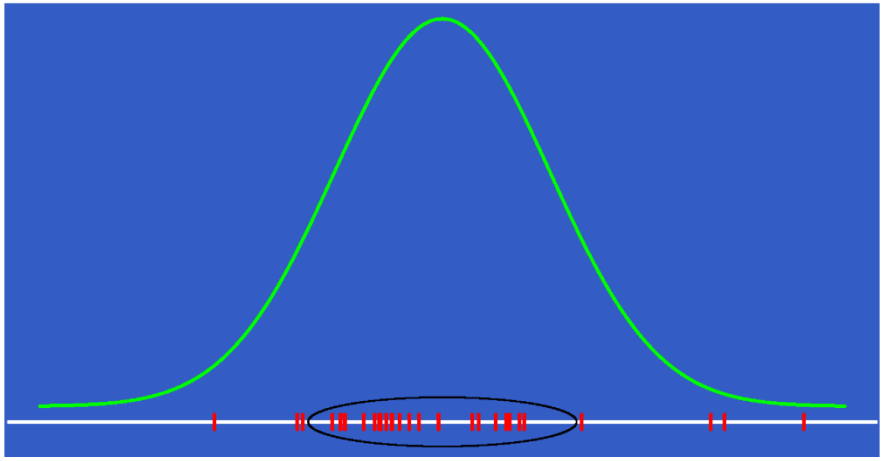
交通大数据

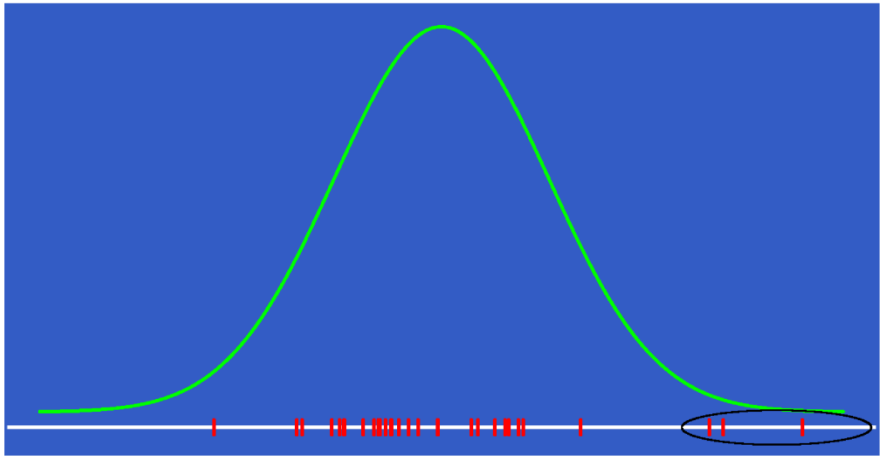
极值统计

- 郭延永
- guoyanyong@seu.edu.cn

1.2 极值理论的起源与发展

极值理论是统计学的一个重要分支，关注的重点是严重背离分布均值的小样本事件，即因发生概率极低而在较短时间内无法观测到的极值事件。





1.2 极值理论的起源与发展

极值理论起源于20世纪初期，主要是在理论层面分析各种随机现象，并从20世纪50年代开始得到广泛关注。期间的代表人物包括：Tippet、Fisher、Gnedenko、Gumbel以及Pickands等。

概述

- 例-1

Pirie港1923-1987年的年最大海平面高度

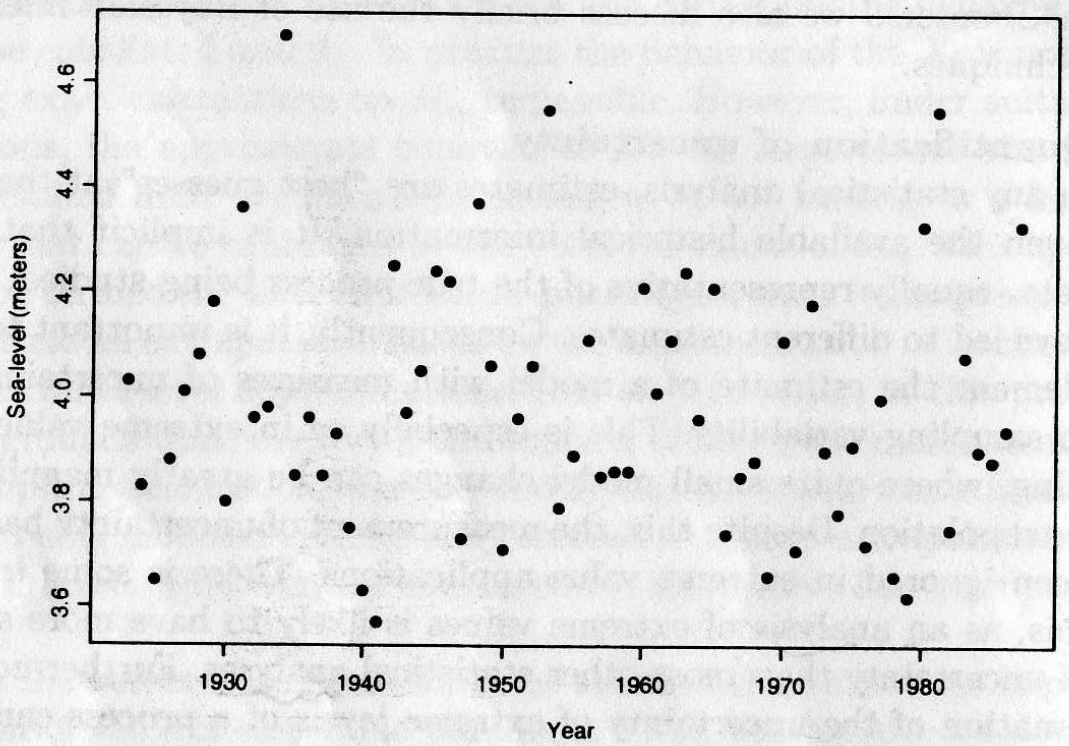
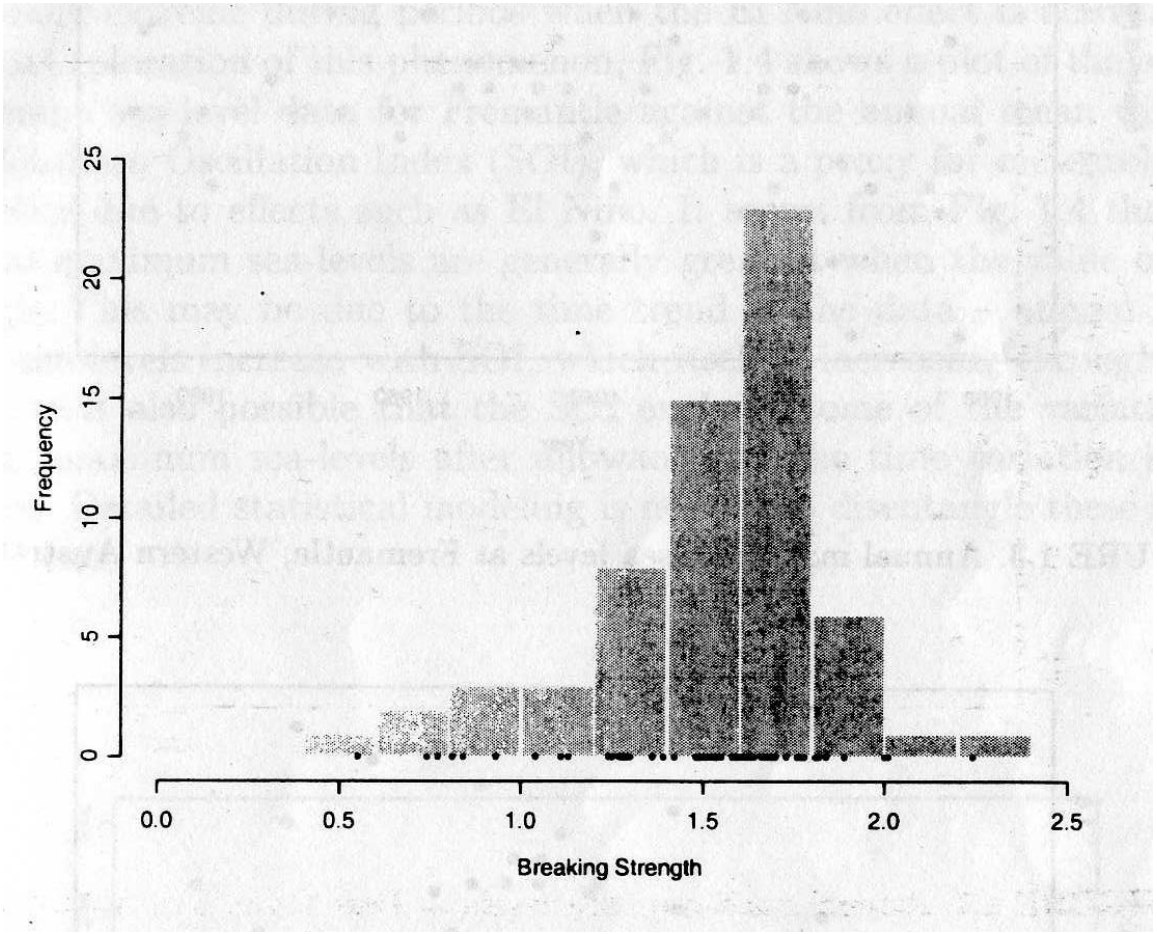


