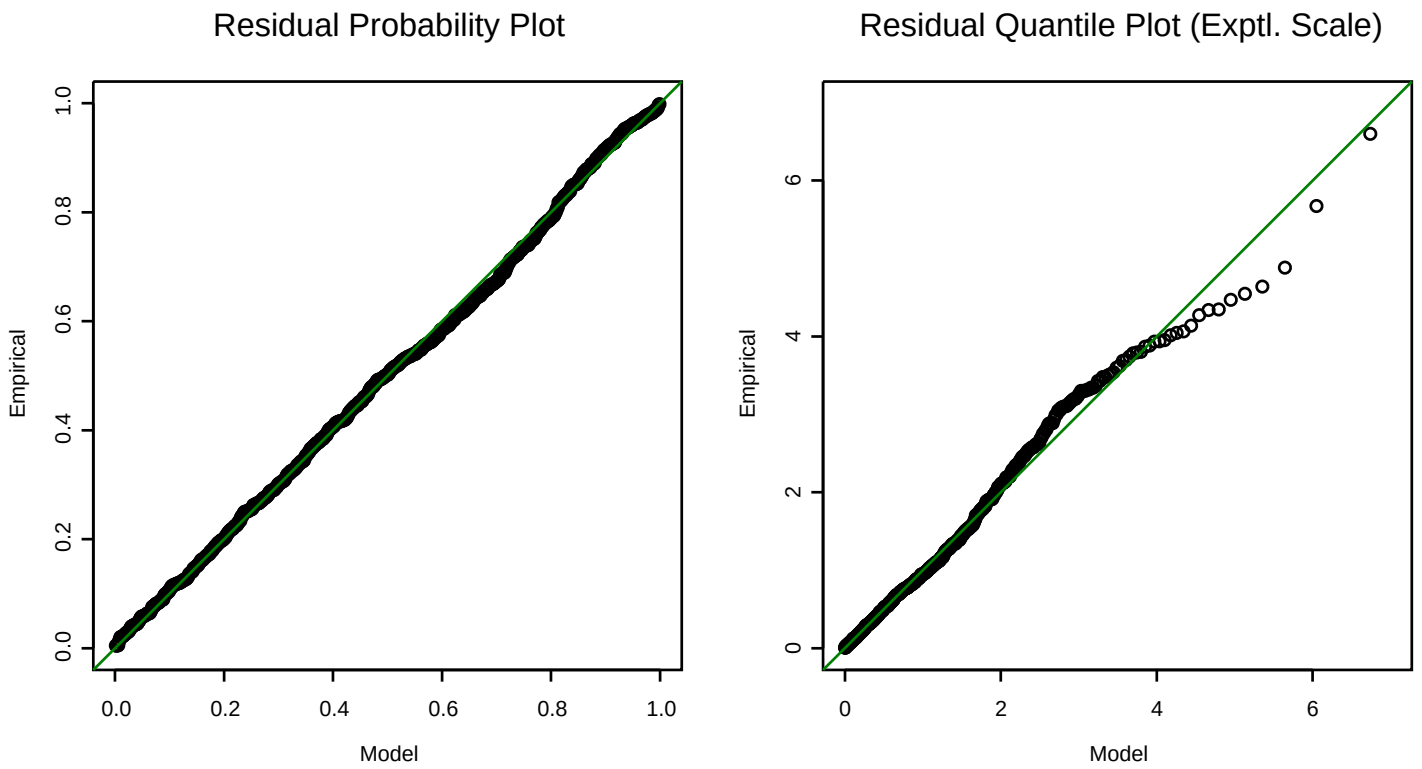
モデル選択

GEVモデルでは非定常なモデルがAICにより選択されている.

パラメータ $(\sigma,\, \xi)$ が時間に依存するモデルを適合.
その結果AICで選択されたのは，形状パラメータは一定で尺度が変化する次のモデル M_{10} である：最大対数尤度は -3598.958 で,

$$\widehat{\sigma}(t) = \exp(2.818 + 0.410t), \quad \widehat{\xi}(t) = 0.226, \quad 0 \le t \le 1.$$

簡単のために138年間を区間 $[0,\, 1]$ に変換している.

