

“统计与数据科学丛书”序

统计学是一门集收集、处理、分析与解释量化的数据的科学。统计学也包含了一些实验科学的因素,例如通过设计收集数据的实验方案获取有价值的信息,为提供优化的决策以及推断问题中的因果关系提供依据。

统计学主要起源对国家经济以及人口的描述,那时统计研究基本上是经济学的范畴。之后,因心理学、医学、人体测量学、遗传学和农业的需要逐渐发展壮大,20 世纪上半叶是统计学发展的辉煌时代。世界各国学者在共同努力下,逐渐建立了统计学的框架,并将其发展成为一个成熟的学科。随着科学技术的进步,作为信息处理的重要手段,统计学已经从政府决策机构收集数据的管理工具发展成为各行各业必备的基础知识。

从 20 世纪 60 年代开始,计算机技术的发展给统计学注入了新的发展动力。特别是近二十年来,社会生产活动与科学技术的数字化进程不断加快,人们越来越多地希望能够从大量的数据中总结出一些经验规律,对各行各业的发展提供数据科学的方法论,统计学在其中扮演了越来越重要的角色。从 20 世纪 80 年代开始,科学家就阐明了统计学与数据科学的紧密关系。进入 21 世纪,把统计学扩展到数据计算的前沿领域已经成为当前重要的研究方向。针对这一发展趋势,进一步提高我国的统计学与数据处理的研究水平,应用与数据分析有关的技术和理论服务社会,加快青年人才的培养,是我们当今面临的重要和紧迫的任务。“统计与数据科学丛书”因此应运而生。

这套丛书旨在针对一些重要的统计学及其计算的相关领域与研究方向作较系统的介绍。既阐述该领域的基础知识,又反映其新发展,力求深入浅出,简明扼要,注重创新。丛书面向统计学、计算机科学、管理科学、经济金融等领域的高校师生、科研人员以及实际应用人员,也可以作为大学相关专业的高年级本科生、研究生的教材或参考书。

朱力行

2019 年 11 月

前言

在当今世界,理解和预测极端事件的发生对于风险评估和管理至关重要.本书由王静龙、王银凤、邓文丽、唐炎林、吴贤毅和朱仲义共同撰写,旨在为读者提供一个全面的视角,以理解和分析这些极端事件背后的统计原理.本书详细介绍了重尾分布的理论基础和应用方法,涵盖了从基础定义到理论分析再到统计推断分析技术的各个方面.

本书的内容分为七章,每章都围绕一个核心议题展开,层层递进地展开论述.第1章作为开篇之章,详细介绍了重尾分布的定义,为后续的深入探讨筑牢根基.从第2章开始,每章都以前一章的内容为基石,逐步拓展研究视野,深化对重尾分布的理解与应用.我们希望读者通过阅读本书,不仅能够获得知识,还能够激发对重尾分布的进一步探索和思考.

第1章首先通过多个领域的实例阐述了重尾分布的背景和必要性,以及其定义和等价陈述.这为读者提供了理解后续章节内容的基础.重尾分布是指那些尾部概率衰减速度很慢的分布.它慢于指数分布和正态分布等人们熟悉的分布,所以重尾分布在刻画极端事件时尤为重要.

第2章深入探讨了帕累托分布,这是一种典型的重尾分布,广泛应用于金融、保险和环境科学等领域.本章详细讨论了帕累托分布的尾部尺度不变性、尾部长尾特性以及极端事件突变特性.此外,还介绍了秩图、幂指数的LS估计、极大似然估计、Q-Q图和Q-Q估计、超越和POT方法以及分布检验等分析工具.

第3章以帕累托分布为基础推广到尺度不变分布,这类分布同样表现出尾部长尾和极端事件突变特性.本章讨论了LS估计图、Q-Q估计图、Hill图、风险值VaR和柯尔莫哥洛夫检验,以及PLFit方法等,进一步丰富了重尾分布的分析手段.

第4章以尺度不变分布为基础推广到渐近尺度不变分布,这是一种在尾部行为上具有特定性质的分布.本章详细讨论了渐近尺度不变分布的闭包属性、若干性质、秩图、Hill图、Hill估计的相容性和渐近正态性、平滑Hill估计和备选Hill估计、Pickands估计和Pickands图、矩估计和矩估计图,以及柯尔莫哥洛夫检验和PLFit方法等.

第5章以渐近尺度不变分布为基础推广到正则变化分布,这是一种在尾部行为上具有特定规律的分布.本章探讨了正则变化分布的若干性质和幂指数的估计

方法,为读者提供了深入的理论知识.

第 6 章进一步将正则变化分布推广到次指数分布.本章分析了次指数分布的长尾特性,并探讨了连续型的次指数分布.

最后,第 7 章将次指数分布推广到长尾分布,总结了长尾分布的相关知识,为读者提供了一个全面的视角来理解和应用重尾分布.

本书适合统计学家、数据分析师、金融分析师、风险管理师、保险业专业人士、环境科学家、经济学家、信息技术专家、公共卫生专家以及文献计量学学术研究人员等阅读.此外,本书还适合大学统计专业高年级学生、相关专业的硕士和博士研究生阅读.通过阅读本书,读者将获得对重尾分布深入的理解,以及如何应用这些知识来分析和预测极端事件,使其成为相关领域专业人士的重要资源.