FIGURE 1.1. Annual maximum sea levels at Port Pirie, South Australia.

概述

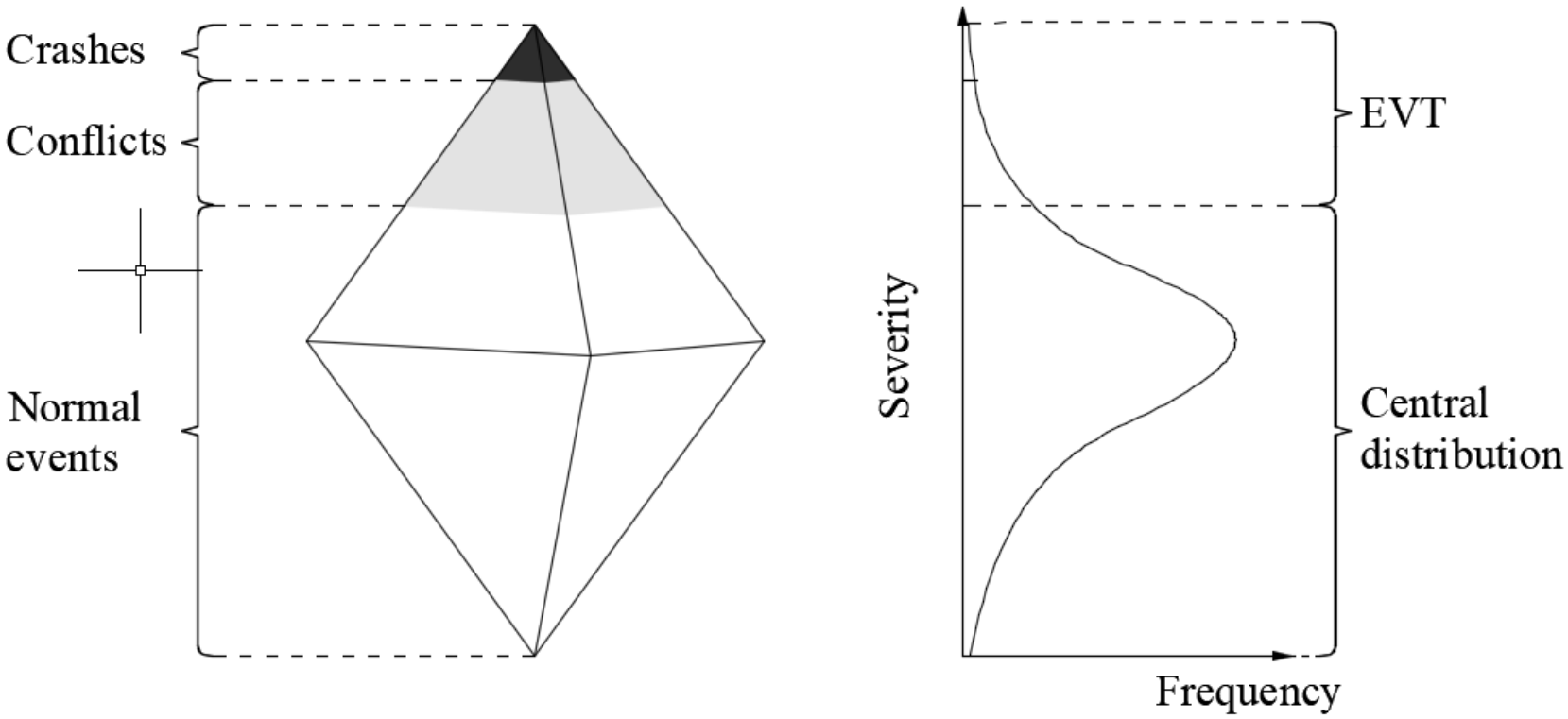
- 例-2

一股由63根玻璃纤维组成的线缆的断裂强度



1.3 交通冲突与极值理论的结合

交通事故是交通冲突的极值事件。



1.2 极值理论的起源与发展

相关研究工作：

Campbell et al. (1996)

Songchitruksa and Tarko (2006)

Zheng et al. (2014a)

Zheng et al. (2014b)

Farah and Azevdo (2017)

Tarko (2012)

Tarko (2018)

…

ELSEVIER

Accident Analysis and Prevention 38 (2006) 811–822

ACCIDENT ANALYSIS & PREVENTION

www.elsevier.com/locate/aap

The extreme value theory approach to safety estimation

Praprut Songchitruksa [a,1], Andrew P. Tarko [b,*]

[a] Texas Transportation Institute, 2929 Research Pkwy, College Station, TX 77843-3135, United States
[b] School of Civil Engineering, Purdue University, West Lafayette, IN 47907, United States

Received 17 June 2005; received in revised form 13 December 2005; accepted 8 February 2006

Abstract

Crash-based safety analysis is hampered by several shortcomings, such as randomness and rarity of crash occurrences, lack of timeliness, and inconsistency in crash reporting. Safety analysis based on observable traffic characteristics more frequent than crashes is one promising alternative. In this research, we proposed a novel application of the extreme value theory to estimate safety. The method is considered proactive in that it no longer requires historical crash data for the model calibration. We evaluated the proposed method by applying it to right-angle collisions at signalized intersections. Evaluation results indicated a promising relationship between safety estimates and historical crash data. Crash estimates at seven out of twelve sites remained within the range of Poisson-based confidence intervals established using historical crash data. The test has yielded large-variance safety estimates due to the short 8-h observation period. A simulation experiment conducted in this study revealed that 3–6 weeks of observation are needed to obtain safety estimates with confidence intervals comparable to those being obtained from 4-year observed crash counts. The proposed method can be applied to other types of locations and collisions as well.

Keywords: Extreme value theory; Traffic conflicts; Surrogate safety measures; Safety estimation; Safety modeling

1. Background

Crash-based safety analysis is hampered by several shortcomings, such as randomness and rarity of crash occurrences, lack of timeliness, and inconsistency in crash reporting. The accidents are rare events and are therefore associated with the random variation inherent in a small number. It is not sufficient to gather the crash data for weeks or months. The typical period to be considered sufficient is as long as 3 years (Nicholson, 1985). Since the current practice for crash-based safety analysis requires years of waiting period for crash data at several locations, safety researchers have sought for alternative approaches to safety estimation without the need to rely on historical crash data.