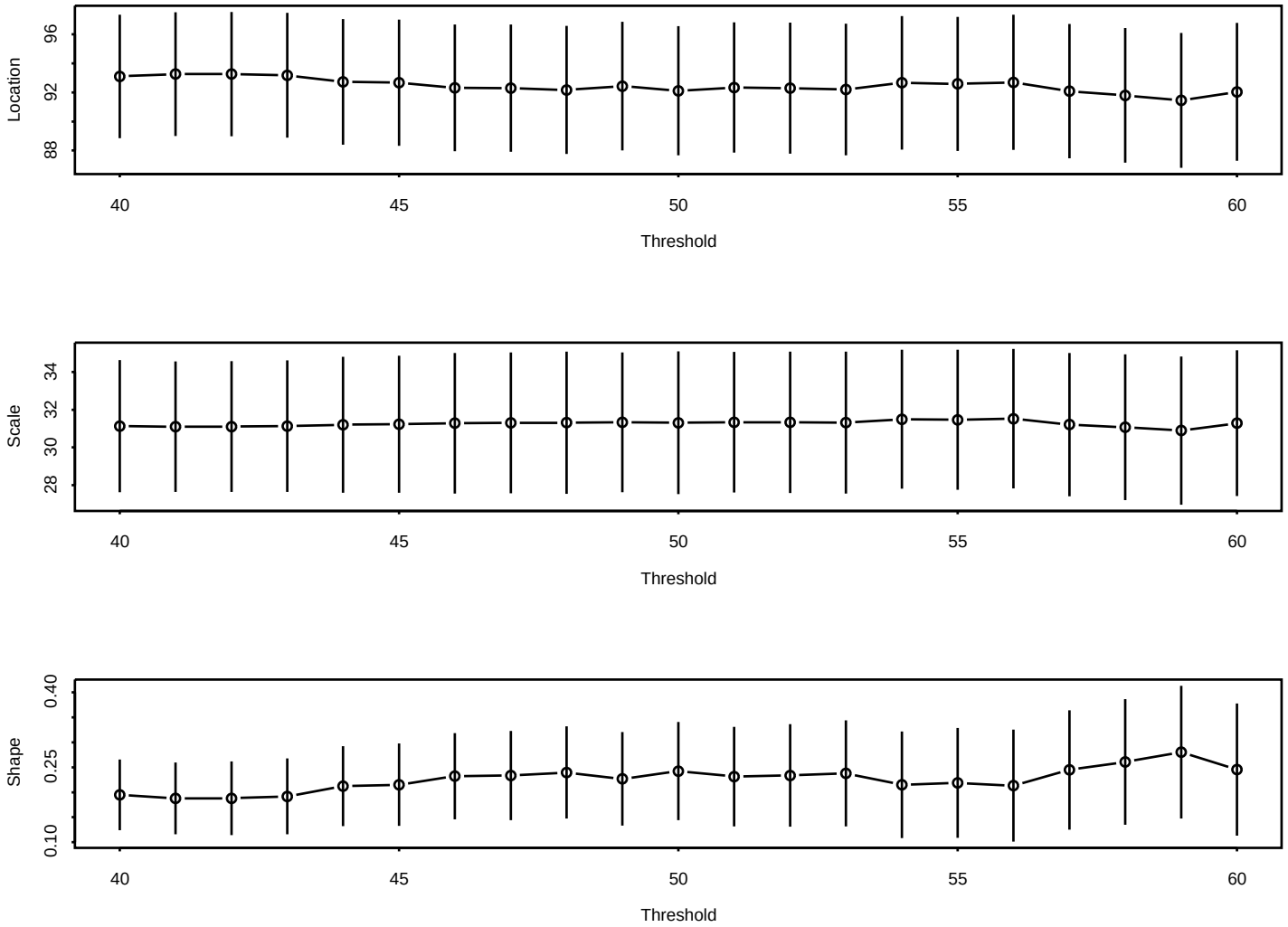


GP 解析モデル M_{10} の診断.

点過程 (PP) モデルによる解析

東京の日降水量データを点過程 (PP) モデルで解析.

データは定常で PP(μ, σ, ξ) に従うとする.



位置，尺度，形状パラメータの推定値プロット.

PP モデルによる解析結果

GP モデルと同じ閾値 $u = 46$ を選択．この値より右では μ，σ，ξ の推定値は一定と見なす．この閾値を用いて PP(μ, σ, ξ) モデルを適用して解析．

最大対数尤度は -2913.872 ，最尤推定値（標準誤差）

$$\widehat{\mu} = 92.30\,(2.23), \quad \widehat{\sigma} = 31.28\,(1.90), \quad \widehat{\xi} = 0.232\,(0.044).$$

