DOI: 10.13718/j.cnki.xdzk.2017.01.015

广义误差-帕累托分布及其在保险中的应用①

马　跃，　彭作祥

西南大学 数学与统计学院，重庆 400715

摘要：使用 LogGED-Pareto 分布模型拟合丹麦火灾保险损失数据. 结果表明 LogGED-Pareto 模型拟合结果优于使用 Cooray 和 Ananda 提出的 Lognormal-Pareto 模型.

关　键　词：LogGED-Pareto 分布；丹麦火灾保险损失数据；极大似然估计；分布拟合

中图分类号：O212.2　　**文献标志码**：A　　**文章编号**：1673-9868(2017)01-0099-04

在精算和保险行业，巨额风险的存在导致赔付分布具有厚尾现象. 使用对数正态与帕累托或 Burr 分布构成的分布模型拟合丹麦火灾保险损失的研究可以见参考文献[1－3]. 该损失数据记录因火灾造成的包括建筑物、家具和个人财产以及利润损失的总和. 实证结果显示，对数正态与 Burr 分布构成的分布拟合效果相对较优[3]. 对丹麦火灾保险损失的相关研究见参考文献[5－6].

本文使用含对数正态分布的对数广义误差分布与帕累托分布的混合拟合丹麦火灾保险损失数据. 下面给出对数广义误差分布-帕累托分布的定义. 随机变量 X 的密度函数具有如下形式：

$$f(x)=\begin{cases} c\cdot\dfrac{\upsilon \mathrm{e}^{-\frac{1}{2}\left|\frac{\log x-\mu}{\sigma}\right|^{\upsilon}}}{2^{1+\frac{1}{\upsilon}}\Gamma\left(\dfrac{1}{\upsilon}\right)x\sigma} & 0<x\leqslant\theta \\ c\cdot\dfrac{\alpha\theta^{\alpha}}{x^{\alpha+1}} & \theta<x<\infty \end{cases} \tag{1}$$

其中：$c>0,\upsilon>0,\sigma>0,\alpha>0,\theta>0$，则称该随机变量服从对数广义误差-帕累托分布，记为 LogGED-Pareto 分布.

若在门限 θ 处 $f(x)$ 可微，则下面的结论成立：

(i) $\upsilon\neq 1$ 且 $\log\theta>\mu$；

(ii) $\dfrac{\mathrm{e}^{-\frac{1}{2}\left(\frac{\log\theta-\mu}{\sigma}\right)^{\upsilon-1}}}{2^{\frac{1}{\upsilon}}\Gamma(\upsilon)}=\left(\dfrac{\log\theta-\mu}{\sigma}\right)^{\upsilon-1}$；

(iii) 记 $\dfrac{\log\theta-\mu}{\sigma}=k$，那么

$$c=\frac{2\Gamma\left(\dfrac{1}{\upsilon}\right)}{\displaystyle\int_0^{\frac{k^{\upsilon}}{2}}\mathrm{e}^{-x}x^{\frac{1}{\upsilon}-1}\mathrm{d}x+3\Gamma\left(\dfrac{1}{\upsilon}\right)}$$

① 收稿日期：2015-07-18
基金项目：国家自然科学基金项目(11171275)；重庆市自然科学基金项目(cstc2012jjA00029).
作者简介：马　跃(1990-)，四川苍溪人，硕士研究生，主要从事极值统计分析的研究.
通信作者：彭作祥，教授.

$$\mu=\log\theta-\frac{\upsilon}{2\alpha}k^{\upsilon}\qquad\sigma=\frac{\upsilon}{2\alpha}k^{\upsilon-1}$$

