Statistics and Application 统计学与应用, 2022, 11(1), 1-9
Published Online February 2022 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2022.111001

Hans汉斯

偏正态与偏*t*分布对巨灾损失数据的拟合分析

徐　鹏

南京邮电大学理学院，江苏　南京

收稿日期：2021年11月15日；录用日期：2021年12月29日；发布日期：2022年1月14日

摘　要

如何更好地拟合巨灾损失数据一直是一个重要的问题。本文将偏正态分布和偏*t*分布应用到实际巨灾损失数据中，选取了福州市近8年火灾造成的经济损失为样本数据，采用极大似然估计的方法对参数进行估计，并且选取常用来拟合损失的分布进行比较。结果表明，偏*t*分布最适合拟合巨灾损失数据。

关键词

偏正态分布，偏*t*分布，极大似然估计，贝叶斯信息准则

The Fitting Analysis of Skew-Normal and Skew-*t* Distribution to the Catastrophe Loss Data

Peng Xu

School of Science, Nanjing University of Posts and Telecommunications, Nanjing Jiangsu

Received: Nov. 15th, 2021; accepted: Dec. 29th, 2021; published: Jan. 14th, 2022

Abstract

How to better fit catastrophe loss data is always an important problem. This paper applies the skew-normal distribution and skew-*t* distribution to the actual catastrophe loss data, and selects the economic loss caused by fire in Fuzhou in recent 8 years as the sample data. Besides, this paper uses the maximum likelihood estimation method to estimate the parameters, and selects the commonly used loss distribution for comparison. The results show that skew *t* distribution is most suitable for fitting catastrophe loss data.

文章引用: 徐鹏. 偏正态与偏 *t* 分布对巨灾损失数据的拟合分析[J]. 统计学与应用, 2022, 11(1): 1-9.
DOI: 10.12677/sa.2022.111001

Keywords

Skew-Normal Distribution, Skew-*t* Distribution, Maximum Likelihood Estimation, Bayes Information Criterion

1. 介绍

1.1. 背景

随着人类活动的不断加剧，全球自然环境也在不断变化。近年来，不同类型的重大自然灾害不断发生，如 2008 年中国汶川 8.0 级大地震、2020 年新型冠状病毒等悲剧，造成了不可估量的人员伤亡和财产损失[1] [2]。从以往灾害损失数据来看，特大灾害破坏力大，发生频率高。此外，地震等巨灾风险事件的发生是典型的“小概率”事件。因此，巨灾造成的风险损失具有“概率小、损失大”的特点。

在大多数情况下，正态分布并不适合于重大灾害经济损失数据的分布模型，因为巨灾损失有偏态的倾向。偏态分布在偏态和峰度方面具有更灵活的优势。此外，偏正态分布和偏 t 分布在统计建模中越来越受欢迎，主要是因为它们易于拟合和解释。

偏正态分布和偏 t 分布具有优美的数学形式和易于处理的特点。偏正态分布和偏 t 分布都是单峰分布。它们比其他分布更集中，因此在拟合单峰和偏置数据方面具有更好的优势。偏正态分布是正态分布的扩展。偏正态分布可以测量数据的浓度特征，显示数据的偏态特征。1908 年 Fernando de Helguero 在他的研究论文中，利用均匀分布对正态分布的密度函数进行扰动，得到了偏正态密度的形式。虽然这种形式不同于标准的偏正态分布数学公式，但它们表达了精确的随机机制。1985 年，Azzalini A 给出了偏正态分布的数学结构和相关性质[3]。构造偏正态分布的目的是提供一种分布类型，它可以具有正态密度、数学可处理性、偏态指数的广泛范围和峰度的“严格包含”。利用均匀分布对正态分布进行扰动，得到了偏正态分布思想的起源。偏 t 分布思想由 t 分布发展到广义 t 分布，最后发展到偏 t 分布。1988 年，McDonald 和 Newey 引入了广义 t 分布的概念来估计回归参数[4]。该文考虑了广义 t 分布的估计量及其相应的性质，以及该估计量对回归估计的渐近效率的影响，并给出了最小距离解释以使回归参数的渐近方差最小。广义 t 分布的使用拓宽了研究人员的思路。此后，许多研究者开始使用广义 t 分布来估计参数。1998 年，西奥多西奥将 t 分布从广义 t 分布扩展到斜 t 分布，并系统地解释了斜 t 分布的数学性质[5]。在论文中，他还给出了具体的应用场景，而偏 t 分布的应用实例包括初步数据统计、最大似然估计、相关财务数据等应用场景。在他的论文中，他指出，经验上，经济领域的数据主要是有偏的，所以在金融领域应用偏 t 分布将显示出良好的发展前景。

