水工学オンライン連続講演会
2023.06.26

極値統計を実務に応用する際の課題

～ Gumbelからの６５年，水文統計をとりまく状況の変化を数理的側面から概観

北野 利一

名古屋工業大学 社会工学専攻／統計数理研究所（客員）

本日の内容
（キーワード）

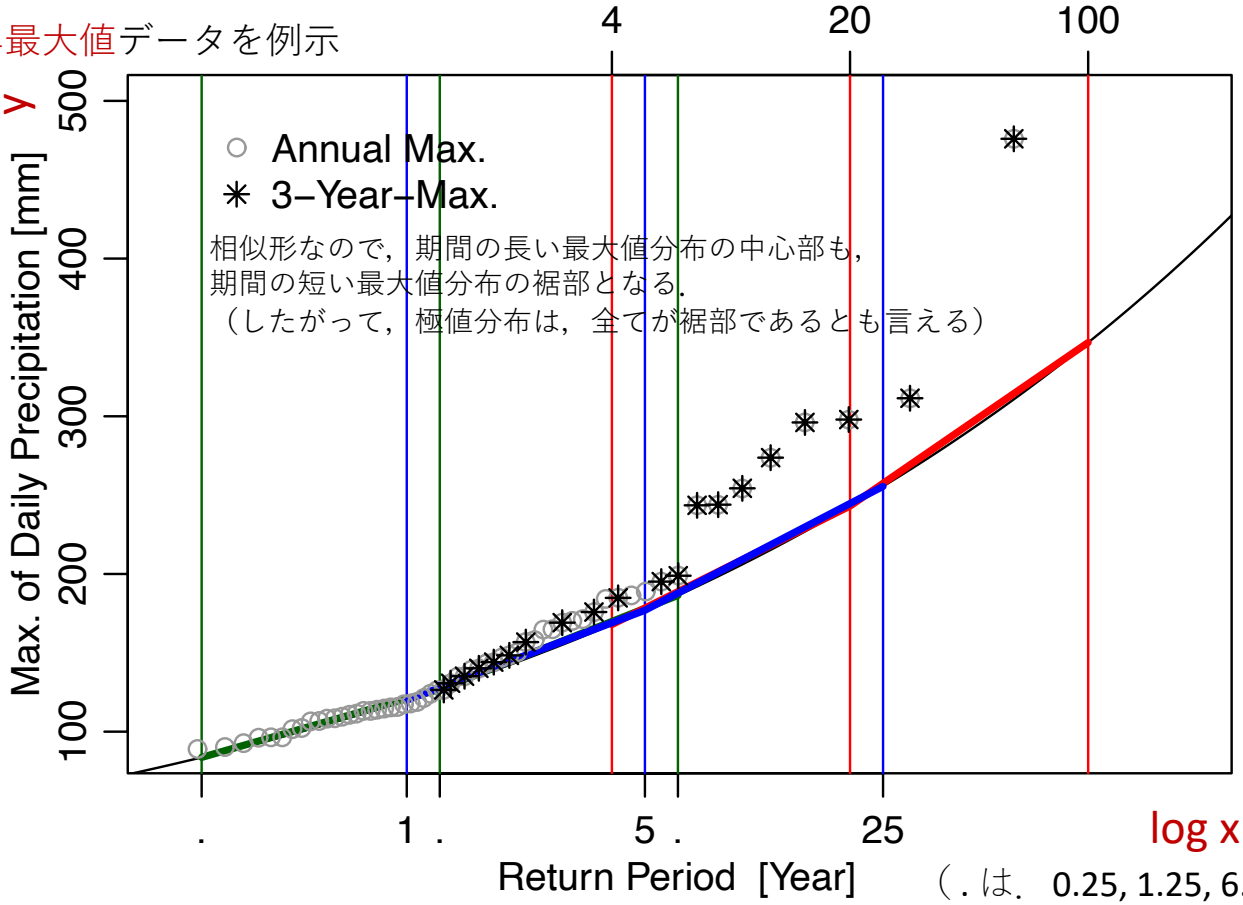
＊ （まずは確率の話）そもそも極値分布とは？
＊ 再現期間 ～ 再現レベル（＝確率外力）との関係
＊ （次に統計の話）統計といえば検定，例えば，
統計的検定の前提を（どの程度）疑う必要があるのか？
＊ 節約原理（オッカムのカミソリ）と G. Box の格言
「すべてのモデルは間違いである．しかし，．．．」
＊ 適合度とモデル選択
「候補分布にあてはめて，そのなかから選ぶ」という発想
からの脱却 → 「プロクルステス寝台」問題（略：寝台問題）
＊ 推定に伴う統計的誤差と本質的な確率変動，そして，
アンサンブル標本を扱う時代の不確実性の捉え方

Genesis： **極値分布の本質** ～ 自己相似性

self-similarity
max-stable

$$\frac{\mathrm{d}}{\mathrm{d}\log x}\log\left(\frac{\mathrm{d}\,y}{\mathrm{d}\log x}\right)=\xi$$

６０年間の年最大値データを例示



（. は．0.25, 1.25, 6.25 年に対応します）

Michelangelo

神（は，
自らを模して，
人間を作った）

アダム
（人間）

Michelangelo

確率分布の本質は，
微分方程式（Systems of frequency curves）
にあり．ガウスによる正規分布の導出も
微分方程式から．

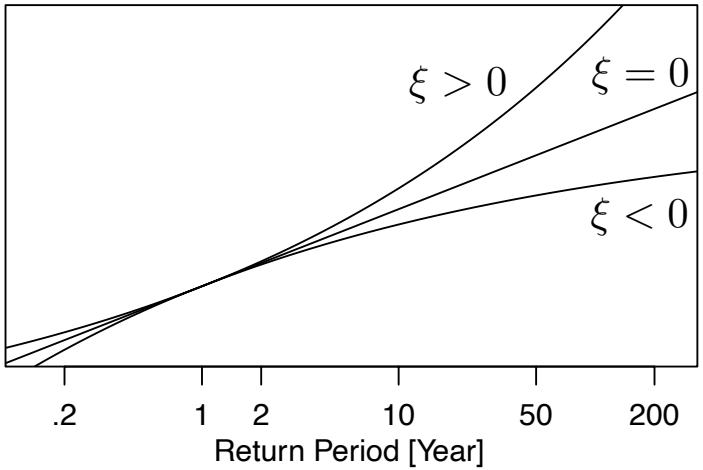
$$-\frac{\mathrm{d}}{\mathrm{d}\log\lambda_n}\log\left(-\frac{\mathrm{d}\,y}{\mathrm{d}\log\lambda_n}\right)=\xi$$