作者对重尾分布的研究兴趣以及在相关领域的长期研究积累,促成了这本书的完成.本书旨在为读者提供一个全面的重尾分布指南,涵盖了从基础概念到最新发展的各个方面.对重尾分布感兴趣的学生、教育工作者、统计研究人员和普通读者,都能从中找到有价值的信息.希望本书能够为重尾分布领域作出有益贡献.

在此,我们要感谢编辑、同行评审者、研究助理等,他们的宝贵意见和建议极大地丰富了本书的内容.我们还要感谢我们的家人,他们的支持和理解是我们完成这项工作的重要动力.

限于水平,书中不妥之处在所难免,恳请大家批评指正.我们恳切希望读者在阅读过程中提出宝贵意见,并与我们联系以分享您的见解和反馈.我们的联系方式是:

王静龙: jlwang@stat.ecnu.edu.cn;

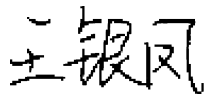
王银凤: dairy-2006@163.com;

邓文丽: wldfudan@126.com;

唐炎林: yltang@fem.ecnu.edu.cn;

吴贤毅: xywu@stat.ecnu.edu.cn;

朱仲义: zhuzy@fudan.edu.cn.



2025 年 10 月 7 日

定稿于华东师范大学丽娃河畔

目 录

“统计与数据科学丛书”序

前言

第 1 章 重尾分布的定义及其等价陈述	1
1.1 重尾分布的定义	3
1.2 重尾分布定义的等价陈述	5
第 2 章 帕累托分布	10
2.1 帕累托原则	10
2.2 帕累托分布尾部的尺度不变性	12
2.3 帕累托分布的尾部长尾和极端事件突变特性	15
2.4 秩图	20
2.5 幂指数 α 的 LS 估计 $\hat{\alpha}_n^{\text{LS}}$ 的弱相容性	23
2.6 帕累托分布参数的极大似然估计	28
2.7 Q-Q 图和 Q-Q 估计	33
2.8 超越和 POT 方法	39
2.9 帕累托分布检验	41
第 3 章 尺度不变分布	48
3.1 尺度不变分布的基本属性	48
3.2 尾部长尾、极端事件突变特性和秩图	54
3.3 LS 估计图、Q-Q 估计图和 Hill 图	56
3.4 风险值 VaR 和柯尔莫哥洛夫检验	61
3.5 PLFit 方法	64
第 4 章 渐近尺度不变分布	77
4.1 常见的渐近尺度不变分布	77
4.2 渐近尺度不变分布的闭包属性	84
4.3 渐近尺度不变分布的若干性质	88
4.4 渐近尺度不变分布的秩图	94
4.5 Hill 估计和 Hill 图	97
4.6 Hill 估计的相容性	99
4.7 平滑 Hill 估计和备选 Hill 估计	106

4.8	Pickands 估计和 Pickands 图	109
4.9	LS 估计和 Q-Q 估计的弱相容性	112
4.10	矩估计和矩估计图	118
4.11	柯尔莫哥洛夫检验和 PLFit 方法	125
第 5 章	正则变化分布	136
5.1	正则变化分布的定义	136
5.2	缓慢变化函数	139
5.3	正则变化分布的若干性质	147
5.4	正则变化分布幂指数 α 的估计	155
第 6 章	次指数分布	162
6.1	次指数分布及其突变和长尾特性	162
6.2	次指数分布的重尾特性	166
6.3	连续型次指数分布	167
第 7 章	长尾分布	178
参考文献		183
“统计与数据科学丛书”已出版书目		

第 1 章 重尾分布的定义及其等价陈述

在概率论与数理统计中, 随机变量 X 的累积分布函数 $F(x) = \Pr(X \leq x)$ 描述了变量取值不超过 x 的概率. 然而, 在许多实际问题中, 还应关心它的互补累积分布函数

$$\bar{F}(x) = 1 - F(x) = \Pr(X > x),$$

通常称 $\bar{F}(x)$ 为 X 的尾分布(tailed distribution), 简称尾分布或尾部. 它刻画了随机变量取值超过某一阈值的概率, 反映了分布的尾部特征, 与累积分布同等重要. 在可靠性分析中, 尾分布 $\bar{F}(x)$ 称为可靠度函数, 表示产品在时间区间 $[0, x]$ 内未发生失效的概率; 在生存分析中, 它则称为生存函数, 用于描述个体在特定时间之后仍然存活的概率. 因此, 尾分布在刻画随机变量的极端行为、风险评估及长期预测中具有至关重要的作用.

本书将围绕尾分布展开深入探讨, 特别关注一类具有重要理论与应用价值的重尾分布(heavy-tailed distribution)族. 这类分布的尾部衰减缓慢, 意味着极端事件发生的概率远高于常见分布(如正态分布、指数分布等)所描述的情况. 重尾分布在金融风险、保险精算、网络流量、自然灾害建模等领域发挥着重要的作用, 理解其性质与行为对于风险管理与决策制定具有重要意义. 下面我们通过具体例子给出重尾分布的前世今生.