An impressive amount of work has been done in the past to search and analyze the traffic characteristics that may complement the crash data. The most acknowledged ones include traffic conflicts (Chin et al., 1992; Chin and Quek, 1997; Glauz and Migletz, 1980; Parker and Zegeer, 1989), critical events, e.g., aggressive lane merging, speeding, and running on red (Kloeden et al., 1997; Porter et al., 1999); acceleration noise (Shoarian-Sattari and Powell, 1987); post-encroachment time (Allen et al., 1978); and time-integrated time-to-collision (Minderhoud and Bovy, 2001). Other proposed measures are volume, speed, delay, accepted gaps, headways, shock-waves, and deceleration-to-safety-time (FHWA, 1981). Although some of the latter measures are safety factors rather than surrogate measures, they are listed to adequately reflect the past work. The attempts to confirm the statistical linkage with accident data yielded mixed findings at best. In addition, not all the proposed indicators in the past research satisfy the desirable properties of surrogate measures. For example, a measure such as time-integrated time-to-collision is so data-intensive that it is attainable only in the simulation environment and is therefore not observable in the field (Minderhoud and Bovy, 2001). The traditional approach to the analysis of a surrogate measure of safety shares several common traits, which we can discuss from two aspects, measurement and evaluation.

The evaluation of surrogate measures of safety has been done in several respects in the past. Traffic conflict attracted the

* Corresponding author. Tel.: +1 765 494 5027; fax: +1 765 496 1105.
E-mail addresses: praprut@tamu.edu (P. Songchitruksa), tarko@ecn.purdue.edu (A.P. Tarko).
[1] Tel.: +1 979 862 3559; fax: +1 979 845 9873.

0001-4575/$ – see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.aap.2006.02.003

概述

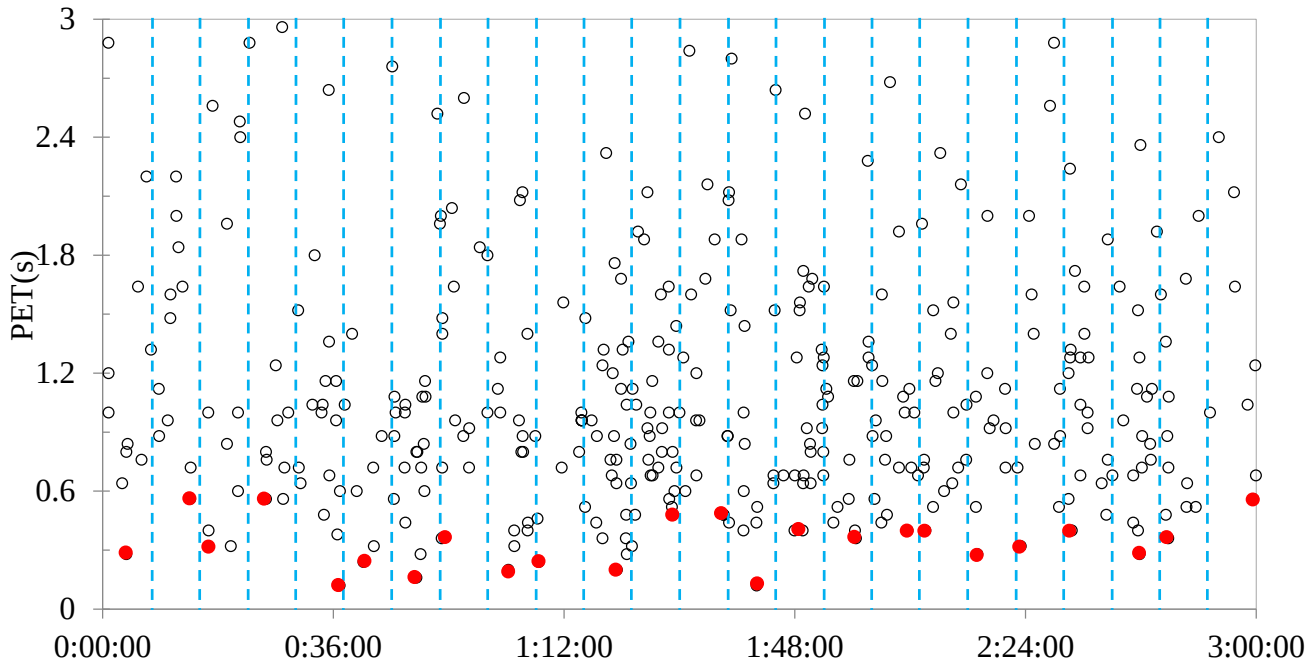
- 极值理论分类

按极值确定方法
- 区组极值理论+广义极值分布
- 超阈值极值理论+广义帕累托分布

按考虑变量的数量
- 一维极值理论
- 多维极值理论(二维、三维及以上)

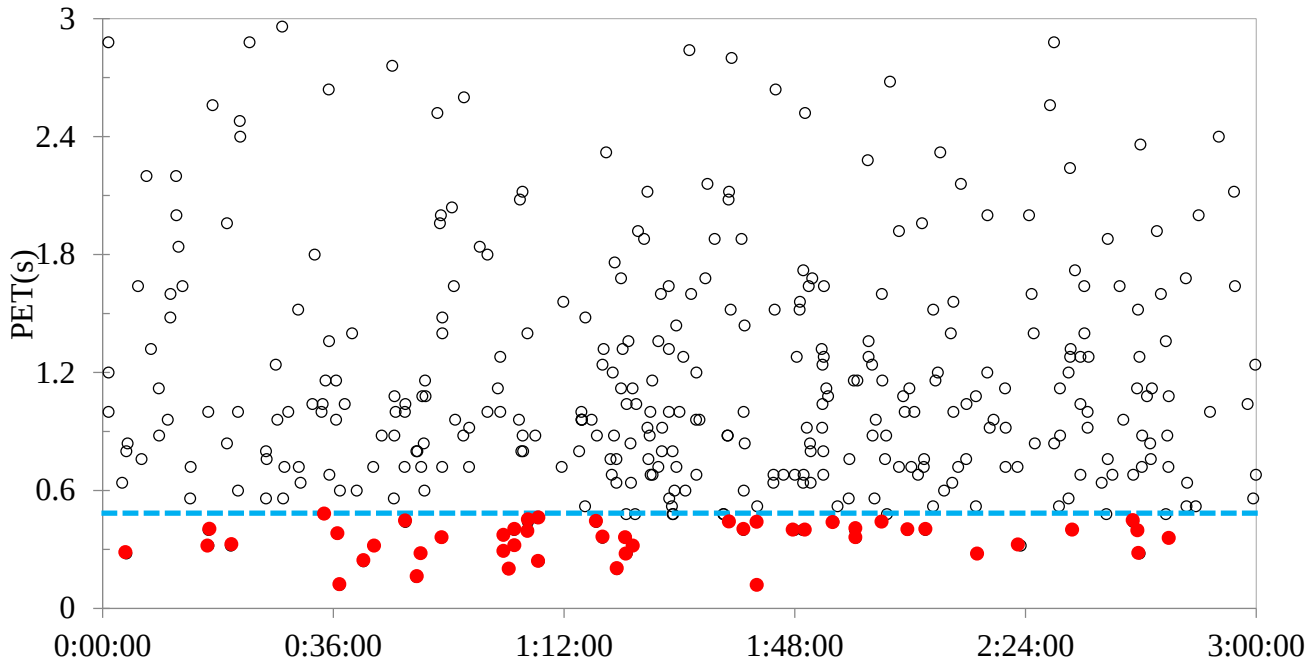
概述

- 区组极值(Block Maxima)



概述

- 超阈值极值(Peak Over Threshold)



3.1 区组极值模型

➢ 模型原理

区组极值是指将样本数据按时间间隔划分成不同区组，然后将每一个区组的最大值视为极值。

假设X_1, X_2, …, X_n是来自同一分布F的一系列独立样本，且$M_n=\max\{X_1, X_2, \ldots, X_n\}$。理论上来讲，$M_n$的分布可以根据$n$个观测样本准确地推导出来，即

$$\Pr\{M_n \le z\} = \Pr\{X_1 \le z, X_2 \le z, \ldots, X_n \le z\}$$

$$= \Pr\{X_1 \le z\} \times \Pr\{X_2 \le z\} \ldots \times \Pr\{X_n \le z\}$$

$$= \{F(z)\}^n$$

3.1 区组极值模型

➢ 模型原理

假设z_+是F分布的上限点，对于任何的$z<z_+$，可以得出当$n\to\infty$时$F^n(z)\to 0$，因此M_n的分布是一个退化分布。为了解决这个问题，引入了规范化处理方法，即

$$M_n^* = \frac{m_n - b_n}{a_n}$$

式中，$\{a_n>0\}$和$\{b_n\}$为一系列常数。通过选择合适的$\{a_n>0\}$和$\{b_n\}$，可以保证随着n的增加M_n^*的位置和尺度均保持稳定，从而避免出现M_n趋向于0的问题。