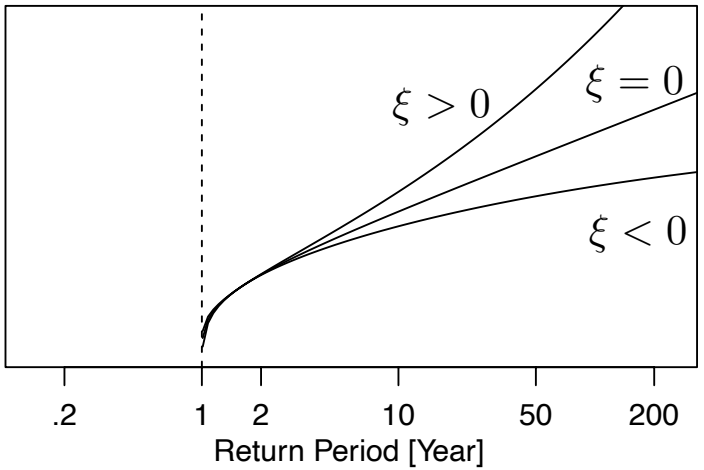
→ 生起率を用いて表すと，ξ がブロックサイズ n に依存せず，
得られる曲線 $y \sim \log\lambda_n$ がブロックサイズ n に対して，
自己相似性があることが表されている．

北野利一：伊勢湾台風級の高潮と確率潮位，2013 年度（第49回）水工学に関する夏期講習会，B コース．B-3-1～B-3-20.
北野利一・高橋倫也・田中茂信: 気候モデルから得られる多数のアンサンブルデータを用いた確率降水量の推定法，土木学会論文集B1（水工学），73, pp. I_1-I_6, 2017.
北野利一 : 成分毎の最大値と閾値を超過する多変量極値：それらの相互関係，それらの単純極値分布と乱数生成法，日本統計学会誌，vol. 51, pp.123-156, 2021.（付録に解説）

＊ 再現期間　$x = 1/\lambda_1(y)$

Logarithmic trend of the increase of the extremes
Gumbel, E. 1958





$$\lambda_1(\mu_n) = 1/n \;\Leftrightarrow\; \lambda_n(\mu_n) = 1 \qquad 1 - P_1(y_n) = 1/n$$

＊したがって，位置母数　μ_n　は，再現期間　n　年の再現レベルに対応する．

＊ 期間最大値分布（いわゆるGEV分布）と生起率関数（添字 n は期間長を表す）

$$P_n(y) = \exp\{-\lambda_n(y)\}$$

$$\lambda_n(y) = \left(1 + \xi\frac{y-\mu_n}{\sigma_n}\right)^{-1/\xi} \quad (\xi \neq 0)$$

$$= \exp\left(-\frac{y-\mu_n}{\sigma_n}\right) \quad (\xi = 0)$$

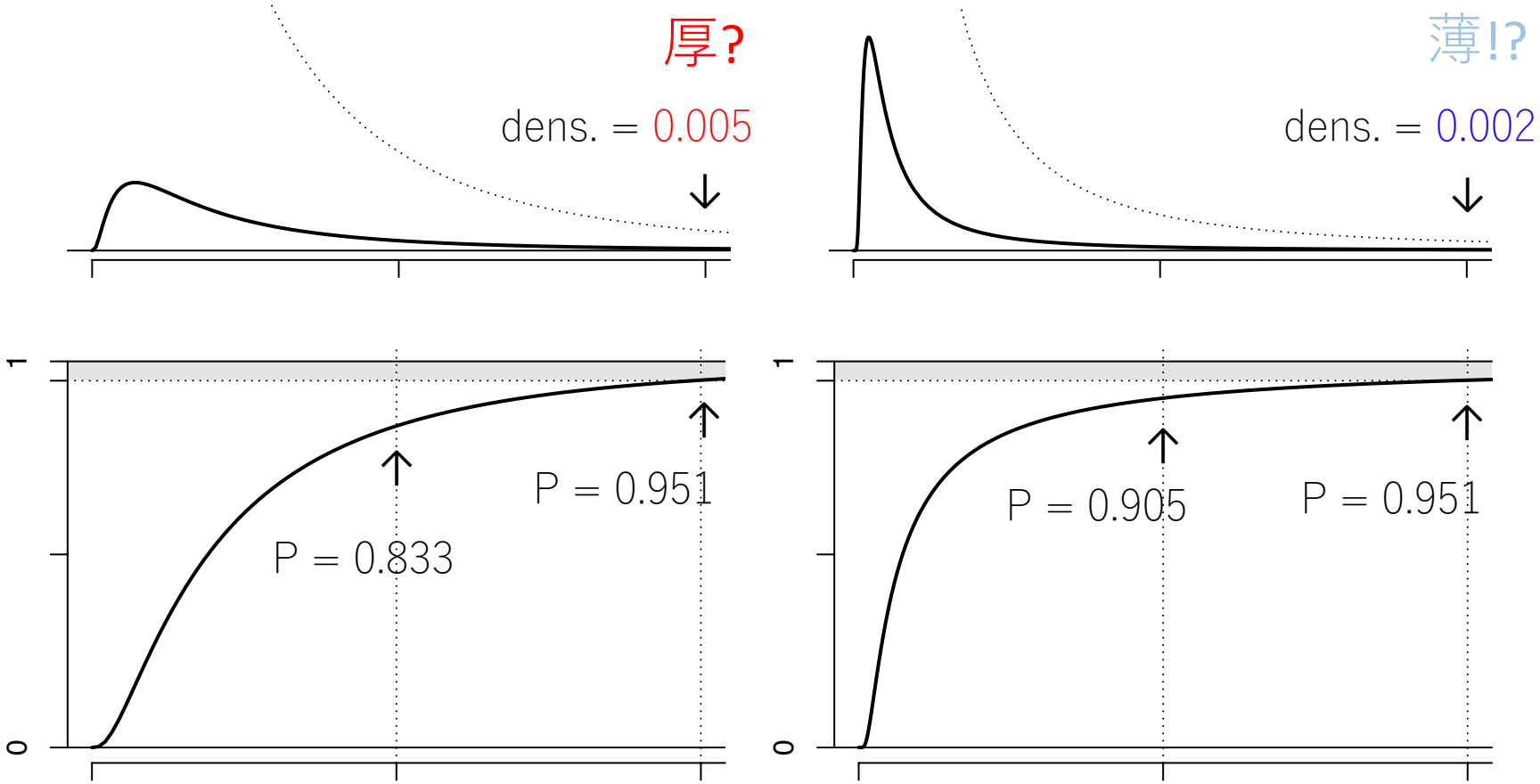
Trinity Theorem　三位一体定理（父・子・精霊；この３つの本性は同一）