在给定的形式下，正态分布可以表示为偏正态分布的一部分。当所构造的偏正态分布的偏态系数为零时，我们恢复了正态分布。正态分布作为一种经典分布，在拟合现实世界中对称且集中的数据方面具有很好的优势。但在现实中，数据往往是有偏的，不能严格满足正态分布极限的要求。Brown N. D.使用偏正态分布拟合智商数据，发现由于在数据选择中忽略了一些变量因素，观察到的数据往往是有偏差的[6]。当使用正态分布来拟合智商数据时，总体平均值被高估了。因此，使用偏正态分布可以得到更好的估计数据。对于单峰和偏态数据，可以较好地估计偏态正态分布。当已知数据有右偏特性时，使用正态

分布进行估计可能会导致均值的高估和标准差的低估。偏正态分布有许多扩展，如多元偏正态分布。总的来说，对于单峰数据具有右偏或左偏的情况，采用偏正态分布拟合模型是可靠的。

除了智商数据，一些研究人员还分析了保险索赔数据。在金融数据中，正态分布有很好的应用，但对于保险索赔，数据往往是有偏见的。一些数据除了偏度的特征外，还可能表现出厚尾的特征。Balance C.等人在建模和评估保险风险数据时，采用了一系列连续分布，包括正态分布、偏正态分布和对数正态分布，对数据进行拟合，以便更好地展示数据分布等信息[7]。

De Leo S. (2021 年) [8]使用偏正态分布拟合新的冠状病毒数据。对于收集的欧洲新冠病毒数据，每百万人确认的数据配有每百万人死亡的数据。通过比较不同偏正态分布的参数，分析各国政府的防疫措施，同时预测疫情的结束时间。在评价欧洲国家时，很难分析表现较好的国家是如何传播病毒的。尽管如此，对该阶段不同国家和地区收集的诊断数据进行探索还是比较容易和有意义的。通过分析数据的分布和参数，可以直观地看到哪些国家的绩效相似，哪些国家的绩效差异较大。当使用数据分布来描述不同国家的性能时，是对于没有达到峰值的数据。偏正态分布可用于估计每百万人中的死亡人数与每百万人中的确诊病例人数的比率。数据的对称性只有在数据达到峰值时才能体现出来。然而，要预测流感的终结，偏正态分布是获得正确答案的基础。

1.2. 偏正态分布

定义：一个随机变量的密度函数为 x

$$f\left(x,\mu,\sigma^2,\lambda\right)=\frac{2}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)\Phi\left(\frac{\lambda\left(x-\mu\right)}{\sigma}\right),x\in R,\mu\in R,\mu>0 \tag{1}$$

其中 $\phi(\ )$ 是标准正态分布的密度函数，$\Phi(\ )$ 是分布函数，记作 $x\sim SN\left(\mu,\sigma,\lambda\right)$ [9]。μ，σ，λ 分别为位置参数、尺度参数、倾斜参数。

1.3. 偏 t 分布

定义：设 Z 和 V 为独立随机变量，V 是服从带参数 $\left(\alpha,\beta\right)$ 的偏正态分布 $SN\left(\alpha,\beta\right)$ 的随机变量，Z 是服从带自由度为 $\left(v/2,v/2\right)$ 的伽马分布 $\Gamma\left(v/2,v/2\right)$ 的随机变量，则称 $V/\sqrt{Z}$ 服从于偏 t 分布，记为 $ST_v\left(\alpha,\beta\right)$，其概率密度函数为

$$f\left(x;\alpha,\beta,v\right)=ct_v\left(x\right)T_{v+1}\left(\alpha x\sqrt{\frac{v+1}{x^2+v}};\beta\right),x\in R \tag{2}$$