两百多年前, 德国数学家约翰·卡尔·弗里德里希·高斯(Johann Carl Friedrich Gauss)在研究天文观测测量误差时提出了著名的正态分布. 这一分布在 19 世纪和 20 世纪初被广泛接受, 原因在于当时人们所观察到的自然和社会数据大多呈现出“中间高、两头低”的钟形曲线特征. 例如, 人的身高、测量误差、考试成绩等数据都集中在均值附近, 极端值极为罕见, 几乎可以忽略不计. 我国自 1986 年至 1990 年, 在国家有关部门的组织下, 进行了中国成年人人体测量的全国性的专项调查. 现以该调查中对不同地区的 5115 名成年男性和 5507 名成年女性的身高数据为例, 经计算, 成年男子的平均身高为 167.48 厘米, 标准差为 6.09 厘米. 据此认为: 我国成年男子身高 $X \sim N(167.48, 6.09^2)$, 根据正态分布的 3σ 原则, 有 99.73% 的成年男子身高数据落在 $167.48 \pm 3 \times 6.09$, 即 149.21 厘米和 185.75 厘米之间. 这意味着几乎完全不可能遇到身高是平均身高两倍, 即身高在 3 米以上的成年男子. 因此, 在那个时代, 正态分布被视为描述现实世界数据的理想模型.

然而,随着统计学和实证研究的深入,人们逐渐发现,许多现实世界中的数据并不服从正态分布,尤其是在经济、金融和保险等领域.例如,个人收入、企业利润、保险理赔金额等数据往往呈现出重尾特征,即尾部概率远高于正态分布所预测的水平.例如,以某保险公司 165 起事故的理赔款额数据为例,见表 1.1. 经计算,平均赔款额为 1342.47×10^2 元,但有 11 笔赔款记录超过平均赔款额两倍,5 笔赔款记录超过平均赔款额三倍,2 笔赔款记录超过平均赔款额四倍,甚至有 1 笔赔款记录超过了平均赔款额五倍.特大赔款会不会发生,这是保险公司企业管理、风险识别最为关心的问题之一.尽管尾部概率不大,也就是说发生特大赔款是个小概率事件,但保险公司绝不能低估尾部的概率.倘若认为赔款额数据服从正态分布,造成的后果之一就是错误地低估尾部,即发生特大赔款的概率.这将给公司的经营带来风险.这种“习以成俗”的现象表明,若任用正态分布来建模,将严重低估极端事件的发生概率,进而导致风险评估和决策失误.因此,重尾分布的研究应运而生,成为现代概率统计、风险管理和精算科学中的重要课题.

表 1.1 某保险公司赔款记录

(单位: 10^2 元)

1143	884	2008	578	949	435	913	1313	1536	790	1672	744
1420	647	1455	1188	1667	695	451	809	969	1139	302	720
458	1014	2304	1390	1361	790	380	1029	911	1053	1132	967
627	727	1188	1885	4220	1459	1541	1261	636	742	601	1120
691	1467	2045	843	1840	879	1293	1375	1349	1425	1390	1067
2057	938	553	717	3883	1964	1234	1324	1060	1079	3450	578
1119	470	972	1764	896	570	3739	906	4013	672	2630	825
512	2770	532	1468	1799	656	931	1575	1337	4926	1568	1172
1210	454	1032	5271	552	1872	1346	1302	1460	1135	848	797
874	1247	1241	1002	1141	1199	790	1036	819	1159	1211	1497
943	843	1167	1031	776	1128	368	1030	1131	502	1133	1092
1588	826	3915	685	1071	2563	1351	1167	1133	933	2270	1644
1002	833	1344	6759	758	914	1028	975	936	1219	2499	1027
1516	2375	1721	489	1147	595	1148	5965	927			

早在近一个半世纪以前,意大利经济学家兼社会学家维尔弗雷多·帕累托(Vilfredo Pareto)就注意到,财富在社会中的分配极不均衡,大部分财富往往集中于少数人之手.具体而言,帕累托指出,社会上约 80% 的财富掌握在 20% 的最富有者手中.这一发现后来称为“帕累托原则”(Pareto principle),亦称“80-20 原则”或“二八定律”.自帕累托提出该原则以来,其适用范围已远远超出经济学领域,成为管理学、社会学、质量控制等多个领域中广泛适用的经验法则.