Ⅱ．（標準）Fréchet 分布（期間最大値）：　$F(x) = \exp\left(-\dfrac{1}{x}\right)$

Ⅰ．（標準）Gumbel 分布（期間最大値）：　$G(z) = \exp\left(-e^{-z}\right)$

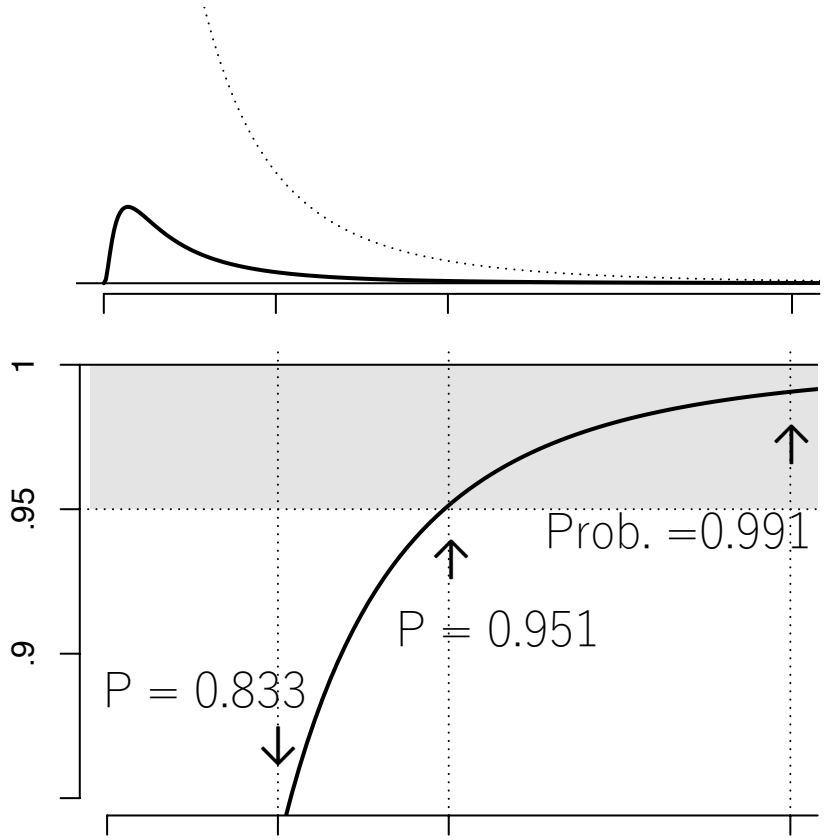
Ⅲ．（標準）Weibull 分布（期間最小値）：　$W(\lambda) = 1 - \exp(-\lambda)$

いわゆるロングテールの本質とは？



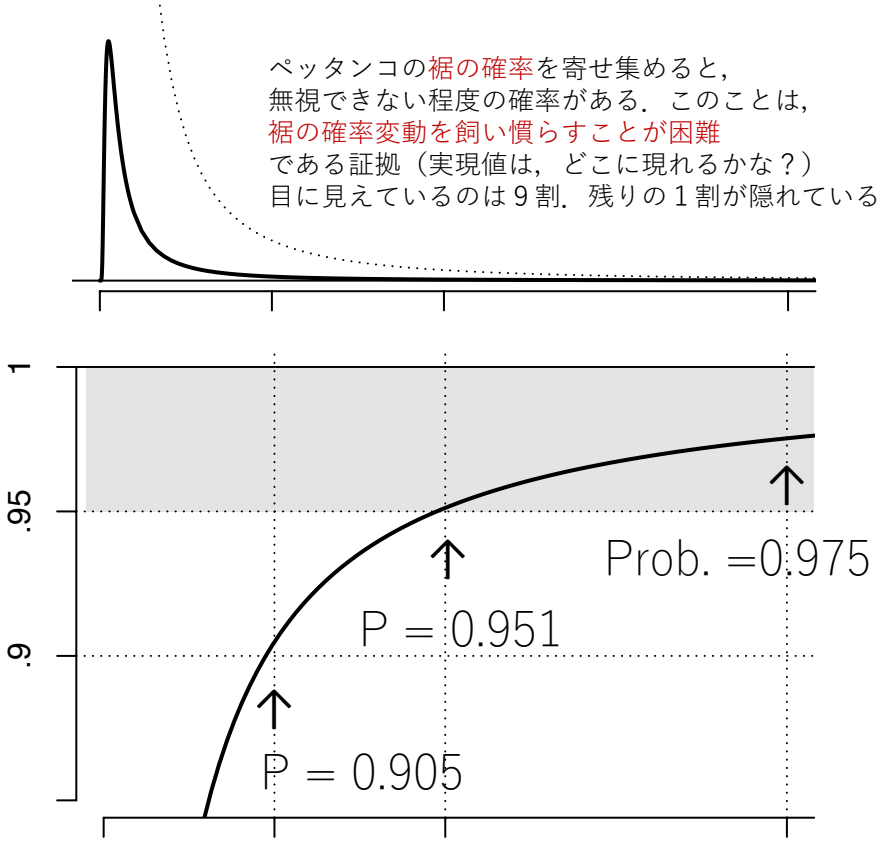
問：左右どちらの分布が，ロングテールですか？

いわゆるロングテールの本質とは？



残り５％から一気に，残り１％に減少
（もう残りがない）

ペッタンコの裾の確率を寄せ集めると，
無視できない程度の確率がある．このことは，
裾の確率変動を飼い慣らすことが困難
である証拠（実現値は，どこに現れるかな？）
目に見えているのは９割．残りの１割が隠れている．



残り５％のうちの半分に減少
（まだ残りがある）

無視できない裾の確率（ここでは 5%）の確率密度が見えない → 相応する実現値は神出鬼没で出現

さらに言えば，極値理論の自己相似性から，極値理論の適用範囲では，どこでも裾になる．
年最大値が極値分布に従うなら，その最頻値も（より小さいブロックサイズの期間最大値分布の）裾なのである！

ロングテールの確率変数は，「飼いならす」ことが本質的に困難となる！

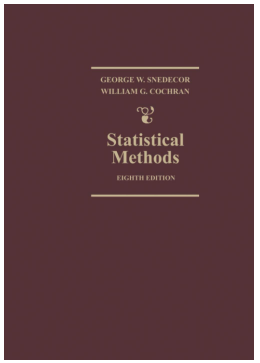
補足： The taming of chance, by I. Hacking, 1990 （正規分布に基づく統計学が様々な分野に応用された成功事例物語）

＊統計的検定の前提を（どの程度）疑う必要があるのか？

母分布を正規分布とした統計的問題の典型例：平均値に関するティ検定

検定統計量　$t = \dfrac{\bar{X} - \mu}{\sqrt{s_n^2/n}}$　　< qt(0.95, df)

（有意水準５％で片側検定の場合）



標準化には，標本平均と標本分散が含まれる．これらの確率的変動が従う分布を想定しないと，有意性を確率で評価できない．

Snedecor & Cochran, Statistical Methods, 1st. Ed. 1937, 8th Ed. 1989
私の本棚のものは，7th Ed. 1980

→ Snedecor & Cochran によれば，正規分布を用いる４つの理由

1．... some are approximately normal, ...　（そもそも正規分布で近似できる？＝理由になっていない？！）

2．a simple transformation ... may induce approx. normality.　（これも，理由になっていない？！）

3．The normal distribution is relatively easy to work with mathematically.　...