3.1 区组极值模型

➢ 极值类型定理

假设X_1，X_2，…，X_n为独立同分布的变量，若存在常数数列$\{a_n>0\}$和$\{b_n\}$，使得当$n\to\infty$时，$\Pr\left(\frac{M_n-b_n}{a_n}\right)\to G(z)$，其中 G(·)是非退化分布函数，则G(·)必将属于以下三种分布类型之一：

I：$G(z)=\exp\left\{-\exp\left[-\left(\frac{z-b}{a}\right)\right]\right\}$, -∞ <z< ∞

II：$G(z)=\begin{cases}0, & z\le b\\ \exp\left\{-\left(\frac{z-b}{a}\right)^{-\alpha}\right\}, & z>b\end{cases}$

III：$G(z)=\begin{cases}\exp\left\{-\left[-\left(\frac{z-b}{a}\right)^{-\alpha}\right]\right\}, & z<b\\ 1, & z\ge b\end{cases}$

3.1 区组极值模型

➢ 广义极值分布

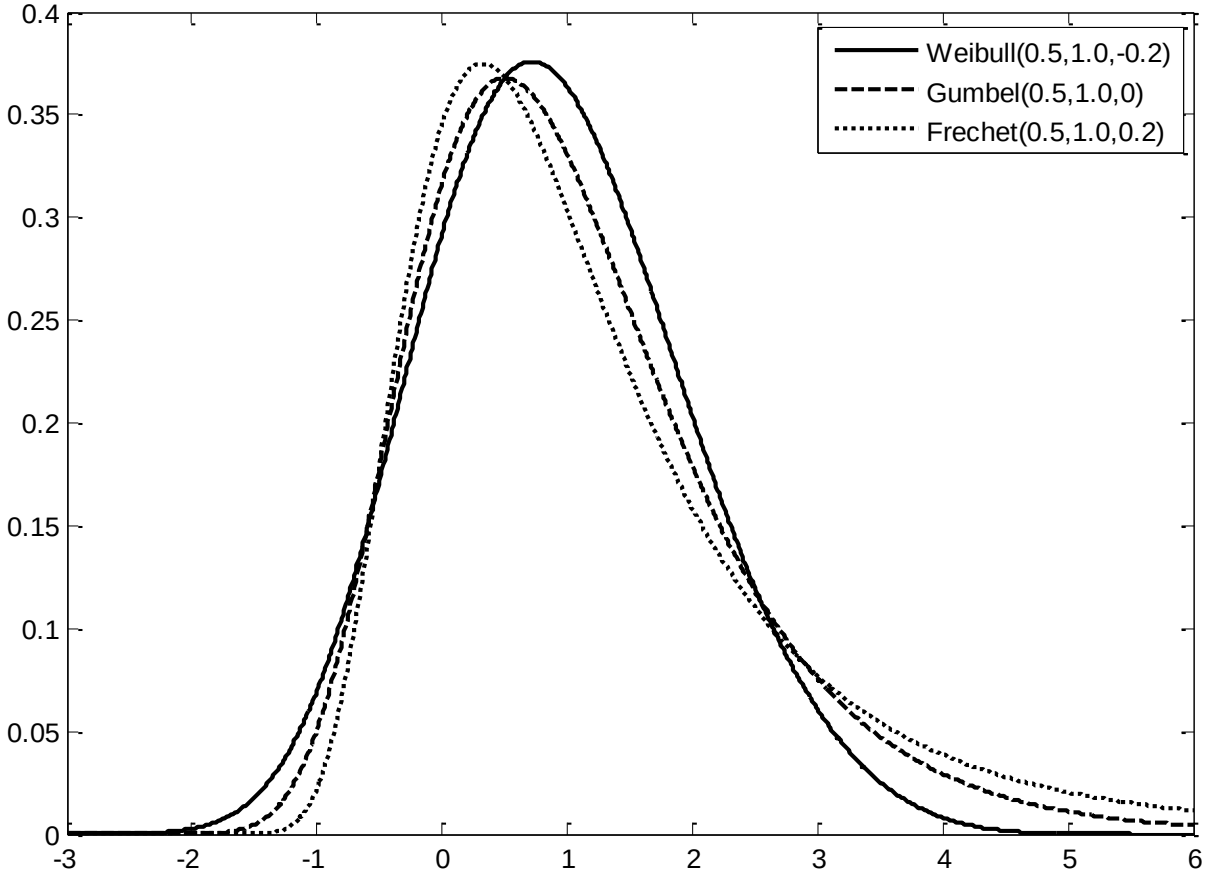
广义极值分布(Generalized Extreme Value, GEV)模型的提出将上述三类分布统一到一个分布之中，从而很好地解决了上述问题。广义极值分布的形式如下：

$$G(z)=\begin{cases}\exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}, & \xi\neq 0\\ \exp\left\{-\exp\left[-\left(\frac{z-\mu}{\sigma}\right)\right]\right\}, & \xi=0\end{cases}$$

式中，$-\infty<\mu<\infty$为位置参数，$\sigma>0$为尺度参数，$-\infty<\xi<\infty$为形状参数，并且该分布需满足$\{z: 1+\xi((z-\mu)/\sigma)>0\}$。$\xi>0$对应的是Fréchet分布，$\xi<0$对应的是Weibull分布，$\xi=0$对应的是Gumbel分布。

3.1 区组极值模型

- 广义极值分布



3.1 区组极值模型

➢ 区组极小值模型

在实际应用中，有时需要对区组极小值进行建模，同样地，令$\widetilde{M}_n=\min\{X_1, X_2, \ldots, X_n\}$且$X_1, X_2, \ldots, X_n$为独立同分布的变量，显然上文所阐述的与$M_n$相关推论可以以相似的方法应用至$\widetilde{M}_n$。

令$Y_i=-X_i(i=1,2, \ldots, n)$，进行这种取负变换也就意味着$X_i$的较小值相当于的$Y_i$的较大值。因此，如果$\widetilde{M}_n=\min\{X_1, X_2, \ldots, X_n\}$而$M_n=\max\{X_1, X_2, \ldots, X_n\}$，那么$\widetilde{M}_n=-M_n$，而当$n$值较大时会有

$$\Pr\{\widetilde{M}_n \le z\} = \Pr\{-M_n \le z\} = \Pr\{M_n \ge -z\} = -\Pr\{M_n \le -z\}$$

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

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

3.1 区组极值模型

➢ r阶广义极值分布模型

由于极值事件的样本量非常小，而基于小样本数据的估计通常会产生较大的偏差。为了解决这个问题，需要寻找能包含更多极值样本而不仅是区组最大值的极值选取方法。现阶段主要有两种方法：一是超阈值极值法，另一种是r阶次序统计方法，即选取每个区组前r个最大值作为极值，通常r会是比较小的值。

假设X_1，X_2，…，X_n是来自总体分布函数F的一系列独立样本，将其按大小顺序排列得到$X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$，称其为次序统计量。定义$M_n^{(k)}$是$\{X_1, \ldots, X_n\}$的$k$阶最大值。假如存在常数数列$\{a_n>0\}$和$\{b_n\}$使得当$n \to \infty$时$\Pr\left(\frac{M_n - b_n}{a_n}\right) \to G(z)$，其中$G(\cdot)$是非退化分布函数，那么$G(\cdot)$就是广义极值分布函数。