其中 $T_{v+1}\left(x;\beta\right)$ 是自由度为 $v+1$、偏度参数为 β 的标准 t 分布的密度函数，c 常数可以表示为

$$c=\left[\frac{1}{2}-\frac{1}{\pi}\arctan\frac{\beta}{\sqrt{1+\alpha^2\left(1+\beta^2\right)}}\right]^{-1} \tag{3}$$

1.4. 目的

本文的目的是利用偏 t 分布、偏正态分布等分布模型对火灾损失数据进行拟合，找到最适合该问题的模型。研究的动机是探索这种统计分布是否能够更好地适合于严重灾害损失数据。本文研究的火灾损失数据是有偏的，高峰度数据以及对数据的偏正态分布和偏 t 分布的拟合是理想的候选分布。主要发现，对于火灾损失数据，偏 t 分布可以很好地拟合数据，而偏正态分布不能很好地拟合火灾数据的偏态特征，对数正态分布和威布尔分布也不能很好地拟合数据的峰度特征。由于这些模型的灵活性、易解释性和易

处理的偏正态分布和偏 t 分布，我们相信这些模型将在现实中显示出良好的应用前景。本文采用偏 t 分布拟合火灾损失数据取得了良好的效果，但利用偏 t 分布分析严重灾害损失数据还需要应用更现实的数据。对于模型参数的估计，本文采用极大似然估计方法进行估计。考虑到分布和数据，这种方法是更好的选择。偏 t 分布和偏正态分布扩大了原有分布的使用范围。贝叶斯方法是一种可以结合先验信息进行分析的方法。对于灾情数据的研究，选择合适的初始分布可以很好地扩展数据的方法。

2. 实例数据分析

火灾每年都给中国许多地区造成严重的破坏。以火灾重灾区福州市为研究对象，收集 2014~2021 年火灾造成的直接经济损失，见表 1。数据来自福州市人民政府官方网站[10]。

Table 1. Economic loss caused by fire in Fuzhou
表 1. 福州市火灾造成的经济损失统计表

日期/年.月	财产损失/万元	日期/年.月	财产损失/万元	日期/年.月	财产损失/万元	日期/年.月	财产损失/万元
2014.01	67.7065	2016.01	192.9	2018.01	45.2	2020.01	97.9
2014.02	30.6110	2016.02	231.7	2018.02	106.3	2020.02	73.1
2014.03	46.9200	2016.03	129.2	2018.03	114.6	2020.03	72.7
2014.04	46.1570	2016.04	136.9	2018.04	67.4	2020.04	134.7
2014.05	28.2150	2016.05	88.5	2018.05	89	2020.05	45.1
2014.06	13.2200	2016.06	116.3	2018.06	51.2	2020.06	170
2014.07	36.9911	2016.07	98.3	2018.07	53.2	2020.07	61.5
2014.08	30.9641	2016.08	128	2018.08	84.6	2020.08	192.9
2014.09	60.4400	2016.09	130.7	2018.09	29.7	2020.09	62.5
2014.10	87.0350	2016.10	81.7	2018.10	95.4	2020.10	159.4
2014.11	63.9380	2016.11	130.9	2018.11	34.5	2020.11	180
2014.12	48.1855	2016.12	387.2	2018.12	146.3	2020.12	105.7
2015.01	92.5292	2017.01	121.4	2019.01	75.3	2021.01	96.4
2015.02	169.7	2017.02	175.3	2019.02	66.6	2021.02	88.7
2015.03	73.87	2017.03	79.2	2019.03	150	2021.03	78.7
2015.04	37.64	2017.04	131.8	2019.04	152.1	2021.04	82.7
2015.05	362	2017.05	100.7	2019.05	80.9	2021.05	57.2
2015.06	37.5	2017.06	200.2	2019.06	40.5	2021.06	86.3
2015.07	44.3	2017.07	125.1	2019.07	115.7	2021.07	120.4
2015.08	37.25	2017.08	67.9	2019.08	159.6	2021.08	110.6
2015.09	134.7	2017.09	80.4	2019.09	137.8	2021.09	79.2
2015.10	112.5	2017.10	141.2	2019.10	99.8		
2015.11	94.9	2017.11	122.7	2019.11	129.1		
2015.12	38.97	2017.12	222.8	2019.12	136		