Such results may hold well enough rough-and-ready to use ...　（**間に合わせだけど役に立つ**，まさにコレ！）

4．sample mean tends to become normal ... as the size of sample increases.　（中心極限定理）

→ さらに言えば，天文学者ガウスは，標本平均が母平均の推定となる確率分布として微分方程式をつくり，その微分方程式を解いて正規分布を導いていることが理由と考える．

統計的な判断には謙虚さも必要である．

**信じるに足る根拠を探す努力よりも，
疑うに足る事実根拠を用意する努力をする．そして，用意できないのなら，深追いせずに，（数理的な）論理を大事にするのが筋か？**

＊ 節約原理（オッカムのカミソリ）

＊ All models are wrong, but some are useful. (G. Box)

＊ 厳密な真理の探究よりも，むしろ筋を通すこと（論理の一貫性）の重視

→ 案外，それが真理への一番の近道かもしれないよ

von Mises の定理

$$\lim_{y \to y_\infty} \frac{\mathrm{d}}{\mathrm{d}y}\left\{\frac{1-F(y)}{f(y)}\right\} = \xi \quad \rightleftarrows \quad \frac{\mathrm{d}}{\mathrm{d}\log x}\log\left(\frac{\mathrm{d}\,y}{\mathrm{d}\log x}\right) = \xi$$

ここでの F と f は，任意の分布関数の累積確率と確率密度を表す（標準フレシェではないことに注意）．

Domain of attraction （吸引領域）

正規分布，対数正規分布，ワイブル分布 → 極値 I 型（Gumbel 分布）

平方根指数分布（江藤分布） → 極値 I 型（Gumbel 分布）

対数ガンマ分布 → 極値 II 型（Fréchet 分布） （なるほど！ 米国の水文統計）

「極値理論の工学への応用」のためのマイルストーン

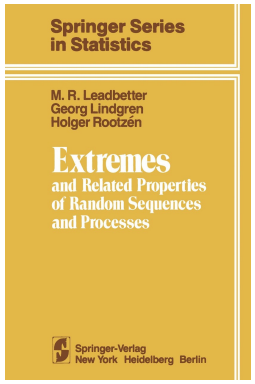
Gumbel, E. J. (1958): Statistics of Extremes の後に，デルタ計画の報告書が出版（1960年）が，しかし，…その後，

de Haan, L. (1990): Fighting the arch-enemy with mathematics で，治水計画に本格的な（近代的な）極値解析が導入．

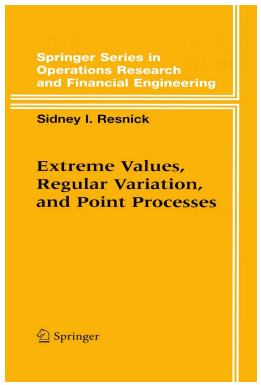
そのころ，日本では，．．．さまざまな候補分布とデータとの適合性の議論

合田良実，宝馨・高棹琢馬 共著 “水文頻度解析における確率分布モデルの評価規準”への討議・回答，土木学会論文集, 405, p.265-272, 1989.

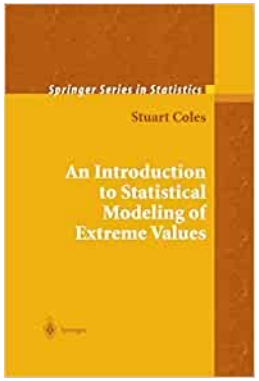
Galambos, J.(1978): The Asymptotic Theory of Extreme Order Statistics の後，現在まで，極値理論の進展



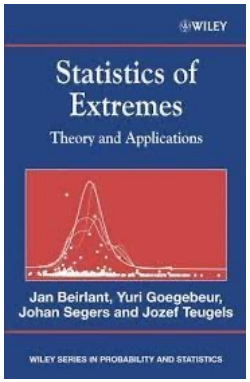
1983 年



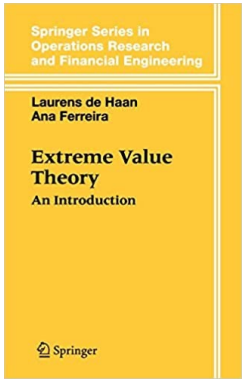
1987 年



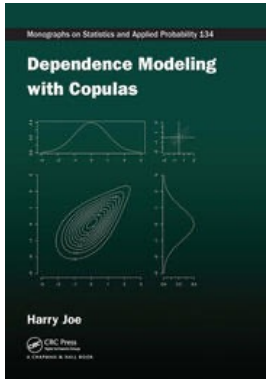
2001 年



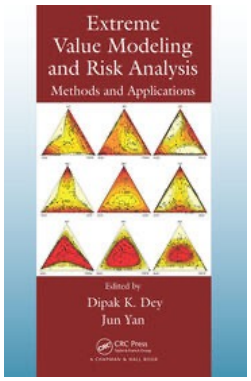
2004 年



2006 年



2015 年



2016 年

この他にも重要な極値理論のテキストが多くある（日本語では，高橋・志村, 2016; 西郷・有本, 2020）

応用水文統計学（岩井・石黒, 1970）にも極値理論は記載

ポアソン分布との関連性：　$P_n(y) = \left.\dfrac{\lambda^k \exp(-\lambda)}{k!}\right|_{k=0,\,\lambda=\lambda_n(y)} = \exp\{-\lambda_n(y)\}$