3.1 区组极值模型

➢ r阶广义极值分布模型

同理，对于固定的k值，也可以得出在集合$\left\{z: 1+\xi\left(\frac{z-\mu}{\sigma}\right)>0\right\}$中，

$$\Pr\left(\frac{M_n^{(k)}-b_n}{a_n}\right)\to G_k(z)$$

$$G_k(z)=\exp\{-\tau(z)\}\sum_{s=0}^{k-1}\frac{\tau(z)^s}{s!}\quad,\quad \tau(z)=\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}$$

上述广义化过程说明，如果一个区组的r阶最大值都像区组最大值一样进行规范化处理，那么r阶最大值的极限分布同区组最大值的极限分布将会具有相同的形式。

3.1 区组极值模型

➢ r阶广义极值分布模型

假设X_1，X_2，…，X_n为独立同分布的变量，若存在常数数列$\{a_n>0\}$和$\{b_n\}$，使得当$n\to\infty$时$\Pr\left(\frac{M_n-b_n}{a_n}\right)\to G(z)$，其中$G(\cdot)$是非退化分布函数。那么，对于固定的$r$值，当$n\to\infty$时，$\mathbf{M_n^{(r)}}=(\frac{M_n^{(1)}-b_n}{a_n},\ldots,\frac{M_n^{(r)}-b_n}{a_n})$有如下联合概率密度分布函数

$$f\left(z^{(1)},\ldots,z^{(r)}\right)$$

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

3.1 区组极值模型

➢ 广义极值分布模型参数估计方法

● 线性矩估计法

线性矩估计法(L-Moment Estimation, LME)是参数估计的一种基本方法，主要利用样本数字特征(矩)来估计母体数字特征(矩)进而获得参数估计值。

设随机变量X的分布函数为$F=(x,\theta)$，$X^{(1)}\leq X^{(2)}\leq \ldots \leq X^{(n)}$为样本的次序统计量，称

$$\lambda_r = r^{-1}\sum_{k=0}^{r-1}(-1)^k\binom{r-1}{k}EX^{(k+1)}, r=1,2,\ldots$$

为r阶线性矩，其中E表示数学期望。

3.1 区组极值模型

➢ 广义极值分布模型参数估计方法

● 线性矩估计法

前3阶样本线性矩的无偏估计为：

$$\begin{cases} \tilde{\lambda}_1 = \sum_{i=1}^{n} x^{(i)} \\ \tilde{\lambda}_2 = \sum_{i>j}^{n} (x^{(i)} - x^{(j)})/n(n-1) \\ \tilde{\lambda}_3 = \sum_{i>j>k}^{n} 2(x^{(k)} - 2x^{(j)} + x^{(i)})/n(n-1)(n-2) \end{cases}$$

定义$\tilde{\tau}_3 = \tilde{\lambda}_3/\tilde{\lambda}_2$为线性矩的偏态系数，当$-0.5 \le \tilde{\tau}_3 \le 0.5$时，$(\mu, \sigma, \xi)$的计算公式为：

$$\begin{cases} \mu = \lambda_1 - \dfrac{\sigma[1-\Gamma(1+\xi)]}{\xi} \\ \sigma = \lambda_2 \xi \left[\left(1 - 2^{-\xi}\right) \Gamma(1+\xi) \right] \\ \xi \approx 7.8590c + 2.9554c^2 \end{cases}$$

式中，$c = \dfrac{2}{3}\tilde{\tau}_3 + 1.37$。

3.1 区组极值模型

- 广义极值分布模型参数估计方法
  - 极大似然估计法

极大似然估计法(Maximum Likelihood Estimation, MLE)是建立在极大似然理论上的一种常见的参数估计方法。假设$\{Z_1, Z_2, ..., Z_n\}$是来自广义极值分布的独立同分布集合。当$\xi \neq 0$时，广义极值分布的对数似然函数为：

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

式中，$1+\xi\left(\frac{x_i-\mu}{\sigma}\right)>0$，$i=1, 2, ..., n$。当$\xi=0$时，广义极值分布变成Gumbel分布，其对数似然函数为：

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

3.1 区组极值模型

- 广义极值分布模型参数估计方法
  - 极大似然估计法

由于广义极值分布的边界是所求参数的函数，致使其可能不能满足极大似然估计所需的正则条件。而非正则性意味着不能直接通过极大化似然函数来得到参数估计值。Smith(1985)的研究表明：

当$\xi>-0.5$时，极大似然估计是正则的，即具有通常的渐进性质；

当$-1<\xi<-0.5$时，虽然存在极大似然估计，但不满足标准的渐进性质；

当$\xi<-1$时，极大似然估计不存在，估计值为不可靠估计。

3.1 区组极值模型

➢ 广义极值分布模型参数估计方法

● 贝叶斯估计法

贝叶斯估计是基于概率密度估计的一种参数估计方法，本质是通过贝叶斯决策得到参数θ的最优估计，使得总期望的风险最小。贝叶斯估计的基本步骤是：

①设置参数θ的先验分布$p(\theta)$；

②根据样本x_i的密度分布$p(z_i|\theta)$得到样本集合的联合分布$p(Z|\theta)=\prod_{i=1}^{n}p(z_i|\theta)$；

③由贝叶斯公式计算θ的后验分布$p(\theta|Z)=\frac{p(Z|\theta)p(\theta)}{\int p(Z|\theta)p(\theta)\mathrm{d}\theta}$；

④得到θ的最优估计$\theta^*=\int \theta p(\theta|z)\mathrm{d}\theta$。

由于无法直接通过后验分布估计出其参数值，通常借助马尔可夫链-蒙特卡罗仿真(Markov Chain Monte Carlo, MCMC)来获得参数值。

3.1 区组极值模型

- 广义极值分布模型参数估计方法
  - 估计方法对比

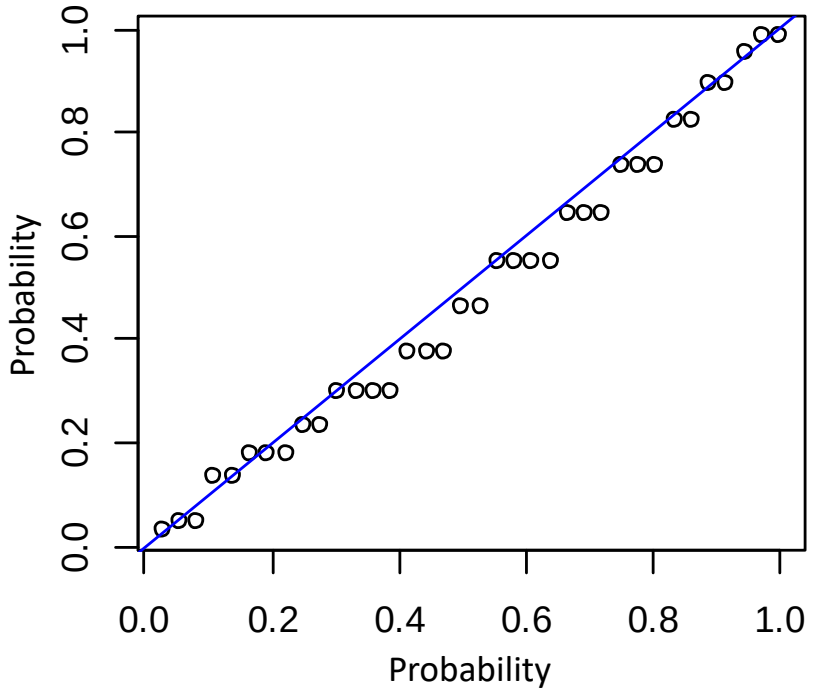
	优点	缺点
线性矩估计法	概念清晰、计算简洁	样本信息利用率低
极大似然估计法	无偏性、有效性、不变性	正则条件限制
贝叶斯估计法	融入先验知识、无正则条件限制	稳定性、收敛性问题