帕累托原则的核心思想是：在许多情况下，约 20% 的因素往往决定了约 80% 的结果。以下是该原则在不同领域中的典型示例：

- (1) 图书馆中，约 20% 的图书承担了约 80% 的借阅次数；
- (2) 在企业中，少数员工频繁缺勤，造成了大部分出勤问题；
- (3) 约 10% 的客户贡献了约 75% 的营业额；
- (4) 在保险理赔中，少数几起案例的损失可能占据了总损失的大部分，例如 40 起索赔案例中，6 起就造成了 80% 的总损失；
- (5) 在生产车间中，少数设备(如 15 台机器中的 3 台)可能引发大部分(约 65%) 的停机事故；
- (6) 产品质量问题中，大部分缺陷往往由少数几个关键原因引起。

这些例子表明，帕累托原则揭示了资源、问题或结果在分布上的高度集中性，强调了识别关键因素对于提高效率和管理效果的重要性。

本节通过多个领域的实例，说明了重尾分布产生的背景和必要性。在计量文献学、保险、金融、意外事故与复杂系统以及网络科学等领域，人们发现许多现象无法用传统的正态分布或指数分布来准确描述。这些现象的共同特点是，极端事件或极端值出现的概率远高于正态分布或指数分布所预测的水平，即分布的尾部比这些常见分布更“厚”。例如，个人收入、保险理赔金额、金融市场波动、网络流量以及意外事故的发生频率等，都表现出明显的重尾特征。因此，为了更好地刻画这些现象，重尾分布应运而生，成为现代统计建模和风险管理中的重要工具。

1.1 重尾分布的定义

重尾分布是一类在理论和应用中都极为重要的概率分布模型。究竟什么是重尾分布？关于重尾分布的定义，文献中存在多种表达形式，其中一种较为直观的定义方法是相对于正态分布而言、基于峰度的比较方法给出的。具体而言，称随机变量 X 是重尾的，若峰度满足

$$\kappa = \mu_4 / \sigma^4 > 3,$$

其中 μ_4 是 X 的四阶中心矩， σ 为标准差。由于正态分布的峰度恰好等于 3，因此，该定义含义为：若某分布的峰度大于 3，则其尾部较正态分布更“厚”，即出现极端值的概率更高；反之，峰度小于 3，则尾部更“薄”。

一个最典型的例子是自由度为 n 的 t 分布，记为 $t(n)$ 。仅当 $n > 4$ 时， $t(n)$ 的四阶矩存在，其峰度为

$$\kappa = 3(n-2)/(n-4) > 3, \quad \text{当 } n > 4 \text{ 时},$$

相对于正态分布而言, t 分布具有更厚的尾部. 然而, 这种基于峰度的定义方式有着明显的缺陷: 首先, 该定义要求随机变量 X 至少存在四阶矩, 但是, 直观上, 尾部厚到一定程度之后, 甚至一阶矩(期望)都有可能不存在. 其次, 更为重要的是, 峰度作为全局性指标, 并不能准确刻画分布尾部的厚度. 事实上, 存在某些对称分布, 见例 1.1, 其峰度小于 3, 但相对于正态分布, 它的尾部却比较厚.

例 1.1 假设某随机变量 X 的密度函数为

$$f(x) = \begin{cases} 2x^{-6}, & x > 1, \\ 0.1, & -1 \leq x \leq 1, \\ 2(-x)^{-6}, & x < -1. \end{cases}$$

通过定义可以看出, 其为对称分布. 经过简单计算可得出它的方差为 $7/5$, 四阶中心矩为 $101/25$, 故其峰度 $\kappa = 101/49 < 3$. 倘若重尾分布可以用峰度来刻画, 那么这个对称分布的尾部就应比正态分布薄一些. 事实上, 当 $x > 1$ 时, 其密度函数 $f(x) = 2x^{-6}$, 按幂函数衰减趋于 0. 而正态分布的密度函数是按指数衰减趋于 0, 所以与例 1.1 的这个对称分布相比较, 正态分布的密度函数就较快地衰减趋于 0 了. 由此可见, 这个对称分布的尾部较正态分布的更厚.

例 1.1 表明, 峰度并不能有效刻画分布尾部的厚度. 顾名思义, 重尾分布的关键在于其尾部的特征. 因此, 判断一个分布是否属于重尾分布, 关键是考察分布尾部的渐近行为特性. 而峰度作为分布整体的一个特征量, 无法准确反映尾部的局部性质. 事实上, 峰度相等的两个分布, 其尾部特征可能存在显著差异, 甚至截然不同.

在实际应用中, 通常更关注随机变量右侧尾部的性质. 当然, 左侧尾部比较厚, 或是两侧尾部都比较厚的情况, 也被认为是重尾分布. 显然, 左右侧尾部的分析和讨论方法完全相仿. 故不失一般性, 以下仅假设 X 是非负实数值的随机变量. 此外, 若存在 $x_0 > 0$, 当 $x \in (x_0, \infty)$ 时, $\bar{F}(x) = 0$, 则该分布显然不属于重尾分布. 因此, 在讨论重尾分布时, 通常假定对所有的 $x \in (0, +\infty)$, 均有 $\bar{F}(x) > 0$, 即分布函数 $F(x)$ 的右端点满足 $\sup\{x: F(x) < 1\} = \infty$.

目前, 在理论研究和实际应用中, 被广泛接受的重尾分布定义是相对于指数分布 $\text{Exp}(\lambda)$ 的尾部 $e^{-\lambda x}$ 而言的(参见本书参考文献[1]). 具体而言, 若分布的尾部衰减速度比指数分布更慢, 则称其为重尾分布. 为更精确地刻画尾部特征, 理论研究中常采用更为严格的数学定义.