（単変量も他変量も同じ）

$$P_n(y_{\mathrm{A}}, y_{\mathrm{B}}) = \exp\left\{-\lambda_{n,(\mathrm{AB})}\right\} \left.\sum_{i=0}^{k_{\mathrm{A}} \wedge k_{\mathrm{B}}} \frac{\lambda_{n,\mathrm{AB}}^{i}}{i!} \frac{\lambda_{n,\mathrm{A|B}}^{k_{\mathrm{A}}-i}}{(k_{\mathrm{A}}-i)!} \frac{\lambda_{n,\mathrm{B|A}}^{k_{\mathrm{B}}-i}}{(k_{\mathrm{B}}-i)!}\right|_{k_{\mathrm{A}}=k_{\mathrm{B}}=0,\ \lambda_{n,(\mathrm{AB})}=\lambda_{n,(\mathrm{AB})}(y_{\mathrm{A}},y_{\mathrm{B}}),\ \cdots}$$

$$= \exp\left\{-\lambda_{n,(\mathrm{AB})}(y_{\mathrm{A}}, y_{\mathrm{B}})\right\}$$

閾値超過極値に対する確率（GP 分布）：　$1 - P_u(y) = \dfrac{\lambda_n(y)}{\lambda_n(u)} = \left(1 + \xi \dfrac{y-u}{\sigma_u}\right)^{-1/\xi}$

$$P_u(y_{\mathrm{A}}, y_{\mathrm{B}}) = \frac{\lambda_{n,(\mathrm{AB})}(y_{\mathrm{A}} \wedge u_{\mathrm{A}}, y_{\mathrm{B}} \wedge u_{\mathrm{B}}) - \lambda_{n,(\mathrm{AB})}(y_{\mathrm{A}}, y_{\mathrm{B}})}{\lambda_{n,(\mathrm{AB})}(u_{\mathrm{A}}, u_{\mathrm{B}})}$$

補足： 多変量極値（２変量の場合を例示），$V(x_A, x_B) = -\log F(x_A, x_B)$ に対して，

斉次性：　$V(x_A, x_B) = V(x_A/m, x_B/m)/m$

が成立することが不可欠．

（これは多変量の生起率の比例性を表す．例えば，再現期間５０年か２００年を超えるイベント数に対し，再現期間１０年か４０年を超えるイベント数は５倍の多さとなる．
これを１次元に縮退させたものから，１次元の極値分布に必要な生起率関数を得ることも可）

この時，横断分布 $H(t)$ を導入すれば，　$V(x_A, x_B) = \displaystyle\int_0^1 \left(\frac{t}{x_A} \vee \frac{1-t}{x_B}\right) H(dt)$

と表される．横断分布の平均は定義域 [0, 1] の中点にあることが要件（唯一の制約条件）．

ここで重要なことは，多変量極値では，外力の大きさを再現期間に変換して，その従属性を検討するのである．
逆に言えば，外力の大きさそのもので従属性を論じることは，本質的な方向性を見誤ることになる．

1995-1997 Neptune Project にて，２＆３変量（多変量）極値理論を海岸堤防の設計（破壊確率の算定）に適用
de Haan, L. & J. de Ronde, Sea and wind: multivariate extremes at work, 1998.

both photos' source: wikipedia

ポアソン確率のヒミツ

表 6.2　プロシア騎兵連隊において馬に蹴られて死んだ兵士数(ボルトキーヴィッチ) deaths from horse kicks 統計学入門：東京大学出版会

死亡数	0	1	2	3	4	5
観測数	109	65	22	3	1	0
理論値	108.7	66.3	20.2	4.1	0.6	0.1

平均
１２２

Ladislaus von Bortkewitsch（1868 –1931）

Siméon Denis Poisson (1781-1840)