此时(1)式可以重新参数化为

$$f(x)=\begin{cases}\dfrac{2\alpha\mathrm{e}^{-\frac{1}{2}\left|\frac{2\alpha}{\upsilon k^{\upsilon-1}}\log\frac{x}{\theta}+k\right|^{\upsilon}}}{\left(3\Gamma\left(\dfrac{1}{\upsilon}\right)+\displaystyle\int_0^{\frac{k^{\upsilon}}{2}}\mathrm{e}^{-x}x^{\frac{1}{\upsilon}-1}\mathrm{d}x\right)\cdot 2^{\frac{1}{\upsilon}}k^{\upsilon-1}x} & 0<x\leqslant\theta\\ \dfrac{2\alpha\theta^{\alpha}\Gamma\left(\dfrac{1}{\upsilon}\right)}{\left(3\Gamma\left(\dfrac{1}{\upsilon}\right)+\displaystyle\int_0^{\frac{k^{\upsilon}}{2}}\mathrm{e}^{-x}x^{\frac{1}{\upsilon}-1}\mathrm{d}x\right)x^{\alpha+1}} & \theta<x<\infty\end{cases}\tag{2}$$

其中 k 满足

$$\frac{\mathrm{e}^{-\frac{1}{2}k^{\upsilon-1}}}{2^{\frac{1}{\upsilon}}\Gamma(\upsilon)}=k^{\upsilon-1}$$

记(1)式对应的分布函数为 $F(x)$，显然有

$$F(x)=\begin{cases}\dfrac{\displaystyle\int_{\frac{(\mu-\log x)^{\upsilon}}{2\sigma^{\upsilon}}}^{\infty}\mathrm{e}^{-t}t^{\frac{1}{\upsilon}-1}\mathrm{d}t}{3\Gamma\left(\dfrac{1}{\upsilon}\right)+\displaystyle\int_0^{\frac{k^{\upsilon}}{2}}\mathrm{e}^{-t}t^{\frac{1}{\upsilon}-1}\mathrm{d}t} & \log x\leqslant\mu,\ 0<x\leqslant\theta\\ \dfrac{\Gamma\left(\dfrac{1}{\upsilon}\right)+\displaystyle\int_0^{\frac{(\log x-\mu)^{\upsilon}}{2\sigma^{\upsilon}}}\mathrm{e}^{-t}t^{\frac{1}{\upsilon}-1}\mathrm{d}t}{3\Gamma\left(\dfrac{1}{\upsilon}\right)+\displaystyle\int_0^{\frac{k^{\upsilon}}{2}}\mathrm{e}^{-t}t^{\frac{1}{\upsilon}-1}\mathrm{d}t} & \log x>\mu,\ 0<x\leqslant\theta\\ \dfrac{\left(3\Gamma\left(\dfrac{1}{\upsilon}\right)+\displaystyle\int_0^{\frac{k^{\upsilon}}{2}}\mathrm{e}^{-t}t^{\frac{1}{\upsilon}-1}\mathrm{d}t\right)x^{\alpha}-2\Gamma\left(\dfrac{1}{\upsilon}\right)\theta^{\alpha}}{\left(3\Gamma\left(\dfrac{1}{\upsilon}\right)+\displaystyle\int_0^{\frac{k^{\upsilon}}{2}}\mathrm{e}^{-t}t^{\frac{1}{\upsilon}-1}\mathrm{d}t\right)x^{\alpha}} & x>\theta\end{cases}\tag{3}$$

1 参数估计

不失一般性，设来自总体的样本值满足 $x_1\leqslant x_2\cdots\leqslant x_n$. 记 θ 的取值在第 m 个样本值和第 $m+1$ 样本值之间，即

$$x_m\leqslant\theta<x_{m+1}$$

那么样本个数为 n 的似然函数表达式为

$$L(\upsilon,\theta,\alpha)=\frac{(2\alpha)^n\theta^{(n-m)\alpha}\left(\Gamma\left(\dfrac{1}{\upsilon}\right)\right)^{n-m}\mathrm{e}^{-\frac{1}{2}\sum_{i=1}^{m}\left|\frac{2\alpha}{\upsilon k^{\upsilon-1}}\log\frac{x_i}{\theta}+k\right|^{\upsilon}}}{\left(3\Gamma\left(\dfrac{1}{\upsilon}\right)+\displaystyle\int_0^{\frac{k^{\upsilon}}{2}}\mathrm{e}^{-x}x^{\frac{1}{\upsilon}-1}\mathrm{d}x\right)^n\cdot 2^{\frac{m}{\upsilon}}k^{m(\upsilon-1)}\cdot\prod_{i=1}^{n}x_i\cdot\prod_{i=m}^{n}x_i{}^{\alpha}}\tag{4}$$

