190-极值分布与广义极值分布（GEV）

在之前的文章《离散选择模型的核心（二项 Logit 模型）》、《多项 Logit 模型（Multinomial Logit, MNL）》以及《Nested Logit 模型》中，我们提到过，Logit 模型的一个重要假设就是：随机项服从极值分布。

实际上，基于**广义极值分布**还可以推导出一系列的 Logit 模型，比如 Paired Combinatorial Logit (PCL)、Generalized Nested Logit (GNL)，等等。

今天咱们就聊一聊 Logit 模型中用到的**极值分布**和**广义极值（Generalized Extreme Value，GEV）分布**。

本文主要参考了史道济教授所著的《实用极值统计方法》。*关注微信公众号“DCM 笔记”，后台回复关键词“极值”，即可获得获取 PDF 版《实用极值统计方法》一书的方法。*

什么是极值分布

我们一般接触到的比较多的分布包括二项分布、正态分布、指数分布等等，很少用到**极值分布**。

在研究极值之前，首先要搞清楚一个最根本的问题：**什么是极值**。

从概率意义上讲，极值表示**随机变量**出现**极端状况**的情形。比如百年一遇的洪水、千年一遇的暴雨，等等。从统计意义上，**极值**是指数据集合中的最大值和最小值 [1]。

极值的特点是**低频高损**——极值虽然发生的概率小，但是一旦发生，造成的损失重大。

我们采用如下定义：

设 $X_1, X_2, ..., X_n$ 是独立同分布的随机变量，且都服从相同的分布函数 $F(x)$；对自然数 n，$M_n = max\{X_1, X_2, ..., X_n\}$ 表示 n 个随机变量的**最大值**，$m_n = min\{X_1, X_2, ..., X_n\}$ 表示 n 个随机变量的**最小值**。

需要理解的是：这里的最大值 M_n 和最小值 n_n 是变量，而不是定值。

举个例子：用随机变量 X_i 表示某地每一个小时内的降雨量，$X_1, X_2, ..., X_{8760}$ 表示一年内 8760 个小时的降雨量（$n = 24 \times 365 = 8760$）；则 $M_{8760} = max\{X_1, X_2, ..., X_{8760}\}$ 表示一年中的最大降雨量。显然，每一年的最大降雨量可能都不一样（比如，1997 年的最大降雨量并不等同于 1998 年的最大降雨量）。亦即 M_{8760} 也是一个随机变量，而不是一个固定值。

极值分布描述是随机变量 M_n 和 m_n 的分布。随机变量 $X_i(i = 1, 2, ..., n)$ 所服从的分布 $F(x)$ 称为**底分布**。

极值分布的 3 种类型

用 $G(x)$ 表示随机变量 M_n 的**分布函数**，即：

$$G(x) = P\{M_n \le x\}$$

$G(x)$ 的函数形式有三种，分别称为 **I 型分布、II 型分布、III 型分布**。

I 型分布： $H_1(x;\mu,\sigma)=e^{-e^{-\frac{x-\mu}{\sigma}}},\quad -\infty<x<\infty$ Gumbel 分布

II 型分布： $H_2(x;\mu,\sigma,\alpha)=\begin{cases}0 & x\le\mu,\\ e^{-\left(\frac{x-\mu}{\sigma}\right)^{-\alpha}} & x>\mu,\end{cases}\quad \alpha>0$ Fréchet 分布

III 型分布： $H_3(x;\mu,\sigma,\alpha)=\begin{cases}e^{-\left(-\frac{x-\mu}{\sigma}\right)^{\alpha}} & x\le\mu,\\ 1 & x>\mu,\end{cases}\quad \alpha>0$ Weibull 分布

其中， μ 称为位置参数（location parameter）， σ 称为尺度参数（scale parameter）。

Ⅰ 型分布又称为 Gumbel 分布， Ⅱ 型分布称为 Fréchet 分布， Ⅲ 型分布称为 Weibull 分布。

极值分布函数 $G(x)$ 看起来总是让人有点望而生畏，可能是因为在指（幂）数函数中又**嵌套**了指（幂）数函数的原因。但实际上了，指数函数有着非常好的求导、积分运算性质。比如 e^x 的导数、积分都是它自身。这可以大大方便运算。