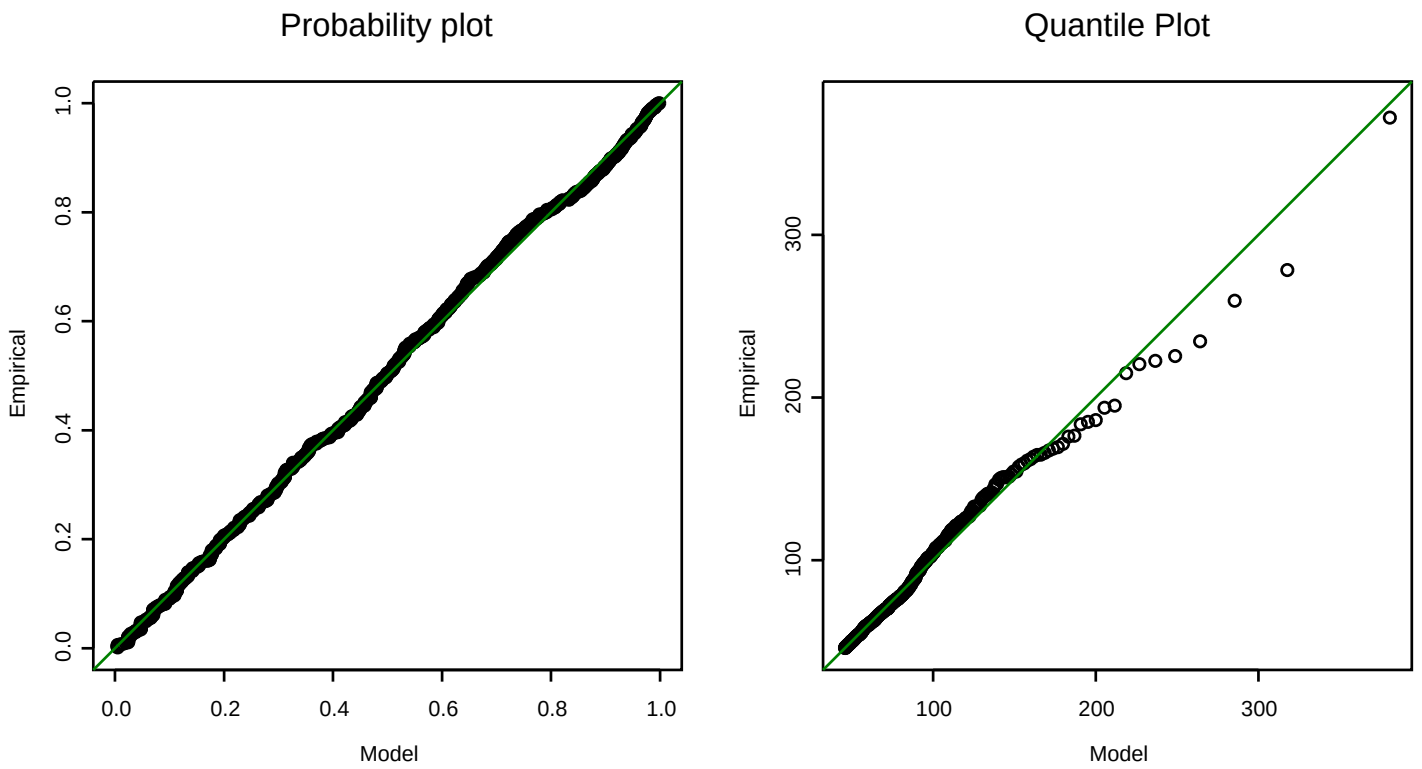
形状パラメータ ξ の推定値は GP モデルでの推定値と等しい．

年最大値のみを用いた場合の最尤推定値（標準誤差）

$$\widehat{\mu} = 95.17\,(3.35), \quad \widehat{\sigma} = 34.08\,(2.58), \quad \widehat{\xi} = 0.114\,(0.075).$$

位置と尺度パラメータはほぼ等しいが，形状パラメータはかなり違う．

データ数に応じて PP モデルでは標準誤差はかなり小さくなっている．


Probability plot
Empirical
Model
0.0
0.2
0.4
0.6
0.8
1.0
Quantile Plot
Empirical
Model
100
200
300

PP解析の診断.