这些数据的一些描述性统计数据如表 2 所示。通过表 2 可以看出：偏度为 1.927，峰度为 6.260，两个系数均大于 1。同时，我们对数据进行单样本 K-S 检验，检验结果为 $p = 0.009 < 0.05$，所以它不符合正态分布。

Table 2. Describes the statistics for the loss of data caused by fire
表 2. 火灾造成的数据损失统计情况

	火灾造成的经济损失/万元
平均值	103.589
标准偏差	62.341
偏态	1.927
峰度	6.260
最小值	13.220
最大值	387.200

为了直接分析数据，本文给出了数据的直方图。从图 1 可以看出，这些数据在右侧具有长尾的特征，不具有正态分布的特征。所以本文使用偏态分布来处理这些数据。

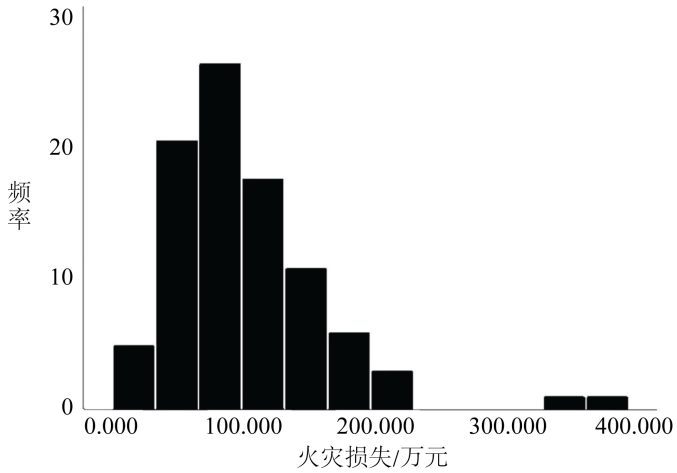

Figure 1. The histogram of the lost data caused by fire
图 1. 火灾造成的经济损失数据直方图

3. 模型参数的估计与比较

正态分布和偏态分布具有较重的数学形式和易于处理的特点。偏正态分布是正态分布的扩展。1908 年 Fernando de Helguero 利用均匀分布对正态分布密度进行扰动，得到一种偏正态密度形式。1985 年，Azzalini A 给出了偏正态分布的数学形式和相应的性质。1988 年，McDonald 和 Newey 引入了分布的概念来估计回归参数。在该文中，作者计算了估计量的广泛分布量、相应的影响以及估计量的渐近效率，最小距离解释了渐近范围。广泛分布的使用扩大了研究人员的问题，许多研究人员致力于估计估计值，开始使用当时的分布图。常用的模型估计方法有极大似然估计、矩估计和贝叶斯估计。根据上述偏正态分布和偏 t 分布的密度函数，本文采用最大似然估计方法估计模型的参数。具体估计结果如表 3 所示。

Table 3. Model parameter estimation results
表 3. 模型参数估计结果

分布函数	参数	估计值
偏正态分布	μ	29.617
	σ	95.643
	λ	7.609
偏 t 分布	α	104.751
	β	73.574
	v	0.7855

本文采用偏正态分布和偏 t 分布拟合数据。将这两种分布的拟合效果与常用的对数正态分布和威布尔分布进行了比较。首先，图 2 显示了所分析的四种分布的数据拟合结果的分布。从这些分析的拟合曲线和实际数据分布来看，偏 t 分布优于其他三种分布。