上述三种分布代表了不同的极值行为 [1]，它们统称为**极值分布**（Extreme Value Distribution）。

从模型表达式的角度来看，三种极值分布类型 $H_1(x;\mu,\sigma)$ 、 $H_2(x;\mu,\sigma,\alpha)$ 、 $H_3(x;\mu,\sigma,\alpha)$ 完全不同；但是，从数学的角度来看，我们可以用统一的形式来表示：

$$H(x;\mu,\sigma,\xi)=exp\left\{-\left(1+\xi\frac{x-\mu}{\sigma}\right)^{-\frac{1}{\xi}}\right\},\quad 1+\xi\frac{x-\mu}{\sigma}>0$$

其中，参数 $\mu,\xi\in R$ ，参数 $\sigma>0$ 。**ξ 称为形状参数（shape parameter）**，当 ξ 取不同的值时对应不同类型的分布：

- 当 $\xi=0$ 时，它表示极值 Ⅰ 型分布。
- 当 $\xi>0$ 时，令 $\alpha=\frac{1}{\xi}$ ，则 $H(x;\mu,\sigma,\xi)$ 表示极值 Ⅱ 型分布，其位置参数和尺度参数分别为 $\mu-\alpha\sigma$ 和 $\alpha\sigma$ 。 $H_2(x;\mu,\sigma,\alpha)$
- 当 $\xi<0$ 时，令 $\alpha=-\frac{1}{\xi}$ ，则 $H(x;\mu,\sigma,\xi)$ 表示极值 Ⅲ 型分布，其位置参数和尺度参数分别为 $\mu+\alpha\sigma$ 和 $\alpha\sigma$ 。 $H_3(x;\mu,\sigma,\alpha)$

以 $\xi=0$ 为例，下面验证当 $\xi\to0$ 时， $H(x;\mu,\sigma,\xi)$ 是否会变成 Ⅰ 型分布。

先用**换元**对 $H(x;\mu,\sigma,\xi)$ 简化处理。记 $Y=\frac{x-\mu}{\sigma}$ ，则

$$H(x;\mu,\sigma,\xi)=exp\left\{-(1+\xi Y)^{-\frac{1}{\xi}}\right\}$$

$$=exp\left\{-(1+\xi Y)^{\frac{1}{\xi Y}(-Y)}\right\}=exp\left\{-\left[(1+\xi Y)^{\frac{1}{\xi Y}}\right]^{-Y}\right\}$$

因为

$$\lim_{t\to0}(1+t)^{\frac{1}{t}}=e$$

所以

$$\lim_{\xi \to 0} H(x; \mu, \sigma, \xi) = \lim_{\xi \to 0} exp\left\{-\left[(1+\xi Y)^{\frac{1}{\xi Y}}\right]^{-Y}\right\}$$

$$= exp\left\{-\left[\lim_{\xi \to 0}(1+\xi Y)^{\frac{1}{\xi Y}}\right]^{-Y}\right\} = exp\left\{-e^{-Y}\right\}$$