3.1 区组极值模型

➢ 模型检验

● PP图

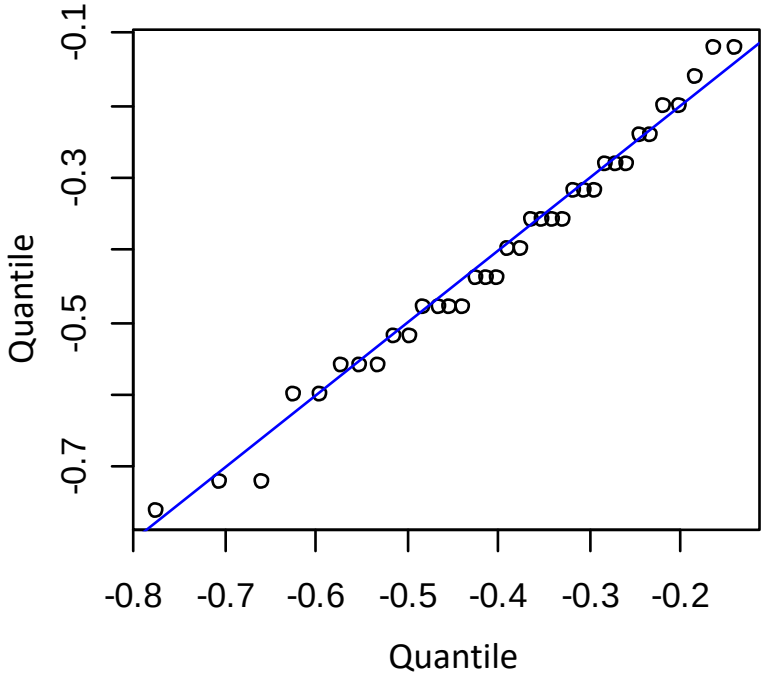
设有一组独立同分布的次序统计量$x_{(1)}\leq x_{(2)}\leq\ldots\leq x_{(n)}$，其分布函数为$F$，那么对应于$x_{(i)}\leq x\leq x_{(i+1)}$的经验分布函数被相应地定义为$\tilde{F}(x)=\frac{i}{n+1}$。假设该组次序统计量的估计分布函数为$\hat{F}$，那么PP图即为由点$\left\{\left(\hat{F}\left(x_{(i)}\right),\frac{i}{n+1}\right):i=1,2,\ldots,n\right\}$所描述的图形。



3.1 区组极值模型

- 模型检验
  - QQ图

QQ 图 是 由 点 $\{(\hat{F}^{-1}(x_{(i)}), x_{(i)}): i =$

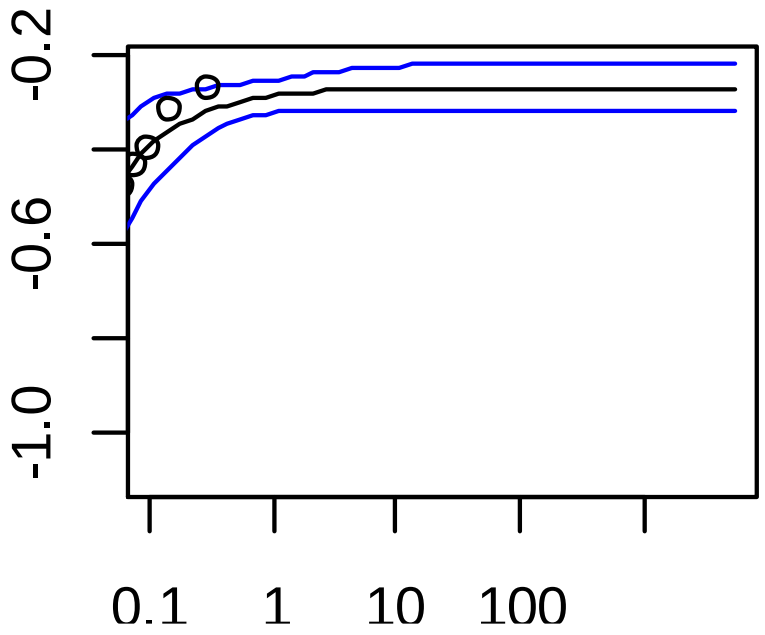


3.1 区组极值模型

➢ 模型检验

● 重现水平图

重现水平图是极值模型诊断检验的一种常用图形，由点$\{(\log y_p, \hat{z}_p): 0 < p < 1\}$所构成，其中$y_p$=-log(1-$p$)；$\hat{z}_p$表示的是重现水平的估计值。如果估计的极值模型是合理的，那么基于该模型的重现水平值和经验的重现水平值应当基本一致，即重现水平图上的点应该落在一条线上。



3.2 超阈值极值模型

➢ 模型原理

假设$X_1, X_2, \ldots, X_n$是来自同一分布F的一系列独立样本并选取一个阈值u，显然当u足够大时所有大于阈值u的样本均可以被视为极值样本。从样本集合中随机选择一个样本X，与其对应的极值事件可由以下条件概率进行描述

$$\Pr\{X > u + y | X > u\} = \frac{1 - F(u + y)}{1 - F(u)}, y > 0$$

如果分布函数F已知，上式所示的阈值超出量的分布函数也可以相应地求出。然而，在实际应用中，F通常是未知的。于是，同区组极值的广义极值分布一样，也需要对超阈值极值分布进行近似估计。

3.2 超阈值极值模型

➢ 广义帕累托分布

根据广义极值分布的相关定理可知，当n足够大时，$F^n(\mathrm{z}) \approx \exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}$。因此，

$$n\log F(z) \approx -\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}$$

当z比较大时，利用泰勒展开可以得到$\log F(z) \approx -\{1-F(z)\}$，将其代入上式并经过适当的变换可得

$$1-F(u) \approx \frac{1}{n}\left[1+\xi\left(\frac{u-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}$$

同样地，对于$y>0$可以得到

$$1-F(u+y) \approx \frac{1}{n}\left[1+\xi\left(\frac{u+y-\mu}{\sigma}\right)\right]^{-1/\xi}$$

3.2 超阈值极值模型

- 广义帕累托分布

将式(2)和（3）代入式(1)可得

$$\Pr\{X > u + y | X > u\} \approx \frac{n^{-1}[1 + \xi(u + y - \mu)/\sigma]^{-1/\xi}}{n^{-1}[1 + \xi(u - \mu)/\sigma]^{-1/\xi}}$$

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

$$= \left[1 + \frac{\xi y}{\tilde{\sigma}}\right]^{-1/\xi}$$

式中，$\tilde{\sigma} = \sigma + \xi(u - \mu)$。

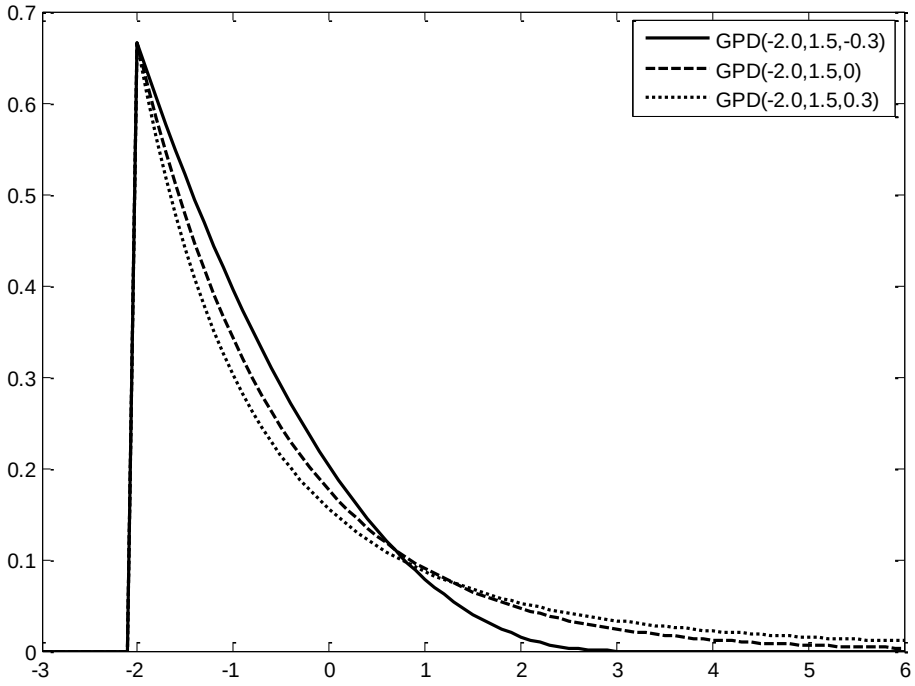
3.2 超阈值极值模型

➢ 广义帕累托分布

对于一个足够大的阈值u，在$X>u$的条件下，阈值超出值$y=(X-u)$的分布近似为以下分布形式：

$$H(y)=\left[1+\frac{\xi y}{\tilde{\sigma}}\right]^{-1/\xi}$$

$$H(y)=1-\exp\left(-\frac{y}{\tilde{\sigma}}\right)$$



3.2 超阈值极值模型

➢ 阈值选取

阈值的选取同样关系到偏差与方差的平衡。如果阈值过大，会导致超阈值极值样本量较少，使得参数估计结果的方差偏大；如果阈值过低，部分非极值事件会被视为极值样本，违背了极值理论的渐进原则，导致产生有偏估计。