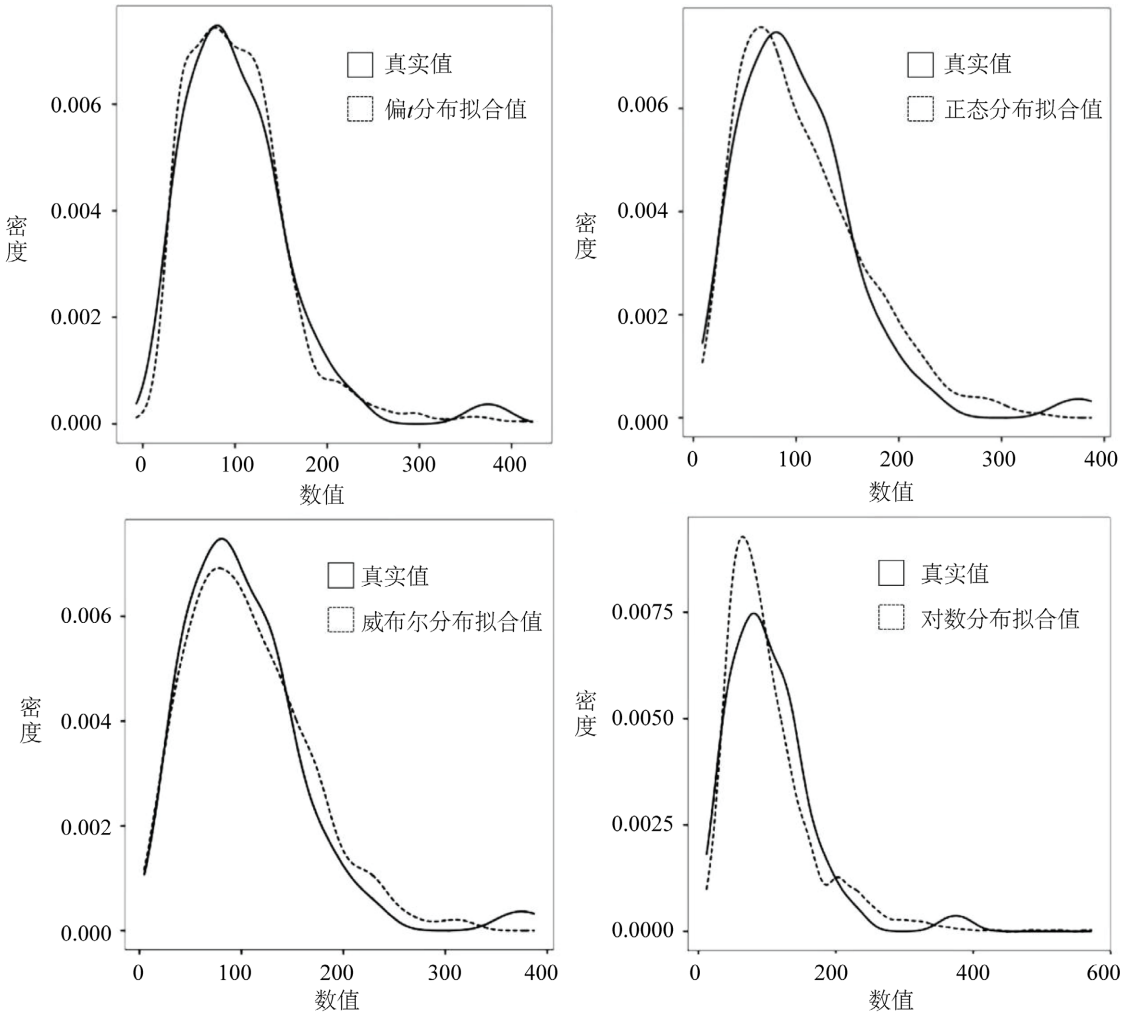

Figure 2. The fitting density curve of each distribution to the loss data caused by fire
图 2. 各分布对火灾损失数据的拟合密度曲线

在数据的显示部分，本文主要显示数据的概率密度图像。在概率密度图像中，一个点从负无穷到某一点所包围的区域构成发生的概率。在图像中，每个单元的面积越大，该部分的数据就越集中。概率密度图像可以直观地显示数据的分布信息。同时还可以从概率密度图像中观察数据的数值特征，如平均值、方差等。概率密度函数的估计有多种核函数可供选择，包括高斯函数、矩形函数、三角函数、余弦函数等。要估计密度函数，R 语言中的 ggplot2 包提供了一种方便的方法。在绘制概率密度图像时，默认采用高斯函数进行估计。具体算法是撅起的质量经验分布函数的正则网格至少 512 点,然后使用快速傅里叶变换与离散卷积这个近似版本的内核,然后使用一个线性近似计算密度在指定点[3]。

上图描绘了四个分布的估计结果，具体的绘制方法是通过最大似然估计法估计给定数据的参数，随机生成 1000 个随机数，绘制得到的理论分布的概率密度函数。将拟合值与实际值绘制在同一张图上，直观地显示模型的估计结果。由上图可以看出，四个图中偏 t 分布的拟合结果是最好的。实际数据围绕拟合曲线波动。对于偏正态分布，模型拟合结果仍有改进空间。从上图可以看出，偏正态分布并不很适合数据的尾部特征。一些数据被高估，而另一部分数据显示出低估的属性。偏 t 分布除了偏态特性外，它还与分布的偏态 t 有关，模型的均值特征拟合不是很令人满意，最大似然估计的均值略小于实际数据的均值。威布尔分布和对数正态分布，为火灾损失数据的峰值。拟合情况不太好，整体适宜条件有较大偏差。不同模型拟合差异的主要原因可能是每个模型的特点不同，每个模型都解释了峰度、偏度等信息。它们的特征是不同的。例如，偏 t 分布可以更好地描述尾部信息，而偏正态分布适用于更稳健的数据。

贝叶斯信息准则和相关系数 R^2 进一步作为选择标准。具体结果如表 4 所示。

Table 4. Fitting results of different distribution functions
表 4. 不同分布函数拟合结果

分布函数	相关系数 R^2	BIC 值
偏正态分布	0.9530	487.9103
偏 t 分布	0.9551	483.4938
对数正态分布	0.9004	557.7275
威布尔分布	0.9502	493.1660

由表 4 可知，偏 t 分布的相关系数 R^2 最接近 1，且偏 t 分布的 BIC 统计量最小，因此可以得出该分布对损失数据拟合效果最好。

对于拟合数据的分布，有更多的分布可供选择，如 beta、柯西、指数等分布。对于有偏数据的拟合，beta 数据也可以完成这一任务。然而，在自然界中，正态分布的频率是很高的，有各种满足正态分布要求的数据。因此，在分析现实世界的数据时，正态分布及其导数分布的概念更有可能适合经验。

除了拟合偏正态分布和偏 t 分布外，本文还拟合了对数正态分布和威布尔分布。当一个数据变量服从正态分布时，对数后的变量服从对数正态分布。对数正态分布也可以解决数据不对称的问题。除了对数正态分布外，本文还拟合了威布尔分布，威布尔分布是可靠性分析和寿命测试的理论基础。这个分布与其他分布密切相关。当其参数 k = 1 时，模型分布为指数分布，当其参数 k = 2 时，模型分布为瑞利分布。分布还可以用于拟合左偏和右偏。对于不同模型的具体拟合，本文使用 R^2、BIC 指标进行估算。R^2 也被称为相关系数，通常在回归模型中用来解释模型的拟合优度。对于本文，对于火灾损失数据，首先，通过最大似然估计法在给定分布下得到模型的参数，从而得到模型的理论分布。然后对采集到的数据计算每个观测值的分位数，并将分位数反求理论分布，得到每个观测值对模型的理论值。残差数据是在每次观测的理论值与实际值不同后得到的。此时可以得到残差平方和和总平方和，并可以计算出最终的平

方和，为相关系数 R^2。R^2 表示基于真实数据估计理想分布的意义，以检验估计分布的拟合能力。本文的结果表明，偏 t 分布的拟合度最好，其次是偏正态分布、威布尔分布，最后是对数正态分布。在具体评价模型拟合结果时，除使用相关系数 R^2 外，本文还使用 BIC 指标进行评价。在模型选择中通常使用 BIC，它通常与 AIC 指标一起使用。主要的区别是 BIC 的惩罚更显著。对于样本容量较大的数据，惩罚更显著，而 AIC 的惩罚固定在一个恒定的水平上。BIC 标准包括索引中包含一层信息。构造的似然函数越大，BIC 值越小，代表的数据越大，说明模型选择得越好。当使用回归分析来分析模型的选择，在确定选择时，必须确保每个模型中的因变量都被选择，这样 BIC 比较结果才有意义。本文利用火灾损失数据拟合的模型也能满足这一点的要求。BIC 值选取的模型结果不能完全代表模型的质量。BIC 价值提供了更多的启发性思考。在防止过拟合的同时计算数据的信息熵，给出了一个惩罚项来更好地比较模型的结果。在评价模型的最终拟合效果时，本文计算了 K-S 检验的 p 值和 AIC 等几个指标，但选择使用 R^2 和 BIC 值进行评估。BIC 指标的计算结果表明，最适合的模型是偏 t 分布，其次是偏正态分布，威布尔，最后是对数正态分布。这个结果与 R^2 选择的结果相同。因此，总体而言，本文的结果表明，对于火灾损失数据的拟合，偏 t 分布更为明显。该分布具有较好的拟合效果。

4. 总结

本文根据巨灾损失数据的特点，介绍了两种新的偏态分布：偏正态分布和偏 t 分布来拟合巨灾风险数据，同时还选取了常用的偏态分布：威布尔分布和对数正态分布进行比较。通过极大似然估计的方法对参数进行估计，用相关系数 R^2 和 BIC 值比较了偏正态分布、偏 t 分布、对数正态分布和威布尔分布的拟合结果。对比分析发现，由于偏态和峰度的限制，偏正态分布的拟合效果并不理想，偏 t 分布可以得到更显著的偏度和峰度，得到更好的相关结果。因此，偏 t 分布适合于本文对巨灾损失数据的处理。不同巨灾数据的具体特征是不同的，也没有一个单一的损失模型适合所有的灾难数据，所以我们必须根据具体情况来考虑。

致谢

此研究由国家自然科学基金项目(31971029)资助。

参考文献

[1] 魏本勇, 苏桂武. 基于投入产出分析的汶川地震灾害间接经济损失评估[J]. 地震地质, 2016, 38(4): 1082-1094.

[2] 唐任伍, 李楚翘, 叶天希. 新冠病毒肺炎疫情对中国经济发展的损害及应对措施[J]. 经济与管理研究, 2020, 41(5): 3-13.

[3] Azzalini, A. (1985) A Class of Distributions Which Includes the Normal Ones. *Scandinavian Journal of Statistics*, **12**, 171-178.

[4] McDonald, J.B. and Newey, W.K. (1988) Partially Adaptive Estimation of Regression Models via the Generalized *t* Distribution. *Econometric Theory*, **4**, 428-457. https://doi.org/10.1017/S0266466600013384

[5] Theodossiou, P. (1998) Financial Data and the Skewed Generalized *t* Distribution. *Management Science*, **44**, 1595-1722. https://doi.org/10.1287/mnsc.44.12.1650

[6] Brown, N.D. (2001) Reliability Studies of the Skew Normal Distribution. Electronic Theses and Dissertations. http://digitalcommons.library.umaine.edu/etd/408

[7] Bolance, C., Guillen, M., Pelican, E. and Vernic, R. (2008) Skewed Bivariate Models and Nonparametric Estimation for the CTE Risk Measure. *Insurance*: *Mathematics and Economics*, **43**, 386-393. https://doi.org/10.1016/j.insmatheco.2008.07.005

[8] De Leo, S. (2021) Impact of COVID-19 Testing Strategies and Lockdowns on Disease Management across Europe, South America, and the United States: Analysis Using Skew-Normal Distributions. *JMIRX Med*, **2**, e21269. https://doi.org/10.2196/21269

[9] Arellano-Valle, R.B., del Pino, G. and San Martin, E. (2002) Definition and Probabilistic Properties of Skew-Distributions. *Statistics & Probability Letters*, **58**, 111-121. https://doi.org/10.1016/S0167-7152(02)00088-3

[10] 中华人民共和国福建省福州市人民政府. 福州市火灾事故统计数据[EB/OL]. http://www.fuzhou.gov.cn/smartSearch/main/search.xhtml?siteId=402849946077df37016077eea95e002f#page=2, 2021-10-12.