使用极大似然方法估计参数. 一种方法是对每一个固定的 υ 值，找到使 $L(\theta,\alpha)$ 最大化的 α 和 θ 的值. 通过改变 υ，最终找到使 $L(\theta,\alpha,\upsilon)$ 最大化的 υ，α 和 θ 的估计值. 另一种方法是先估计 α[4]，并得到门限 θ 的范围；再使用变动 υ，估计 θ，使得 $L(\theta,\hat{\alpha},\hat{\upsilon})$ 最大.

估计 α 的算法[4] 如下：

1) 设 m 为超过阈值 θ 的个数，记 $k_0^{*}=[2m^{\frac{2}{3}}]$；

2）令 $\hat{\gamma}_n^H(k_0^*, m)=\frac{1}{k_0^*}\sum_{i=0}^{k_0^*-1}\log\frac{X_{(n-i,n)}-X_{(n-m,n)}}{X_{(n-k_0^*,n)}-X_{(n-m,n)}}$；

3）计算 k_0 的最优估计 $\hat{k}_0=\left(\frac{(1+\hat{\gamma}_n^H(k_0^*, m))^2}{2\hat{\gamma}_n^H(k_0^*, m)}\right)^{\frac{1}{2\hat{\gamma}_n^H(k_0^*, m)}}\cdot k^{\frac{2\hat{\gamma}_n^H(k_0^*, m)}{2\hat{\gamma}_n^H(k_0^*, m)+1}}$；

4）令 $\hat{\gamma}_n^H(\hat{k}_0, m)=\frac{1}{\hat{k}_0}\sum_{i=0}^{\hat{k}_0-1}\log\frac{X_{(n-i,n)}-X_{(n-m,n)}}{X_{(n-\hat{k}_0,n)}-X_{(n-m,n)}}$；

5）修正 $\hat{\gamma}_n^H(\hat{k}_0, m)=\hat{\gamma}_n^H(\hat{k}_0, m)-\sqrt{\frac{\hat{\gamma}_n^H(\hat{k}_0, m)}{2\hat{k}_0}}$，得 $\hat{\alpha}^H=\frac{1}{\hat{\gamma}_n^H(\hat{k}, m)}$.

2　丹麦火灾保险损失数据分析

对丹麦火灾保险损失数据的拟合，使用 R 软件的宏包 Stats 中函数 nlm 完成极大似然估计. 对 α 的估计，使用上述算法完成 Hill 型估计. 估计结果见表 1. 虽然参数估计值差异不大，但考虑后尾性，使用先估计尾部的方法应更可取一些.

表 1　拟合丹麦火灾保险损失数据模型的参数估计值

估计法	$\hat{\upsilon}$	$\hat{\alpha}$	$\hat{\theta}$	对数似然函数值
Hill 型估计	2.411	$\hat{\alpha}_{\mathrm{H}}=1.334\,947$	1.436 042	−3 875.245 298
极大似然估计	2.316 056	$\hat{\alpha}_{\mathrm{ML}}=1.403\,441$	1.409 483	−3 872.073 628

下面进行 LogGED-Pareto 与 Lognormal-Pareto 的拟合比较. 由于 Lognormal-Pareto 为 LogGED-Pareto 的特殊情况，后者对丹麦火灾保险损失数据的拟合应比前者更精确. 可以通过使用皮尔逊的 χ^2 检验和 Kolmogorov 与 Smirnov 的 K-S 检验进行验证. 两检验统计量的定义如下：

假设样本量为 n 的样本观测值可以分成 k 类，各自出现的频数分别为 $n_1,n_2,\cdots,n_k$，且 $\sum_{i=1}^{k}n_i=1$，p_i 为拟合分布在 i 类上的概率值，$i=1,2,\cdots,k$，则 χ^2 检验统计量为：

$$\chi^2=\sum_{i=1}^{k}\frac{(n_i-np_i)^2}{np_i} \tag{5}$$

在显著水平为 p 的情形下，若 $\chi^2<\chi_p^2(k-1-s)$，则拟合分布通过 χ^2 检验，其中 s 为已估参数个数.