- 数值方法

均方差法、阈值自动选取法、多重阈值选取法

- 图解方法

Hill图、**平均剩余寿命图**和**阈值稳定性分析图**等。

3.2 超阈值极值模型

➢ 平均剩余寿命图

假设变量Y~GPD(σ,ξ)，那么当ξ<1时，该变量的均值为

$$E(Y)=\frac{\sigma}{1-\xi}$$

当$\xi\geq1$时，均值为无限值。对于一组序列X_1，X_2，…，X_n，假设其超出阈值u_0的样本服从广义帕累托分布，则可知

$$E(X-u_0|X>u_0)=\frac{\sigma_{u_0}}{1-\xi}$$

式中，σ_{u0}表示超出阈值u_0的样本所对应的尺度参数。如果超出阈值u_0的样本服从广义帕累托分布，那么对任意一个阈值u>u_0，其阈值u超出值同样服从广义帕累托分布，只是极值样本的尺度参数会发生变化，即$\sigma_u=\sigma_{u_0}+\xi(u-u_0)$。

3.2 超阈值极值模型

➢ **平均剩余寿命图**

可以得到

$$E(X-u|X>u)=\frac{\sigma_u}{1-\xi}=\frac{\sigma_{u_0}+\xi(u-u_0)}{1-\xi}=\frac{\sigma_{u_0}-\xi u_0}{1-\xi}+\frac{\xi}{1-\xi}u$$

所以，当$u>u_0$时，$E(X-u|X>u)$是阈值u的线性函数，而$E(X-u|X>u)$是超出阈值u的样本的超出部分的均值的经验估计值。由此可以得出，在超出阈值u的样本服从广义帕累托分布的条件下，这些经验估计值应随着u的改变而线性变化。

下式所描述的一系列点

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

所构成的图即为平均剩余寿命图。

3.2 超阈值极值模型

➢ 阈值稳定性分析图

由超阈值极值的基本理论可知，如果超出阈值u_0的样本服从广义帕累托分布，那么对任意一个阈值$u>u_0$，其阈值u超出值同样服从广义帕累托分布，并且其形状参数一致，但尺度参数会发生变化。

对于阈值u，其尺度参数将变为$\sigma_u = \sigma_{u_0} + \xi(u - u_0)$。当$\xi \neq 0$时，$\sigma_u$会随着阈值$u$的变化而变化。这里定义

$$\sigma^* = \sigma_u - \xi u$$

式中，σ^*—修正后的尺度参数，其不随阈值u的变化而变化。

因此，如果超出阈值u_0的样本服从广义帕累托分布，任意一个阈值$u>u_0$，其估计得到参数ξ和σ^*是基本不变的。

3.2 超阈值极值模型

➢ 阈值稳定性分析图

由超阈值极值的基本理论可知，如果超出阈值u_0的样本服从广义帕累托分布，那么对任意一个阈值$u>u_0$，其阈值u超出值同样服从广义帕累托分布，并且其形状参数一致，但尺度参数会发生变化。

对于阈值u，其尺度参数将变为$\sigma_u = \sigma_{u_0} + \xi(u - u_0)$。当$\xi \neq 0$时，$\sigma_u$会随着阈值$u$的变化而变化。这里定义

$$\sigma^* = \sigma_u - \xi u$$

式中，σ^*—修正后的尺度参数，其不随阈值u的变化而变化。

因此，如果超出阈值u_0的样本服从广义帕累托分布，任意一个阈值$u>u_0$，其估计得到参数ξ和σ^*是基本不变的。

3.2 超阈值极值模型

➢ 阈值选取流程

综合平均剩余寿命图和阈值稳定性分析图，确定阈值选取的基本流程如下：

①依据平均剩余寿命图，选择一个阈值范围$\boldsymbol{R_1}$，该范围内阈值u的平均剩余寿命线近似线性；

②依据阈值稳定性分析图，确定一个范围$\boldsymbol{R_2}$，该范围内修正后的尺度参数和形状参数能够基本保持稳定；

③取上述两个范围的交集，$\boldsymbol{R}=\boldsymbol{R_1}\cap\boldsymbol{R_2}$，以集合$\boldsymbol{R}$的上确界$u_+$为最终值。

3.2 超阈值极值模型

➢ **超阈值模型估计方法**

- 线性矩估计法

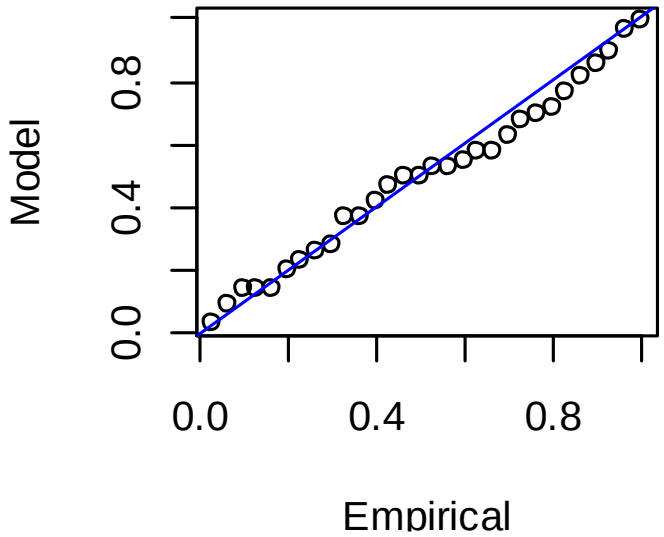
$$\begin{cases} \xi = \dfrac{1}{2}\dfrac{E^2(y)}{s^2-1} \\ \sigma = \dfrac{1}{2}E(y)\dfrac{E^2(y)}{s^2+1} \end{cases}$$