定义 1.1 称随机变量 X 是重尾的, 若对所有的 $\lambda > 0$, 都有

$$\limsup_{x \rightarrow \infty} \frac{\Pr(X > x)}{e^{-\lambda x}} = \limsup_{x \rightarrow \infty} e^{\lambda x} \bar{F}(x) = \infty,$$

重尾分布族记为 $\mathcal{H}$.

按此定义, 需要对所有的 $\lambda > 0$, 都有 $\limsup_{x \rightarrow \infty} e^{\lambda x} \bar{F}(x) = \infty$, 重尾分布的尾部比任意一个指数分布的尾部都要厚, 指数分布就显然不是重尾分布了. 为说明正态分布也不是重尾分布, 首先证明关于正态分布的尾分布的一个结果(参见本书参考文献[2], Chapter 7). 设 Z 是服从标准正态分布 $N(0,1)$ 的随机变量, 则根据洛必达(L'Hôpital)法则, 有

$$\lim_{x \rightarrow \infty} \frac{\Pr(Z > x)}{e^{-x^2/2} / (\sqrt{2\pi}x)} = \lim_{x \rightarrow \infty} \frac{1}{1 + x^{-2}} = 1, \quad (1.1)$$

$$\bar{F}(x) = \Pr(Z > x) \sim \frac{e^{-x^2/2}}{\sqrt{2\pi}x}, \quad x > 0.$$

由式(1.1)知, 对任意的 $\lambda > 0$, 都有 $\limsup_{x \rightarrow \infty} e^{\lambda x} \bar{F}(x) = 0$, 故 Z 的尾部比任意一个指数分布都还要薄, 可见它不是重尾分布. 假设随机变量 X 是重尾分布, 则 X 经仿射变换后的随机变量 $Y = cX + b$ 仍是重尾分布, 同理于随机变量 X 不是重尾分布的情况. 所以, 对于随机变量 Z 服从标准正态分布 $N(0,1)$, 其不是重尾分布, 可以推知: 任意一个正态分布 $N(\mu, \sigma^2)$ 也不是重尾分布.

注 1.1 符号“ $\sim$ ”, 除了可用来表示变量服从某种分布, 例如“ $X \sim N(0,1)$ ”意思是说随机变量 X 的分布是标准正态分布 $N(0,1)$. 除此之外, $a(x) \sim b(x)$ 还可表示为, 当 $x \rightarrow c$ 时, $\lim_{x \rightarrow c} \frac{a(x)}{b(x)} = 1$, 称 $a(x)$ 和 $b(x)$ 同阶, 其中 $c \in [-\infty, \infty]$. 本书若无特别说明, 则 $c = \infty$.

常见的重尾分布往往存在极限, 并不需要在该定义中取上极限. 通常遇到的情况是 $\lim_{x \rightarrow \infty} e^{\lambda x} \bar{F}(x) = \infty$, 较为简单.

1.2 重尾分布定义的等价陈述

下面的定理给出了定义 1.1 的两个等价陈述, 见本书参考文献[1].

定理 1.2 对于随机变量 X , 以下的三个陈述相互等价.

- (1) X 是重尾的, 即对所有的 $\lambda > 0$, 都有 $\limsup_{x \rightarrow \infty} e^{\lambda x} \bar{F}(x) = \infty$;
- (2) X 的矩母函数 $M(t) = E(e^{tX}) = \infty$, 其中 $t > 0$;

$$(3) \liminf_{x \rightarrow \infty} \frac{-\ln \bar{F}(x)}{x} = 0.$$

常见的重尾分布往往是存在极限的, 并不需要在定理 1.2(3)中取下极限. 所以通常遇到的情况是 $\lim_{x \rightarrow \infty} \frac{-\ln \bar{F}(x)}{x} = 0$, 较为简单.

证明定理 1.2 前, 首先, 分析重尾分布的这两个新等价定义的含义. 由于矩母函数是可逆的, 而且它具有良好的分析性质, 故可以用它来确定随机变量 X 的概率分布. 既然定理 1.2(2)的意思是说, 重尾分布不存在矩母函数, 所以它的分析就需要使用和非重尾分布不同的分析工具. 倘若随机变量 X 存在矩母函数, 则可导出切尔诺夫(Chernoff)界, 即当 $t > 0$ 时, 有

$$\bar{F}(x) = \Pr(X > x) = \Pr(e^{tX} > e^{tx}) \leq \frac{E(e^{tX})}{e^{tx}} = \frac{M(t)}{e^{tx}},$$

故 $\bar{F}(x) \leq \min_{t>0} \frac{M(t)}{e^{tx}}$. 若存在 $\mu > 0$, 使得 $\min_{t>0} \frac{M(t)}{e^{tx}} = \frac{M(\mu)}{e^{\mu x}}$, 则其尾部有切尔诺夫界: $\bar{F}(x) \leq ce^{-\mu x}$, $c > 0$. 倘若服从重尾分布的随机变量 X 的尾分布 $\bar{F}(x)$ 也有切尔诺夫界: $\bar{F}(x) \leq ce^{-\mu x}$, $c, \mu > 0$, 那么

$$\frac{-\ln \bar{F}(x)}{x} \geq \frac{-\ln c}{x} + \mu, \quad \liminf_{x \rightarrow \infty} \frac{-\ln \bar{F}(x)}{x} \geq \mu > 0,$$

而定理 1.2(3)告诉我们, 对于重尾分布而言, μ 必等于 0. 这说明, 重尾分布的尾部是没有切尔诺夫界的. 综上所述, 对于重尾分布的尾分布 $\bar{F}(x)$ 的分析, 需要有与非重尾分布不同的分析方法.