记样本量为 n 的样本观测值对应的经验分布函数 $F_n(x)$，拟合分布函数为 $F(x)$，则单样本 K-S 检验的统计量为：

$$D_n=\max_{x\in\mathbb{R}}|F(x)-F_n(x)| \tag{6}$$

在显著水平为 p 的情形下，若 $D_n<D(n, p)$，则拟合分布通过 K-S 检验，其中 $D(n, p)$ 为临界值.

使用统计检验方法对两种模型分别进行统计检验，检验结果见表 2. 其中，在计算卡方值时将丹麦火灾保险数据以上限分别为 1.25,1.75,2.25,2.75,3.25,3.75,4.25,4.75,5.25,5.75,6.25，∞ 进行分组[1].

表 2　丹麦火灾保险损失数据模型的参数估计值检验

分　布	参　数　估　计	对数似然函数	K-S	DF	卡方值
Lognormal-Pareto	$\hat{\theta}=1.385\,128$, $\hat{\alpha}=1.436\,332$	−3 877.844 425	0.028 7	9	12.488
LogGED-Pareto(ML)	$\hat{\upsilon}=2.316\,056$, $\hat{\theta}=1.409\,483$, $\hat{\alpha}=1.403\,441$	−3 872.073 628	0.025 8	8	10.428 7
LogGED-Pareto(H)	$\hat{\upsilon}=2.411$, $\hat{\theta}=1.436\,042$, $\hat{\alpha}=1.334\,947$	−3 875.245 298	0.026 5	8	12.951 7

由表 2 知，在 p 值等于 0.05 的情况下，Lognormal-Pareto 与 LogGED-Pareto 均通过拟合分布卡方检验($\chi^2(8)=15.507\,3$，$\chi^2(9)=16.919\,0$). 在 p 值等于 0.05 情况下，Lognormal-Pareto 未通过 K-S 检验($D(n, p)=0.028$). 因此拟合丹麦火灾保险损失数据，使用对数广义误差与帕累托联合分布比使用对数正态与帕累托联合分布更加精确，这与直观的结果是相符的.

参考文献:

[1] COORAY K, ANANDA, M A. Modeling Actuarial Data with a Composite Lognormal-Pareto Model [J]. Scandinavian Actuarial Journal, 2005, 2005(5): 321－334.

[2] SCOLLNIK D. On Composite Lognormal-Pareto Models [J]. Scandinavian Actuarial Journal, 2007, 2007(1): 20－33.

[3] NADARAJAH S, BAKAR S. New Composite Models for the Danish Fire Insurance Data [J]. Scandinavian Actuarial Journal, 2012, 2014(2): 1－8.

[4] ALVES M I F. A Location Invariant Hill-type Estimator [J]. Extremes, 2001, 4(3): 199－217.

[5] RESNICK S I. Discussion of the Danish Data on Large Fire Insurance Losses [J]. Astin Bulletin, 1997, 27(1): 139－151.

[6] PIGEON M, DENUIT M. Composite Lognormal-Pareto Model with Random Threshold [J]. Scandinavian Actuarial Journal, 2011, 2011(3): 177－192.

[7] VASUDEVAY R, KUMARI J V. On General Error Distributions [J]. ProbStat Forum, 2013, 6(10): 89－95.

Modeling Actuarial Data with LogGED-Pareto Model

MA Yue, PENG Zuo-xiang

School of Mathematics and Statistics, SouthwestUniversity, Chongqing 400715, China

Abstract: In this paper, we use LogGED-Pareto model to fit the Danish fire insurance data. The results show that the logGED-Pareto model is better than the Lognormal-Pareto model proposed by Cooray and Ananda.

Key words: LogGED-Pareto model; Danish fire insurancedata; maximum likelihood estimation; goodness-of-fit

责任编辑 张 枸