モデル選択

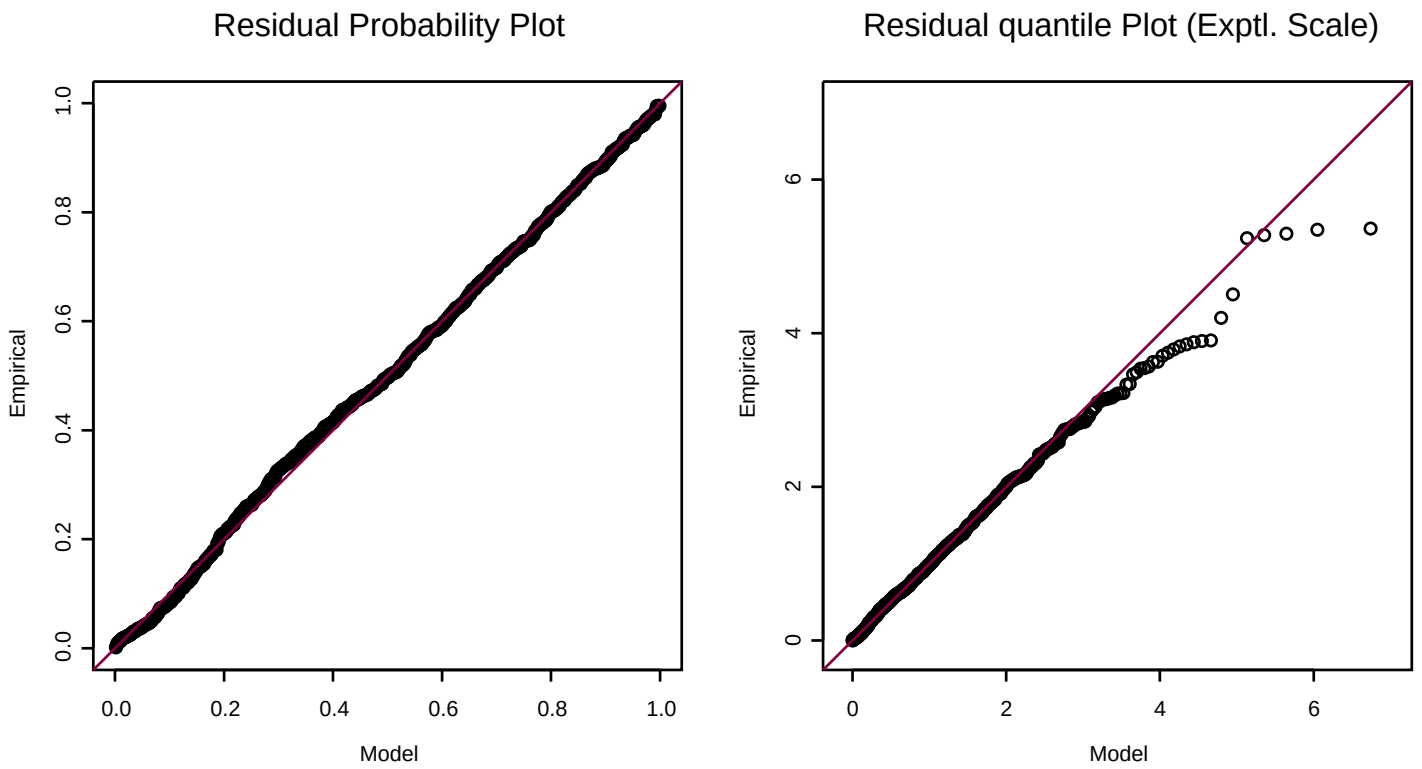
GEV モデルや GP モデルでの解析結果では非定常なモデルが選ばれた.
ここでも 12 個のモデルの比較を行う.

AIC で選ばれたのはモデル M_{110} で，最大対数尤度は -2909.82

$$\widehat{\mu}(t) = 83.42 + 17.85\,t, \quad \widehat{\sigma}(t) = \exp(3.24 + 0.39\,t),$$

$$\widehat{\xi}(t) = 0.226, \qquad 0 \le t \le 1.$$

簡単のために 138 年間を区間 $[0,\ 1]$ に変換している.



PP 解析モデル M_{110} の診断図.

6　おわりに

極値統計学の基本仮定

次の３つの仮定は同値.

ブロック最大データに一般極値分布が適合できる.

閾値超過データに一般パレート分布が適合できる.

閾値を超えるデータに点過程モデルが適合できる.

推定は最尤法で行う．最尤法は，推定値の標準誤差が簡単に求まり，
非定常の場合も扱うことが出来る柔軟な推定法である.

極値データ解析結果の保証のために診断は重要である.

高橋のレジメに校正ミスがあります．修正をお願いします．

p. 02–10　下から10行目　　$\sigma(t) \ \rightarrow \ \log \sigma(t)$

p. 02–14　下から8行目　　m年間　$\rightarrow$　n_y年間

下から5行目　　1年間の観測数を　$\rightarrow$　観測年数を

下から1行目　　$x_1, \ldots, x_n \quad \rightarrow \quad x_1, \ldots, x_{N(A)}$

p. 02–15　上から2行目　　$\lambda(t,\, x) = [1 + \xi(x - \mu)/\sigma]^{-1/\xi - 1}/\sigma$