定理 1.2 的证明

(1) $\Rightarrow$ (2). 设 X 是重尾的, $\limsup_{x \rightarrow \infty} e^{\lambda x} \bar{F}(x) = \infty$, 其中 $\lambda > 0$. 故有严格上升的序列 $\{x_1 < x_2 < \cdots\}$, 满足条件 $\lim_{k \rightarrow \infty} x_k = \infty$, 使得 $\lim_{k \rightarrow \infty} e^{\lambda x_k} \bar{F}(x_k) = \infty$, 则

$$M(\lambda) = E(e^{\lambda X}) = \int_0^\infty e^{\lambda x} dF(x) \geq \int_{x_k}^\infty e^{\lambda x} dF(x) \geq e^{\lambda x_k} \bar{F}(x_k),$$

则 $M(\lambda) \geq \lim_{k \rightarrow \infty} e^{\lambda x_k} \bar{F}(x_k) = \infty$. 因而由条件(1)推得了条件(2).

(2) $\Rightarrow$ (3). 设 X 满足条件(2), 它的矩母函数 $M(t) = E(e^{tX}) = \infty$, 其中 $t > 0$. 使用反证法证明, 由条件(2)必然推得条件(3)成立. 倘若条件(3)不成立, 下面证明, 条件(2)也不可能成立. 设

$$\liminf_{x \rightarrow \infty} \frac{-\ln \bar{F}(x)}{x} > 0,$$

故存在 $\mu > 0$ 和 $x_0 > 0$, 使得在 $x > x_0$ 时, $\frac{-\ln \Pr(X > x)}{x} \geq \mu$, 因而

$$\Pr(X > x) \leq e^{-\mu x}, \quad \text{当 } x > x_0 \text{ 时}, \quad (1.2)$$

取 $t \in (0, \mu)$, 则对一个非负随机变量 Y 而言, $E[Y] = \int_0^\infty \Pr(Y > y) dy$, 所以

$$\begin{aligned} M(t) &= E[e^{tX}] = \int_0^\infty \Pr(e^{tX} > x) dx \\ &= \int_0^{e^{tx_0}} \Pr(e^{tX} > x) dx + \int_{e^{tx_0}}^\infty \Pr\left(X > \frac{\ln(x)}{t}\right) dx. \end{aligned}$$

由于 $\Pr(e^{tX} > x) \leq 1$, 所以 $\int_0^{e^{tx_0}} \Pr(e^{tX} > x) dx \leq e^{tx_0}$. 依据式 (1.2), 当 $x \geq e^{tx_0}$ 时, $\frac{\ln(x)}{t} > x_0$,

$$\Pr\left(X > \frac{\ln(x)}{t}\right) \leq x^{-\mu \frac{\ln(x)}{t}},$$

故而

$$M(t) \leq e^{tx_0} + \int_{e^{tx_0}}^\infty e^{-\mu \frac{\ln(x)}{t}} dx = e^{tx_0} + \int_{e^{tx_0}}^\infty x^{-\frac{\mu}{t}} dx.$$

由于 $t \in (0, \mu)$, $\mu/t > 1$, 因而 $\int_{e^{tx_0}}^\infty x^{-\frac{\mu}{t}} dx < \infty$. 从而知 $M(t) < \infty$, 条件(2)不成立. 这给我们带来了矛盾. 所以由条件(2)推得条件(3).

(3) $\Rightarrow$ (1). 设 X 满足条件(3), $\liminf_{x \rightarrow \infty} \frac{-\ln \bar{F}(x)}{x} = 0$. 则有严格上升的序列 $\{x_1 < x_2 < \dots\}$, 满足条件 $\lim_{k \rightarrow \infty} x_k = \infty$, 使得 $\lim_{k \rightarrow \infty} -\frac{\ln \bar{F}(x_k)}{x_k} = 0$. 给定 $\lambda > 0$, 则存在 k_0 , 使得当 $k > k_0$ 时,

$$-\frac{\ln \bar{F}(x_k)}{x_k} < \frac{\lambda}{2}, \quad \bar{F}(x_k) > e^{-\frac{\lambda x_k}{2}}, \quad \frac{\bar{F}(x_k)}{e^{-\lambda x_k}} > e^{\frac{\lambda x_k}{2}}.$$

因而 $\lim_{k \rightarrow \infty} e^{\lambda x_k} \bar{F}(x_k) = \infty$, $\limsup_{x \rightarrow \infty} e^{\lambda x} \bar{F}(x) = \infty$, 则由条件(3)推得条件(1). 定理 1.2 得证.