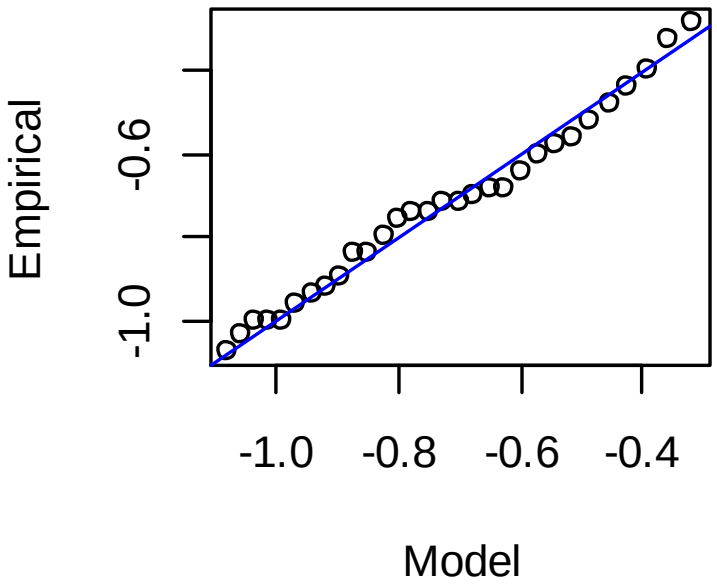
式中，$E(x)$、s^2——样本的均值和方差。

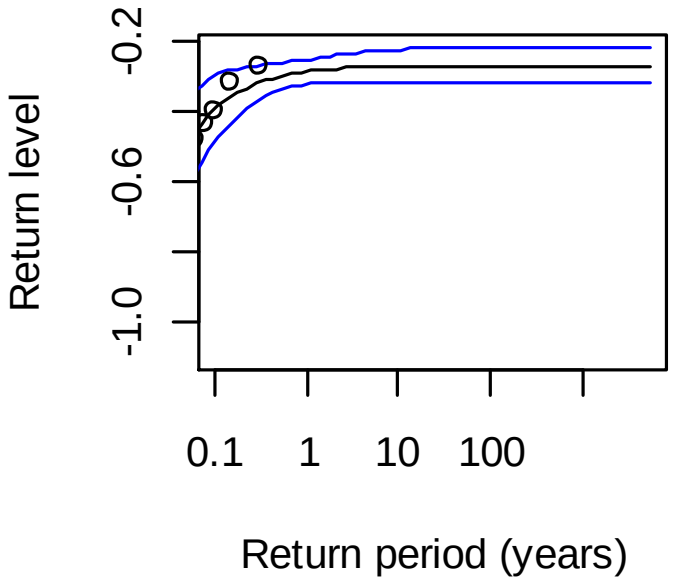
- 极大似然估计法
- 贝叶斯估计法

3.2 超阈值极值模型

- **超阈值模型检验**
  - PP图
  - QQ图
  - 重现水平图







3.3 非独立同分布极值处理

- 非独立极值分析

非独立序列的广义化描述是平稳序列。

平稳序列是指样本变量之间可能是相关的，且相关性存在多种形式，其中一种最基本的假设是极值事件不存在长相关性，即仅相隔时间较短的序列之间存在相关性，而相隔时间较长的序列可以看做是近似独立的。

3.3 非独立同分布极值处理

➢ 非独立区组极值分析

设$X_1, X_2, \ldots, X_n$是一个平稳序列，$\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_n$是与$X_1, X_2, \ldots, X_n$具有相同分布的独立序列，令$M_n = \max\{X_1, X_2, \ldots, X_n\}$以及$\tilde{M}_n = \max\{\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_n\}$。通过适当的规范化处理，引入常数列$\{a_n>0\}$和$\{b_n\}$，使得当$n\to\infty$时，有

$$\Pr\left\{\frac{\tilde{M}_n - b_n}{a_n} \le \mathrm{z}\right\} \to G_1(z)$$

其中$G_1(\mathrm{z})$为非退化分布函数，当且仅当

$$\Pr\{(M_n - b_n)/a_n \le \mathrm{z}\} \to G_2(z)$$

且

$$G_2(z) = G_1^{\tau}(z)$$

式中，$0<\tau\le 1$是极值指标，对于独立序列$\tau=1$。

3.3 非独立同分布极值处理

➢ 非独立区组极值分析

上述推导过程说明，短时相关性对区组极值极限分布的影响仅相当于用G_1^τ代替G_1。如果G_1表示的是以(μ, σ, ξ)且$\xi \neq 0$为参数的广义极值分布，那么

$$G_1^\tau(z) = \exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}^\tau = \exp\left\{-\left[1+\xi\left(\frac{z-\mu^*}{\sigma^*}\right)\right]^{-1/\xi}\right\}$$

式中，$\mu^* = \mu - \frac{\sigma}{\xi}(1-\tau^{-\xi})$；$\sigma^* = \sigma\tau^\xi$。

3.3 非独立同分布极值处理

➢ 非独立超阈值极值分析

超阈值极值理论表明只有独立的超阈值极值才服从广义帕累托分布，因此最直接的方法就是剔除具有相关性的极值样本，从而得到一个近似独立的极值集合。最常用的一种消除极值样本短时相关性的方法是分块-除串法，其基本步骤如下：

①根据某一准则将超阈值极值划分为若干个块；

②找出每个块的最大值；

③以每个块的最大值为极值(或剔除块最大值两侧若干个极值样本并将剩余的样本作为极值)，并假设这些极值服从广义帕累托分布；

④用广义帕累托分布进行极值拟合。

3.3 非独立同分布极值处理

➢ 非平稳极值分析

非平稳序列是指该序列的统计特性(如均值、方差等)随时间的改变而改变，而其变化多受到外部因素的影响。这里将可能与非平稳序列的极值行为相关的变量称为协变量。为了将协变量的影响融入到建模过程中，将极值分布(广义极值分布或广义帕累托分布)的参数表示为如下形式：

$$\omega = h\left(\mathbf{X}^{\mathbf{T}}\boldsymbol{\beta}\right)$$

式中，ω代表的是极值分布模型参数μ，σ和ξ中的任意一个；h是一个特定的函数；$\boldsymbol{\beta}$是参数向量；$\mathbf{X}$是协变量矩阵。

3.3 非独立同分布极值处理

➢ 非平稳极值分析

在本文的研究中，称h为关联函数。比如，广义极值分布位置参数μ的时间线性趋势可以通过恒等关联函数表示，即：

$$\mu = [1\ t]\begin{bmatrix}\beta_0\\ \beta_1\end{bmatrix}$$

同样地，尺度参数σ的指数关联函数为：

$$\sigma = \exp\left([1\ X_1\ X_2]\begin{bmatrix}\beta_0\\ \beta_1\\ \beta_2\end{bmatrix}\right)$$

3.3 非独立同分布极值处理

➢ 非平稳极值分析

模型比选：一种较为简单的方法是基于模型似然度偏差的统计分析。假设有两个模型$M_0 \subset M_1$，两者的偏差统计量为

$$D = 2\{l_1(M_1) - l_0(M_0)\}$$

式中，$l_0(M_0)$和$l_1(M_1)$是模型M_0和M_1的对数似然度。如果D值较大，则说明模型M_1能够比模型M_0更好的反映样本总体。

严格的判断依据如下：在α显著性水平检验下，如果$D>c_\alpha$那么就拒绝模型M_0。其中，c_α是在k自由度下卡方分布的$(1-\alpha)$分为数，k是模型M_1和模型M_0的变量个数之差。

3.4 多元极值模型

➢ 分量最大值模型

设二元随机向量$(X,\ Y)$的分布函数为$F(x,\ y)$，$F \in \mathrm{MDA}(H)$，有标准Fréchet边缘分布，$(X_1, Y_1), \ldots, (X_n, Y_n)$是$(X,\ Y)$的独立同分布样本。记$M_{n,x} = \max\limits_{1 \le i \le n}\{X_i\}$和$M_{n,y} = \max\limits_{1 \le i \le n}\{Y_i\}$，取规范化常数$a_{n,x} = a_{n,y} = n$和$b_{n,x} = b_{n,y} = 0$，得到规范化向量$M_n^* = (M_{n,x}/n, M_{n,y}/n)$。如果当$n \to \infty$时，

$$\Pr\left\{\frac{M_{n,x}}{n} \le x, \frac{M_{n,y}}{n} \le y\right\} \to H(x,y) \tag{1}$$

H是非退化函数，则H具有如下形式：

$$H(x,y) = \exp\{-V(x,y)\}, x > 0, y > 0 \tag{2}$$

3.4 多元极值模型

- 分量最大值模型

其中，

$$V(x,y)=2\int_0^1 \max\left(\frac{\omega}{x},\frac{1-\omega}{y}\right)dK(\omega) \tag{3}$$

且$K(\omega)$是$[0,1]$上均值为1/2的分布函数，满足

$$\int_0^1 \omega\, d\, K(\omega)=1/2 \tag{4}$$

式(1)中的分布函数$H(x,y)$称为二元极值分布，其与区间$[0,1]$上满足式(4)的分布函数K一一对应。由于K不具有有限参数形式，因此二元极值分布也不能用有限参数形式表示。另外，K可以可微，也可以不可微。当K可微时，设其密度为k，有$V(x,y)=2\int_0^1 \max\left(\frac{\omega}{x},\frac{1-\omega}{y}\right)k(\omega)d\omega$。