平均（生起数）が分かれば，生起確率を求めることができる

事象が生じない確率 $p(k=0) = e^{-\lambda}$

２００年間で１２２回生じるイベントが生じない年は，１０９年あることが分かる．

$$e^{-\frac{122}{200}} = \frac{108.7}{200} (= 0.543) > 0.5$$

３分の１の法則

（平均１回生じる事象は，
生じないことより，生じる可能性が
２倍ある）

$$e^{-1} \approx \frac{1}{3} = 1 - \frac{2}{3}$$

３の法則

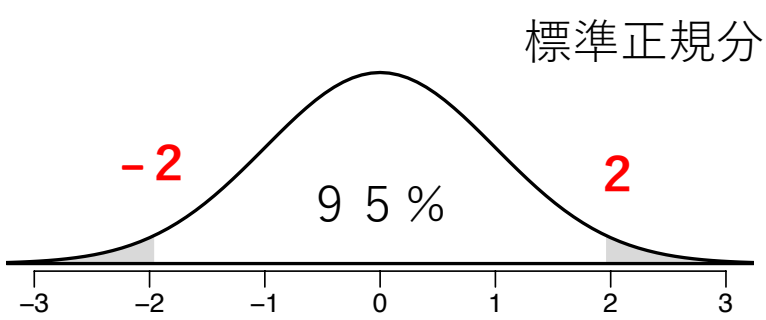
（平均３回生じる事象でも，
たまたま生じていないこともある）

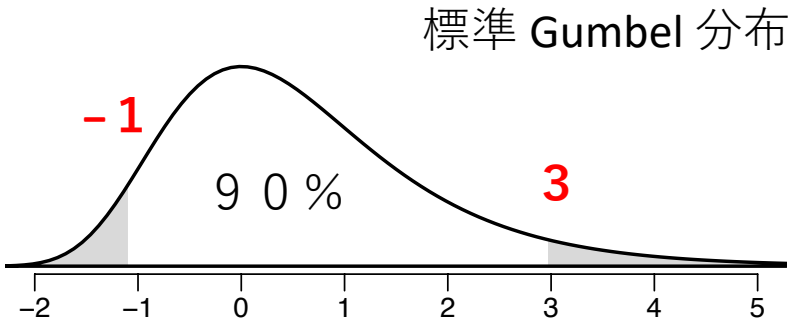
$$e^{-3} \approx \frac{1}{20} = 0.05$$

平均生起数 λ と外力 y を結びつけて考えるのが，極値統計理論．

追加イメージ（1）

（3の法則と1／3の法則を用いて分かること）





追加イメージ（2）

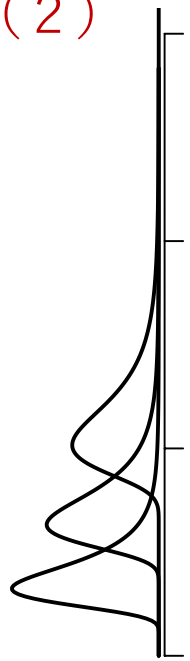
外挿とは？

単なる延長
というより，
もう少し深い
背景がある．

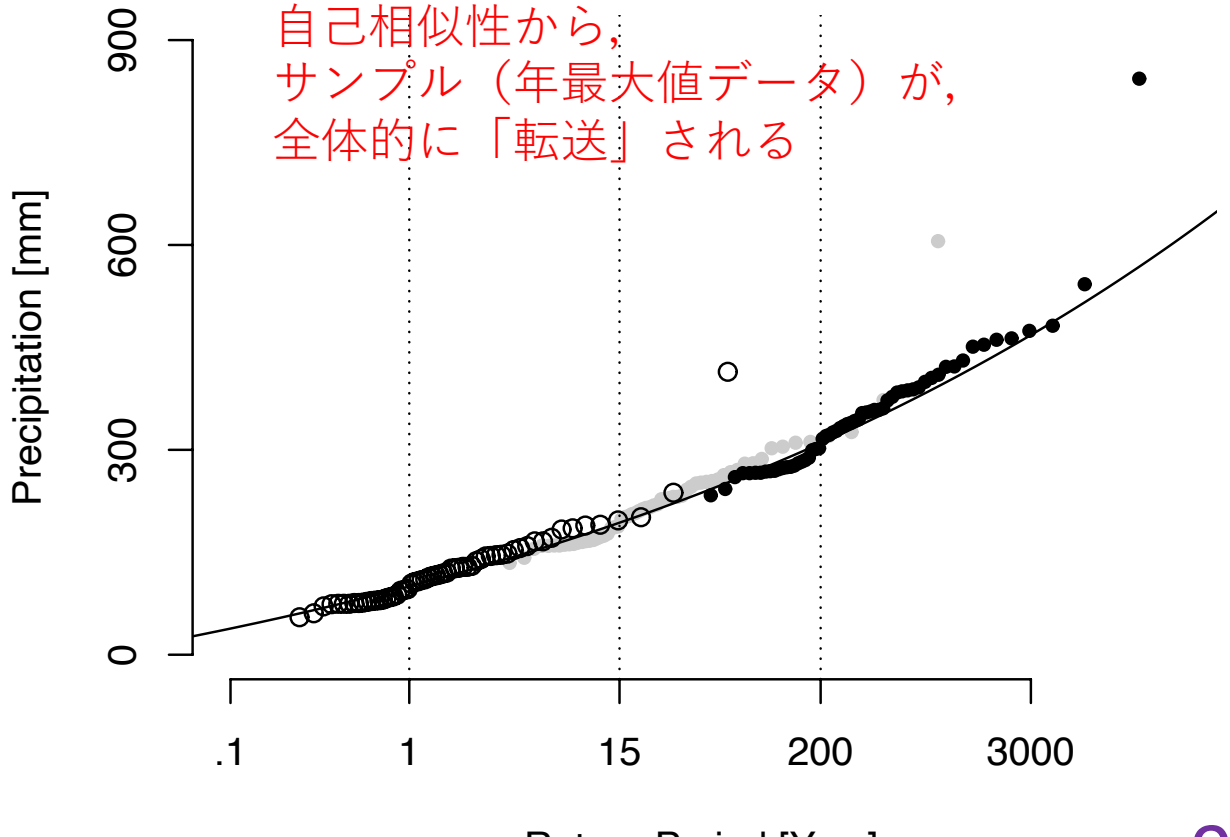
200-yr. max.

15-yr. Max.

1-yr. Max.

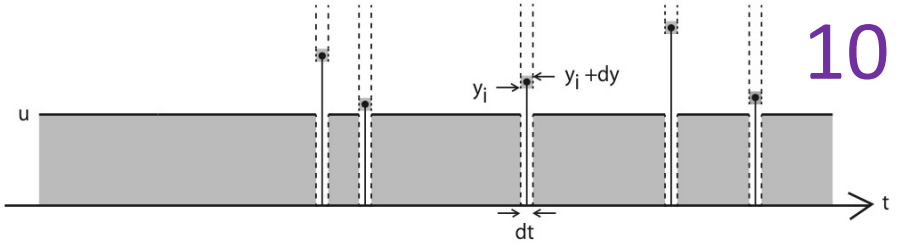


自己相似性から，
サンプル（年最大値データ）が，
全体的に「転送」される



ポイントプロセスモデル（点過程モデル）

GP よりも柔軟表現なのに現在の水工で普及していない．
ただし，r や u を幾らにとるか悩ましい問題は変わらない．



$$L(\mu_R, \sigma_R, \xi) \propto \prod_{i=1}^{n} \exp\left\{-\frac{n\lambda_R(Y_{i,r})}{R}\right\} \prod_{j=1}^{r}\left\{-\frac{\mathrm{d}\lambda_R(Y_{i,j})}{\mathrm{d}y}\right\}$$

（各ブロック上位 r 番めまでの最大値）

$$L(\mu_R, \sigma_R, \xi) \propto \exp\left\{-\frac{n\lambda_R(u)}{R}\right\} \prod_{j=1}^{m}\left\{-\frac{\mathrm{d}\lambda_R(Y_j)}{\mathrm{d}y}\right\}$$

（閾値 u を超過した極大値）

ここで言いたいことは，「外力に対する分布に当てはめる」というよりも，「生起数に対するポアソン分布に当てはめている」ということである．このようなことができるのは，極値理論（から導かれる生起率関数）を適用するからである．極値理論とは異なる理論体系を構築するのは，非常に困難であり，荊棘の道をすすむことになるだろう．ただし，極値統計理論の適用にあたり，適用すべき極値（すなわち，裾のデータ）を慎重に抽出する必要がある．
データに合う分布を求めるのではなく，分布に合うようにデータの刈り込みをする必要がある．

Goodness of fit 適合性 と Model selection モデル選択

適合性を検定する場合，適合していることを帰無仮説にとり，適合していないこと（対立仮説）を検出しようと検定している（したがって，積極的に「適合」を検討しているわけではない）

モデルとデータの距離を測る情報量基準に基づいて，モデルの選択を行う．
比較するに値するモデル間に対して，比較を行うべきであろう（Compare apple to orange）．

データに語らせる手法（data-driven approach）というと，きこえが良く，より客観的であることを印象付けるが，統計的解析には，**基本的に，主観が必ず入る（はずだ）**．

統計的有意性とP値に関する米国統計協会の声明（2016）

The ASA Statement on p-Values: Context, Process, and Purpose, by Wasserstein, R. L. & N. A. Lazar, The American Statistician, 20, 129-133, 2016

日本計量生物学会，佐藤俊哉（訳），2017 掲載

科学的結論の土台となっているのは「統計的有意性」という概念であり，通常，P値とよばれる指標で評価される．

P値は，データと特定の統計モデル（仮説を含む）が矛盾する程度を示す指標のひとつ．

P値が0.05ということは、**その仮説が95%の確率で正しいことを意味するものではない**。これは、帰無仮説が真であり、他の全ての仮定が正しいならば、観察されたような極端な結果が得られる確率が5%あることを意味する。
そしてまた、**P値は知見の重要性を示すものでもない**。

「可否」による2分類の決定は実用的であるが，P値だけで決定が正しいかどうかが保証されるものではない．

科学的な主張や結論を正当化するために、データ解析や科学的推論を機械的で明白なルール（「P≤0.05」といった）におとしめるようなやり方は、**誤った思いこみ**と**貧弱な意思決定**につながりかねない。
二分割された一方の側で、結論が直ちに「真実」となったり、他方の側で「誤り」となったりすることは**ありえない**。

P値は有用な統計指標ではあるが、誤用と誤解がまかり通っている。このことにより、一部の学術雑誌ではP値の利用を控えさせたり、一部の科学者や統計家がP値の使用をやめるよう勧めたりしているが、**その際の主張はP値が導入されたときから本質的に変わっていない。**

ある科学誌は、P値を使った論文の発表を禁止するという思い切った対策を打っている。生物統計学者Andrew Vickersは、この科学誌の試みを、交通事故を減らすために自動車を運転しないように呼びかけるようなもので、逆効果かもしれないと言う。おそらく、呼びかけたい人々の多くがこのメッセージを無視するからだ。Vickersは、「**統計をレシピではなく科学として扱う」ことを研究者に教えるべきだ**と言う。

統計学者 Andrew Gelman は、**P値についての理解が深まったとしても、統計を利用して、あり得ないレベルの確実性を作り出したいという人間の衝動がなくなることはないと警告**する。「**人間は、現実には手にすることができないものを追い求めます**」と彼は言う。「**確かさが欲しいのです**」。

Statisticians issue warning over misuse of P values, by Monya Baker, Nature, 531, p.151 (2016)

昔の significant の意味：なにかいつもとちがうことが起こっているからちょっと注意してみた方がよいですよ，
現在の significantの意味：なにか重要なことが起こっている（に変化）
そして，統計的に意味があることは，〇〇学的にも意味がある，という誤解に．

山中伸弥，青井貴之，佐藤俊哉，医療統計学の専門家を交えた鼎談，蛋白質核酸酵素，54，1792～1803

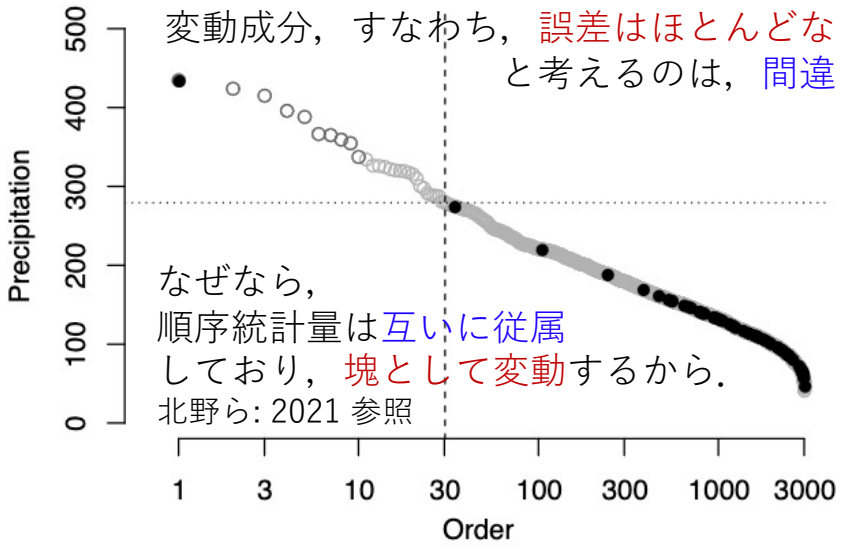
水文統計における「プロクルステスの寝台」問題

人里離れた谷道を行く旅人に，「日も暮れてきたので，しばらく休んで行きませんか．私の寝台は大変珍しいもので，どんな旅人にもよく合うので，旅の疲れもあっという間に癒えるでしょう」と丁重な物言いで自宅に招き，寝台に寝かせた．そして，旅人の体が寝台からはみ出したら，はみ出た分を切り落とし，寝台の長さに足りなければ，頭と脚に縄を結んで無理やり引っ張って，寝台に合うように引き伸ばす．これが，盗賊プロクルステスが，旅人を襲ったやり方で，そのため，ストレッチャー（伸張賊）とも呼ばれていたのだが，英雄テセウスによって，プロクルステスが旅人にやったことと同じやり方で，プロクルステスを寝台に寝かせ，退治したと伝えられている[1,2]．このことから，「プロクルステスの寝台」とは，無理にでも基準に合わせることや杓子定規を戒めるための成句になっている[3]．

1) Baldwin, J.: Old Greek Stories, 1895. https://www.gutenberg.org/ebooks/11582 （杉谷代水 訳，希臘神話，富山房，1909; 復刻版 2011）
2) Bulfinch, T.: The age of fable, 1855. （野上弥生子訳，ギリシア・ローマ神話，岩波書店，1991 ）
3) 例えば，世界大百科事典 第2版，平凡社，https://kotobank.jp/word/プロクルステス-127913

多数アンサンブル標本を扱う時代の**不確実性の捉え方**

滑らかに連続するところに，
変動成分，すなわち，誤差はほとんどないだろう
と考えるのは，間違いである．

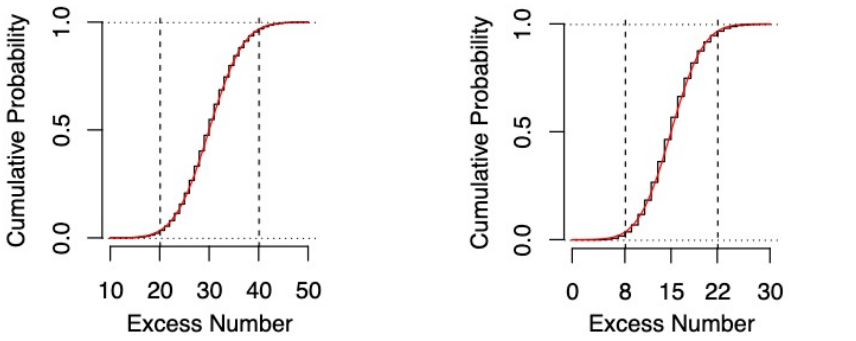


なぜなら，
順序統計量は互いに従属
しており，塊として変動するから．
北野ら: 2021 参照

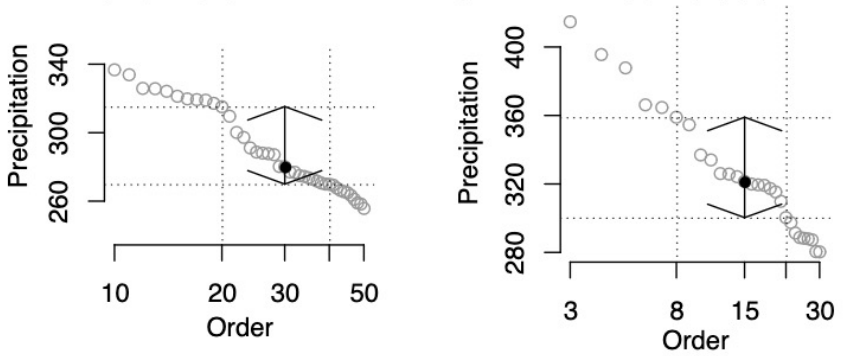
図 -1　3000 年分の年最大日降水量の順序統計量
（順位を示す横軸は対数スケール，
●印は全体の最大値を含むアンサンブルを表示）

$$P(k) = \sum_{j=0}^{k} \binom{3000}{j} 0.01^{j}\, 0.99^{3000-j} \qquad (1)$$

ノンパラメトリックもきこえは良いが，きちんと議論するには，間接量を用いて，メンドクサイ計算が必要．
北野ら: 2017



(a)　総数 3000 における超過確率に対する順位の区間推定
（超過確率 1/100 を左に，1/200 を右に表示）



(b)　確率降水量の区間推定
（再現期間 100 年を左に，200 年を右に表示）
図 -2　ノンパラメトリック法による区間推定

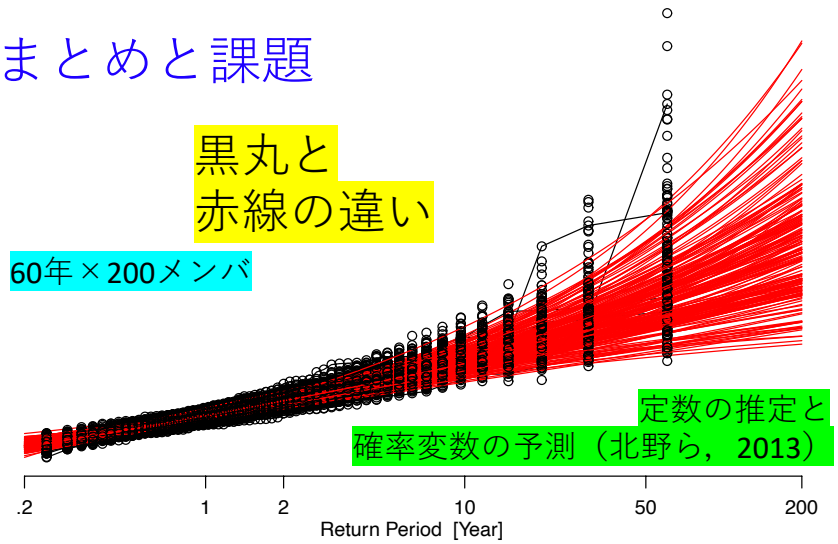
大事なことは，rank と order (stats.)と
そして，誤差に対する理解

誤差変動の基本公式：$\mathrm{V}(\overline{X}_n) = \dfrac{\sigma^2}{n}$
（統計学の基本）

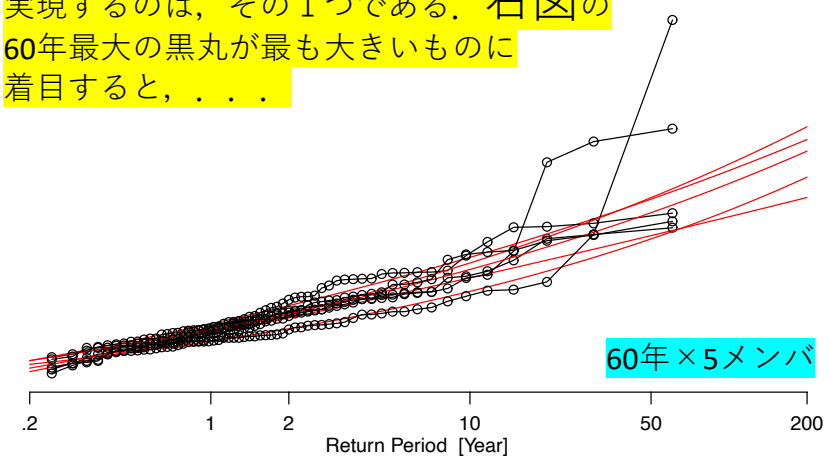
再現レベルに対しては，$\mathrm{V}(\hat{\mu}_R) \approx \dfrac{\sigma_R^2}{K}$
（経験度 K，北野ら: 2008）

北野利一・渡部哲史・小林健一郎: 単純極値変数と超過数の視点から見た降水量の極値の数理特性，土木学会論文集B1（水工学），77, pp.I_1153-I_1158, 2021.
北野利一・高橋倫也・田中茂信: 確率降雨量の統計解析におけるモデル適合性の限界，土木学会論文集B1（水工学），67, pp.I_277-I_282, 2008.

まとめと課題



さらに総体で見れば左図だが，
実現するのは，その１つである．右図の
60年最大の黒丸が最も大きいものに
着目すると，．．．



＊ 極値を扱うためには，極値理論の理解が大切である．

欧米では，歴史的に，極値理論の専門家が水工学研究のプロジェクトに含まれてきた．少なくとも統計の専門家との交流は不可欠（経済・ファイナンスに関わる確率・統計の専門家に意見を求めることも一手）．

＊ できるかぎり，極値理論（点過程を含む）を適用できるような努力をする．

→「プロクルステスの寝台」問題：データの身の丈に合わせた試行錯誤も良いが，確率分布を比べることは簡単では無い（compare apple to orange）．データを極値分布（点過程の尤度）という寝台に合わせるように，もう１つの試行錯誤（cut-and-try）が必要である．極値理論に基づいた結果を最終的に採用するためには，プロクルステスを退治したテセウスの知恵と工夫が必要である．

＊ 多変量極値への展開のための周辺分布である単変量との接続を意識する．

さらに，非定常性の課題の１つとして，Hurst (1951) が発見したロングメモリは，極値の問題と考えている．

＊ 推定誤差と本質的な確率変動，そして，悩ましい記録更新（とび抜けた新記録値）

寝ても覚めても我々を悩まされるのは，神出鬼没な新記録値である．それゆえに，我々は無いものねだりの確からしさを求めがちであることを自覚した意思決定をしなければならない．このことは，合田と宝の討議の時代と問題意識は同じである．この 65 年間の極値理論の進展は，その取組み方を変えるだろう．