先前我们已经验证了正态分布不是重尾分布, 但对数正态分布却是重尾分布. 若 $\ln X \sim N(\mu, \sigma^2)$, 则 X 服从对数正态分布, $X \sim \text{LN}(\mu, \sigma^2)$, 其密度函数为

$$f(x) = \frac{1}{\sigma x \sqrt{2\pi}} \exp \left\{ -\frac{(\ln x - \mu)^2}{2\sigma^2} \right\}, \quad x > 0.$$

对数正态分布 $\text{LN}(0,1)$ 的密度函数的图形见图 1.1. 正态分布的密度函数是单峰对称, 对数正态分布也是单峰, 但右侧具有明显的偏性. 依据定理 1.2(3), 不难验证对数正态分布是重尾分布.

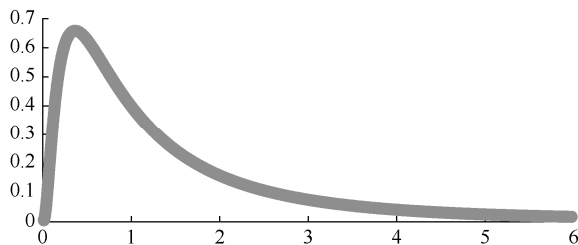


图 1.1 对数正态分布 $\text{LN}(0,1)$ 的密度函数

若 $X \sim \text{LN}(\mu, \sigma^2)$, 则 $e^{-\mu} X \sim \text{LN}(0, \sigma^2)$. 所以欲验证对数正态分布 $\text{LN}(\mu, \sigma^2)$ 是重尾分布, 仅需验证 $\text{LN}(0, \sigma^2)$ 是重尾分布. 若 X 服从对数正态 $\text{LN}(0, \sigma^2)$ 分布, 则由式(1.1)知

$$\begin{aligned} \bar{F}(x) &= \Pr(X > x) = \Pr\left(\frac{\ln X}{\sigma} > \frac{\ln x}{\sigma}\right) \\ &\sim e^{-(\ln x)^2/(2\sigma^2)} / \left(\sqrt{2\pi} \frac{\ln x}{\sigma}\right), \end{aligned} \quad (1.3)$$

则由于 $\sigma > 0$, 所以

$$\frac{-\ln \bar{F}(x)}{x} \sim \left[\frac{(\ln x)^2}{2\sigma^2} + \ln \left(\sqrt{2\pi} \frac{\ln x}{\sigma} \right) \right] / x \rightarrow 0,$$

由定理 1.2(3)知, 对数正态分布 $\text{LN}(0, \sigma^2)$ 是重尾分布, 从而对数正态分布 $\text{LN}(\mu, \sigma^2)$ 也是重尾分布. 从中我们发现了这样一个有趣的现象, X 是重尾的对数正态分布, 而 $Y = \ln X$ 却服从非重尾的正态分布. 这可能与 X 的对数变换 $\ln X$ 有关, 它将 $X=1, e=2.718, e^2=7.389, e^3=20.086, \dots$ 变换为 $\ln X=0, 1, 2, 3, \dots$. 这相当于压缩了重尾分布的尾部, 让它更快地趋于 0, 尾部就变得轻了. 对数变换, 有可能将重尾分布变换成非重尾分布. 这也就是说, X 是重尾分布, $Y = \ln X$ 有可能不是重尾分布. 倘若如此, 那我们在对来自于重尾分布的数据 $x_1, \dots, x_n$ 作统计推断时, 令 $y_i = \ln x_i, i=1, \dots, n$, 那么 $y_1, \dots, y_n$ 就有可能视为来自于非重尾分布的

数据. 然后将基于数据 $x_1, \dots, x_n$ 对重尾分布作统计推断的问题, 转化为基于数据 $y_1, \dots, y_n$ 对非重尾分布作统计推断的问题. 一般来说, 对非重尾分布作统计推断的问题, 相对容易些.

下面验证: 当形状参数 $\alpha \in (0, 1)$ 时, 韦布尔(Weibull)分布 $W(\alpha, \lambda)$ 是重尾分布. 而当 $\alpha \geq 1$ 时, 它不是重尾分布. 韦布尔分布 $W(\alpha, \lambda)$ 的密度函数和尾部分布分别为

$$f(x) = \alpha \lambda x^{\alpha-1} e^{-\lambda x^\alpha}, \quad \bar{F}(x) = e^{-\lambda x^\alpha}, \quad \alpha > 0, \quad \lambda > 0,$$

由于对任意的 $\mu > 0$,

$$\lim_{x \rightarrow \infty} \frac{\bar{F}(x)}{e^{-\mu x}} = \lim_{x \rightarrow \infty} e^{\mu x - \lambda x^\alpha} = \begin{cases} \infty, & 0 < \alpha < 1, \\ 0, & \alpha > 1, \end{cases}$$

且当 $\alpha = 1$ 时, $W(1, \lambda)$ 是指数分布 $\text{Exp}(\lambda)$, 所以 $\alpha \geq 1$ 时的韦布尔分布 $W(\alpha, \lambda)$ 不是重尾分布. 仅当 $0 < \alpha < 1$ 时, $W(\alpha, \lambda)$ 分布才是重尾分布.