$$= exp\left\{-e^{-\frac{x-\mu}{\sigma}}\right\} = H_1(x; \mu, \sigma)$$

$H(x; \mu, \sigma, \xi)$ 即为**广义极值分布（Generalized Extreme Value）**，简称为 **GEV 分布。**

Gumbel 分布

离散选择模型中用的比较多的是 **I 型极大值的分布**（即 Gumbel 分布）：

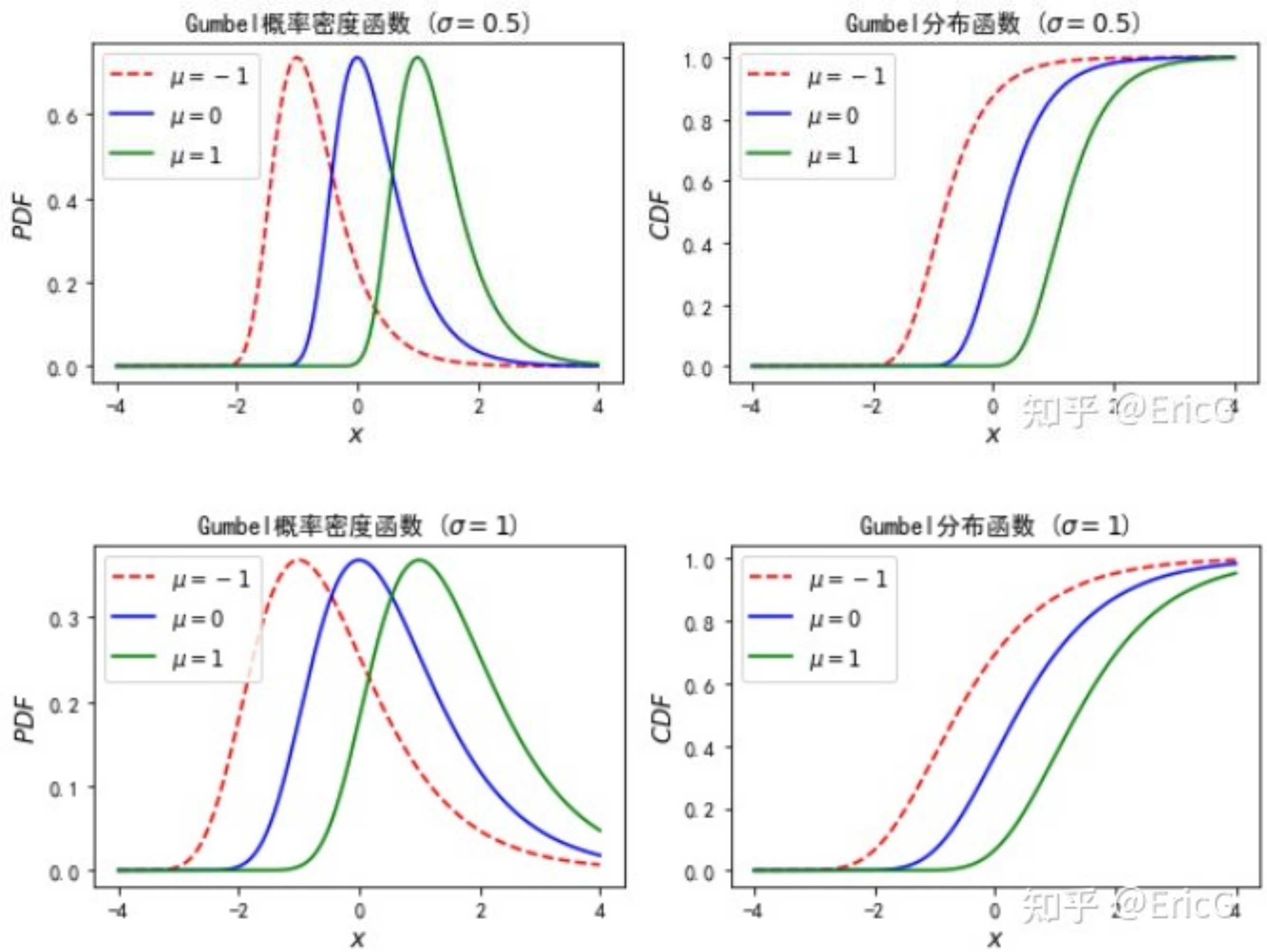
$H_1(x; \mu, \sigma) = e^{-e^{-\frac{x-\mu}{\sigma}}}, \quad -\infty < x < \infty$

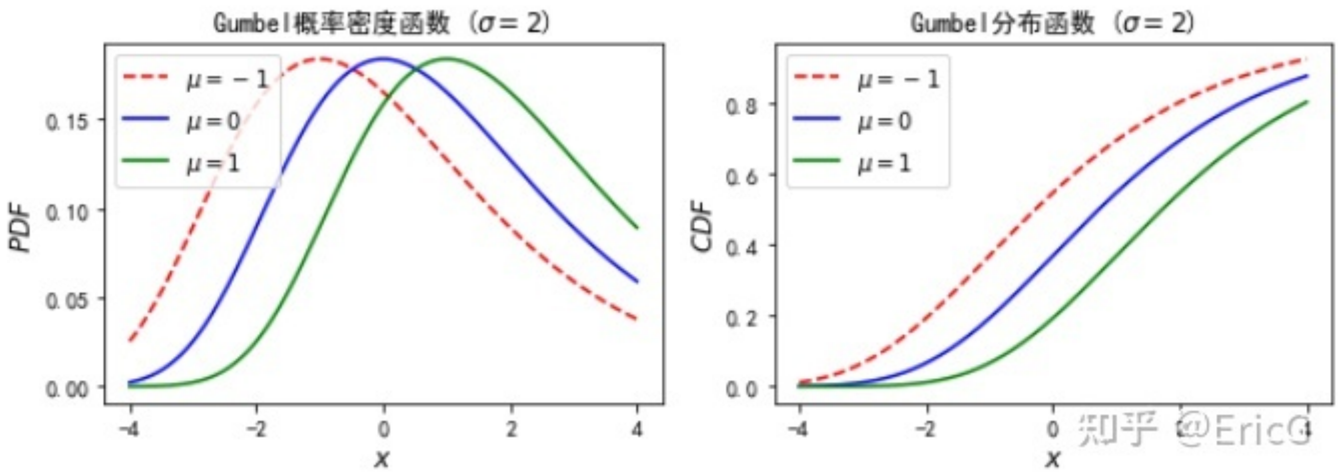
相应的概率密度函数为：

$$\text{pdf:} \quad h_1(x; \mu, \sigma) = H_1^{'}(x; \mu, \sigma) = \frac{1}{\sigma} \cdot e^{-\frac{x-\mu}{\sigma}} \cdot e^{-e^{-\frac{x-\mu}{\sigma}}}$$

$$= \frac{1}{\sigma} \cdot e^{-\frac{x-\mu}{\sigma}} \cdot H_1(x; \mu, \sigma) \quad = \frac{1}{\delta} \cdot e^{-\frac{x-\mu}{\delta}} \cdot \text{CDF}$$

下图展示了当位置参数 μ 、尺度参数 σ 取不同的值的时候，Gumbel 分布的概率密度函数和分布函数。





需要注意是，μ 和 σ 并不等于 Gumbel 分布的**均值**及**标准差**；实际上，对于 Gumbel 分布，它的均值及方差分别为：

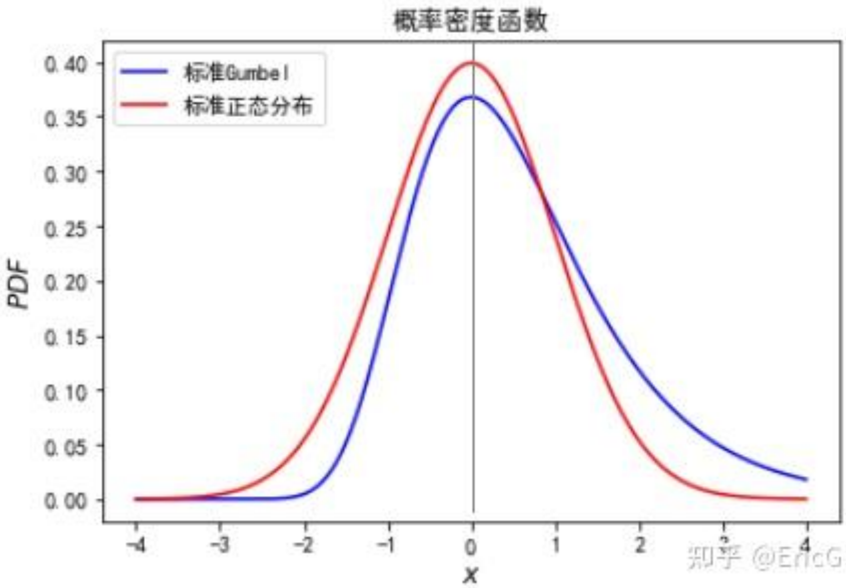
$$\begin{cases} E\left(M_n\right)=\mu+\gamma\sigma \\ Var\left(M_n\right)=\dfrac{\pi^2}{6}\sigma^2 \end{cases}$$

其中，γ 为欧拉常数；$\gamma\approx0.57721$。

当 $\mu=0$ 、$\sigma=1$ 时，$H_1\left(x,0,1\right)=e^{-e^{-x}}$ 又称为**标准 I 型极大值分布**；它的均值及方差分别为：

$$\begin{cases} E\left(M_n\right)=\gamma\approx0.57721 \\ Var\left(M_n\right)=\dfrac{\pi^2}{6}\approx1.645 \end{cases}$$

下图是**标准 I 型极大值分布**和**标准正态分布**之间的对比。从图中可以看出，跟正态分布不同，Ⅰ 型极大值分布为一**偏态分布**。



小结

极值理论博大深奥，本文仅讨论了极值分布的一些**基本知识**以及和离散选择模型相关的应用。有兴趣想要更深入地研究极值理论的朋友可以参考史道济教授的《实用极值统计方法》一书。

关注微信公众号“DCM 笔记”，后台回复关键词“极值”，即可获得获取 PDF 版《实用极值统计方法》一书的方法。

参考文献：

[1] 史道济，《实用极值统计方法》