韦布尔与指数分布可以通过幂函数变换相互转换:

$$X \sim \text{Exp}(\lambda) \Leftrightarrow X^{1/\alpha} \sim W(\alpha, \lambda), \quad X \sim W(\alpha, \lambda) \Leftrightarrow X^\alpha \sim \text{Exp}(\lambda).$$

若随机变量 X 是服从 $\alpha \in (0, 1)$ 的重尾 $W(\alpha, \lambda)$ 分布, 幂变换后的 $Y = X^\alpha$, $0 < \alpha < 1$, 却是服从非重尾的指数分布 $\text{Exp}(\lambda)$. 由此可见, 除了对数变换 $X \rightarrow \ln X$, 有可能将重尾分布变换成非重尾分布, 还有如幂指数 $\alpha \in (0, 1)$ 的幂变换 $X \rightarrow X^\alpha$, 也有可能将重尾分布变换成非重尾分布. 这可能与 X 变换为 X^α ($0 < \alpha < 1$) 有关, 例如幂指数为 $1/2$ 的幂函数变换就将 $X = 1, 4, 9, 16, \dots$ 变换为 $X^{1/2} = 1, 2, 3, 4, \dots$. 这相当于压缩了重尾分布的尾部让其更快地趋于 0, 尾部就变得轻了. 但在 $\alpha \geq 1$ 时, X 变换为 X^α , 例如变换为 X^2 , 它就将 $X = 1, 2, 3, 4, \dots$ 变换为 $X^2 = 1, 4, 9, 16, \dots$. 这相当于拉长了重尾分布的尾巴, 让它更慢地趋于 0, 尾部就变得更重了. 所以在对来自于重尾分布的数据 $x_1, \dots, x_n$ 作统计推断时, 除了可以令 $y_i = \ln x_i$, 也可以令 $y_i = x_i^\alpha$ ($0 < \alpha < 1$), $i = 1, \dots, n$, 那么 $y_1, \dots, y_n$ 就有可能视为来自于非重尾分布的数据. 然后将基于数据 $x_1, \dots, x_n$ 对重尾分布作统计推断的问题, 转化为基于数据 $y_1, \dots, y_n$ 对非重尾分布作统计推断的问题.

第 2 章 帕累托分布

在第 1 章中, 我们已经借助定理 1.2 从理论层面剖析了重尾分布的三个等价定义, 并通过其证明过程厘清了它们之间的逻辑关联. 然而, 理论推导的抽象性使我们难以直观地感受到重尾分布所固有的内在特性. 为弥补这一不足, 本章开始转向具体实例的观察与分析. 通过详尽探讨一系列典型的重尾分布实例, 不仅能够直观地感受到重尾分布的特有属性, 还能深入理解其在不同场景下的表现形式, 从而为重尾分布的理论研究与应用实践奠定坚实的直观基础.

2.1 帕累托原则

在连续性重尾分布族中, 帕累托(Pareto)分布是最具代表性的范例之一. 假设随机变量 X 服从幂指数为 $\alpha > 0$ (形状参数)、 $\beta > 0$ (尺度参数) 的帕累托分布, 记作 $X \sim \text{Pareto}(\beta, \alpha)$, 或简记作 $X \sim P(\beta, \alpha)$, 其概率密度函数、累积分布函数和尾分布函数分别为

$$f(x) = \frac{\alpha}{\beta} \left(\frac{\beta}{x} \right)^{\alpha+1}, \quad F(x) = 1 - \left(\frac{\beta}{x} \right)^{\alpha}, \quad \bar{F}(x) = \left(\frac{\beta}{x} \right)^{\alpha}, \quad x \geq \beta.$$

图 2.1(1)给出了幂指数 $\alpha=1.5$ 、尺度参数 $\beta=1$ 的帕累托分布的密度函数曲线: 该曲线呈典型的 L 形、单调递减且右偏特点.

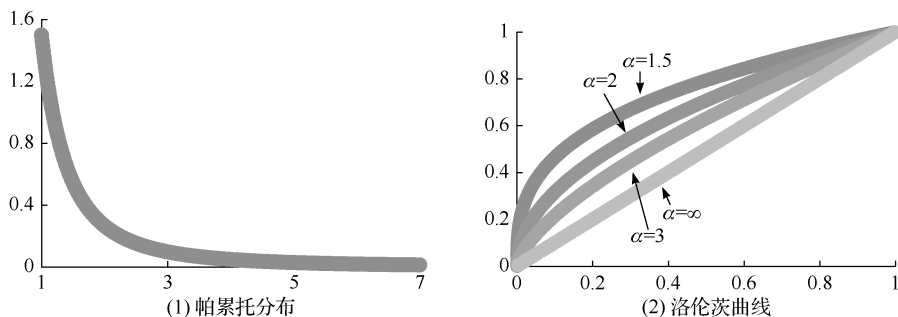


图 2.1

帕累托分布的尾部具有精确的幂指数为 α 的幂律(power law)尾形式: